# How to Write a Data Management Plan for a National Science Foundation (NSF) Proposal

Submitted by James Brunt on February 18, 2011 - 12:19am

- Data Management and Cybersecurity

The National Science Foundation (NSF) has made good the announcement in last May's press release to require a data management plan with every NSF proposal. You will be happy to know that writing a data management plan is not difficult. While constructing the text to meet the NSF requirements does demand some attention to detail, the real challenge is that the data management plan has to be non-fiction, describing procedures that will actually take place. The NSF receives about 40,000 proposals each year (source: Wikipedia). It occurred to me to wonder how those 40,000 potential investigators were going to approach this new requirement. A quick scan of blogs by scientists between now and last May when the intention was announced reveals that much single-investigator science has no process or procedures in place that could safely be called data management. The data life cycle for these projects ends with the publication of results in a peer-reviewed journal. The purpose of this briefing is to provide you with a solid outline for a data management plan to include in your NSF proposals and some resources that will help you on your way to leveraging your valuable research products through preservation and reuse.

As of January 18, 2011, all proposals to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (seeAAG Chapter VI.D.4). The NSF policy includes the sharing of results, primary data, physical samples and collections. This policy also mentions that NSF will enforce this policy through a variety of mechanisms and provide appropriate support and incentives for data cleanup, documentation, dissemination, and storage. NSF suggests that the plan "may" contain:

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.

NSF stops short of dictating what data management practices you should engage in. This means if there are community standards they will be applied through peer review pressure. While in some communities this means you can probably get away with two sentences saying how much you don't need a data management plan, that's not true in the ecological community where there are standards of practice and experienced informatics-oriented colleagues on the review panels. Some NSF directorates and divisions have issued advice to proposers that contain more specific suggestions (e.g. SBE, EAR, MPS). In addition, institutions are beginning to post resources for their constituents that can be of use in developing a data management plan (e.g., MIT, UWM).

If you are reading this first hand then you are in luck because you are in some way associated with an LTER site. LTER proposals have been going in with data management plans and backed up by data management procedures for the last 30 years. This means that there is expertise for you to draw on to prepare your plan and more importantly resources to guide you down the road to fulfilling your plan. (Note: It has been expressed by an NSF source that a PI adopting their LTER site research


This material is based upon work supported by the National Science Foundation under Cooperative Agreements DEB#9634135 (3/15/97 - 2/29/04), DEB#0236154 (3/1/03 - 2/28/10), #DEB-0832652 (5/1/09 - 4/30/15 (Core funding)), DEB#0936498 (9/1/09 - 8/31/14 (ARRA funding)).

data management plan for their proposed projects to other NSF programs would be viewed favorably.) If you've received this via a colleague or through the magic of Google then I hope that I can give you some added confidence in the composition of your data management plan.

The National Science Board in its 2005 recommendations to NSF, NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century, intended these data management plans to be quite comprehensive. With this 2-page directive, however, NSF is particularly interested in data management with regard to the dissemination and sharing of research results. While the instructions below reflect desirable data management practices, there are several essential issues among them that deserve more weight in your write-up for NSF. I will identify these in the text below. As with LTER proposals, any specific solicitation instructions trump this 2-pager in terms of expectations but must still include the essential information below.

Step 0. Label the page - "Data Management Plan"

Step 1. Collection - Describe the data to be collected during the proposed period of operation. These are the actual observations, not the final derivative product. This can be prose if simple or a table if more complex. Name the type of data (e.g., mass of seeds, counts of inflorescences), the instrument or collection approach (e.g., visual count recorded on paper), and the sampling design (e.g., number of plots, replicates, frequency of collection). If actual data are interpreted, note the interpretation (e.g., impedance interpreted as soil moisture). If data volumes are significant (e.g., >1Gb/day) indicate an estimate of the totals. Describe any quality control measures that will be put in place as part of data collection.

Step 2. Processing - Describe the disposition of the raw data post-collection. How will data be transmitted from field or instrument to institution? How regularly, by whom, and where will data be stored? How will the security of those data be ensured? A previous article describes several rules of thumb for data security (LTER Data Management and Cybersecurity Briefing #1).

Step 3. Analysis - Describe in general any descriptive or analytical statistics that will be run against the data for quality assurance, derivation, aggregation, etc. Mention the names of analytical packages (e.g, SAS, SPSS, MatLab, R).

Step 4. Documentation - Documentation is required to ensure the longevity of data. The documentation of your study is best done during the process, not after. This step describes the accumulation of the documentation text, while Step 8 describes the encoding of this text into a metadata language for publication. Here you will describe what metadata/documentation will be created at each stage of the data life cycle and by whom. For example, "Changes made to the data to correct errors will be described and revised during the data manipulation process by the budgeted graduate student". Examples of good metadata can be seen in the LTER data catalog or consult with your Site Information Manager. What is the metadata content standard you will use to document these data? Most ecological metadata is based on recommendations contained in Michener et al. 1997.

Step 5. Products (Essential) Describe the data or other products that you will be making available from the study. These may or may not be the raw data described in step 1. This is another place where a table might be useful.

Step 6. Policy (Essential) Describe the policies under which these data will be made available (See LTER Data Access Policy for example) and how you will deal with privacy or other sensitive data issues (e.g., location of endangered species).

Step 7. Archival (Essential) Describe how and where you will make these data and metadata available to the community in perpetuity. Here again you have an advantage by being associated with an LTER site. LTER sites maintain archival infrastructure for making data and metadata accessible and can give you tips and maybe some direct support. If not, most institutional libraries operate digital repositories that will provide this service for their constituents.


This material is based upon work supported by the National Science Foundation under Cooperative Agreements DEB#9634135 (3/15/97 - 2/29/04), DEB#0236154 (3/1/03 - 2/28/10), #DEB-0832652 (5/1/09 - 4/30/15 (Core funding)), DEB#0936498 (9/1/09 - 8/31/14 (ARRA funding)).

Step 8. Curation (Essential) - Preparation of metadata and data for publication is a time consuming process. This should be acknowledged in the data management plan and in the budget. In this step you will describe the structural standards that you will apply in making data and metadata available. For example, for most ecological data, documentation will need to be structured in Ecological Metadata Language (EML) to be included in community repositories. There are best practices available from the LTER community for EML. However, you can avoid direct contact with EML and best practices documents by registering your datasets online with the Knowledge Network for Biocomplexity (See Step 9.)

Step 9. Publication (Essential) - After making sure you have a secure place for your data products to reside, you need to register them with community repositories. Include a description here of the institutional repository(s) where you will register your data. Your LTER site can register and publish your data. If that is not appropriate for your study, the LTER Network operates as a node on the Knowledge Network for Biocomplexity (KNB) where these data can be independently registered. KNB offers an online repository form and a guide for completing the form. The NSF DataNet projects, in particular DataONE, will hopefully soon offer another outlet for data publication.

For specific datasets you may consider formally publishing the data. Ecological Archives is a peer-reviewed data journal operated by the Ecological Society of America that accepts well described datasets and their textual description for publication. There are others operated in various ways by scientific societies. Avoid only committing your data to commercial journal repositories for what I hope are obvious reasons.

Other considerations:

The information contained in the plan regarding “plans for preservation, documentation, and sharing of data” is also required to be part of the Project Description - - so it seems that placement of an appropriate reference to the 2-page plan in the project description would be prudent.

Make sure your proposed budget addresses the data management plan. Costs of documenting, preparing, publishing, disseminating and sharing research findings and supporting material are allowable charges against the grant.

Data management plans and procedures should become standardized for a lab, institute, or even community such that in time there is boilerplate material available that reflects institutionalized procedures.

Ultimately the success of any given plan will lie in the hands of the reviewers and the makeup of the panel, but as with any new initiative those 40,000 proposals that go in first tend to set the tone for the future. Finally, just before going to press I read in a reliable source that DataONE and others are developing a software tool that will write data management plans for you. Until that time, I hope you find this information useful.

Comments and discussion are encouraged and should be directed to the online forum so that the community may benefit.



‹ On the Road: Fear and Living with Public Computing. Protecting Your Digital Research Data and Documents ›


This material is based upon work supported by the National Science Foundation under Cooperative Agreements DEB#9634135 (3/15/97 - 2/29/04), DEB#0236154 (3/1/03 - 2/28/10), #DEB-0832652 (5/1/09 - 4/30/15 (Core funding)), DEB#0936498 (9/1/09 - 8/31/14 (ARRA funding)).