NSF ENG Data Management Plan Template

for the University of Michigan College of Engineering

Overview

NSF proposals submitted after January 18th, 2011 must include a Data Management Plan (DMP), a supplementary document of not more than two pages. Proposers may use part of the 15-page Project Description for additional data management information.

The purpose of the DMP requirement is to provide clear, effective, and transparent implementation of the long-standing NSF Policy on Dissemination and Sharing of Research Results, which may be found in the Award Administration Guide, Section VI.D.4. A general description of the contents of a DMP can be found in NSF's Proposal and Award Policies and Procedures Guide. Many NSF directorates and divisions have issued their own guidance on preparing the DMP; consult the one(s) relevant to your proposal [4]. There may be additional requirements provided in the solicitation; these can generally be found in a section entitled "Proposal Preparation Instructions." Contact your program officer with any questions.

This document is intended to be used for preparing CoE ENG proposals. Requirements are based on ENG documentation, examples are taken from past CoE proposals, and advice is geared to CoE faculty. For guidance from other NSF directorates and divisions see the U-M library's NSF DMP information page.

Instructions

The left column contains a series of questions to be answered by the preparer, as well as example answers. The right column contains advice and commentary on the questions and examples, with links to supplementary documentation and related resources where appropriate. If possible, read through the entire document before answering the questions. Although it refers to UK funder requirements, another good preliminary read is [30].

Few proposals need to address every question, and few PIs will have all the answers at the application stage. The DMP should be thought of as a sketch of procedures to be fleshed out after the grant has been awarded. If you need further assistance, see Appendix A for a list of U-M people who can assist with DMP preparation.

Color Key

Quote from NSF or ENG guidance
Positive Example
Negative Example

Note: Most examples were chosen to highlight a specific aspect of the question, and many exhibit a mix of positive and negative aspects. We have tried to point these out when there is potential for confusion.

Credits

The questions in this document are adapted from Purdue University Libraries' *Data Management Plan Self-Assessment Questionnaire,* Purdue University, West Lafayette, IN 2/4/2011; and *NSF ENG Directorate Data Management Plan Template* from the Scientific Data Consulting Group at the University of Virginia Library, version 2.5, 3/30/2011. NSF guidance is taken from NSF and ENG documentation. Examples are taken from data management plans submitted by PIs in the U-M College of Engineering.

This document was created by Jacob Glenn, Physics & Astronomy Librarian, Shapiro Science Library, with input from the CoE's Office of the Associate Dean for Research, CoE Research Administration, CAEN, DRDA, the Office of Tech Transfer and the library's Copyright Office. Send feedback to sciencelibrary@umich.edu.



DMP excerpts are copyrighted by their respective authors. The rest of this work is ©2011, Regents of the University of Michigan, and is subject to a Creative Commons Attribution 3.0 license. Details and exceptions.

Contents

I. Roles and Responsibilities 1

II. Expected Data 2

III. Data Formats and Metadata 6

IIIa. Data Formats 6

IIIb. Metadata 8

IV. Data Sharing, Access & Rights 10

IVa. Data Sharing and Access 10

IVb. Rights and Conditions for Re-Use 15

V. Data Archiving and Preservation 16

Va. Period of Data Retention 17

Vb. Archiving and Preservation Strategy 18

References 21

Appendix A: DMP Contacts at the University of Michigan 23

Appendix B: Software Licensing 24

Appendix C: Applying a License to your Software 27

Appendix D: A Data Sharing and Archiving Timetable 28

I. Roles and Responsibilities

The DMP should clearly articulate how "sharing of primary data" is to be implemented. It should outline the rights and obligations of all parties as to their roles and responsibilities in the management and retention of research data. It must also consider changes to roles and responsibilities that will occur should a principal investigator or co-PI leave the institution. Any costs should be explained in the Budget Justification pages.

http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

Summary: *Explain how you will assign responsibilities for managing your data in the manner described by the remainder of your DMP. Outline the rights and responsibilities of all project participants as to their roles in the the management and retention of research data generated during the project. Also consider any changes to these roles and responsibilities that would occur if a PI or co-PI should leave his or her current institution, and describe the procedure for transferring responsibility should this happen during the anticipated lifespan of any data you plan to preserve.*

NSF DMP FAQ:

http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp

1. What are the responsibilities of staff and investigators for implementing the present DMP? Include time allocations, project management responsibilities, training requirements, and contributions of non-project staff. Name specific individuals where possible.

See DataONE Best Practices, *Recognize stakeholders in data ownership:*

https://www.dataone.org/content/recognize-stakeholders-data-ownership

"For the proposed collaborative research, Dr. __ at Pennsylvania State University will take the lead and responsibility for coordinating and assuring data storage and access. However, Dr. __ at the University of Michigan will also be involved in managing, storing, and disseminating the results of the project, particularly in regards to data acquired through the experimental testing associated with the ESMA isolation devices."

This paragraph names specific individuals and states their roles in managing specific types of data.

"All the investigators involved in the proposed project have equal rights to access the data generated through this sponsored project. They also have the same obligations to share data with each other and to publish the results in a timely manner."

This boilerplate paragraph is a good start, but needs to be elaborated with more specific information about project roles.

2. How will the PI(s) verify that the data generated are being managed according to this plan? At what point(s) in the project will this happen? Who is responsible for checking that the plan is being followed?

3. Is there a formal process for transferring responsibility for the data should a PI or co-PI leave his or her institution?

Spell out how this will happen and who will be responsible for ensuring that it goes smoothly.

"Should the PI leave the University of Michigan, the grant would likely be transferred. If not, the co-PI will assume leadership of the project and responsibility for data storage."

The co-PI is designated explicitly, a good start. This statement should be elaborated to include specific procedures for transfer.

"Should the PI move to a different institution, the web site can be easily moved to the co-PI's web space."

"There will be no change to the scope of this data management plan should the PI or co-PI leave the University of Michigan."

The scope of the data management plan needs to include provisions for what will be done in this case, so stating that it will not change is insufficient.

4. Who will have responsibility for decisions about the data once all the original personnel are no longer associated with the project? Is there a procedure in place for transferring responsibility once the original personnel are no longer available?

This question should be answered with reference to your long-term data preservation strategy as described in Section V.

5. Who will bear the costs associated with data preparation, management and preservation?

The detailed costs of implementing the plan should be specified in the Budget Justification portion of your proposal. Also see DataONE Best Practices, *Provide budget information for your data management plan:* https://www.dataone.org/content/provide-budget-information-your-data-management-plan

"We do not expect the project to produce more data than can be appropriately managed by lab personnel."

This may be true, but it should be justified, at least by reference to other portions of your data management plan.

II. Expected Data

"The DMP should describe the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project. It should then describe the expected types of data to be retained."

http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

Summary: *Describe the data you will produce in the course of the project (see the ENG directorate guidelines for what types of data to include). State roughly how much data you'll generate and at what rate, if possible. If you will be generating multiple data sets, answer the questions below for each data set. If you know you won't be keeping all the data you generate, state what you will and won't retain and why.*

1. What is the general nature of the data you will be generating? Experimental measurements, qualitative data, simulations?

Provide a general description of the content, including anticipated size and volume of data if known.

"This research project will generate data resulting from sensor recordings (i.e. earth pressures, accelerations, wall deformation and displacement and soil settlement) during the centrifuge experiments. In addition to the raw, uncorrected sensor data, converted and corrected data (in engineering units), as well as several other forms of derived data will be produced. Metadata that describes the experiments with their materials, loads, experimental environment and parameters will be produced. The experiments will also be recorded with still cameras and video cameras. Photos and videos will be part of the data collection."

"A total storage demand of 50 GB is anticipated at the University of Michigan, and 50 GB at Auburn University."

"Based on the previous viscoelastic turbulent channel flow simulations, the amount of resulting binary data is estimated around 40 TB per year. Some text format data files are also required for post-processing in the laboratory and are anticipated to be around 1 TB per year."

"In one year, we will perform approximately 2 to 3 simulations. This means ~100 3D plots, 30 restart files, 1000 EUV, X-ray and LASCO-like images, 10 satellite files, 1000 2D plot files (total of about 150 GB of data per year)."

These three examples all illustrate parts of a good answer to this question. The first lists the various types of data that will be generated, the second states how much data will be created in total, and the third estimates the volume of data to be created per year. Your plan should address all of these elements, if possible.

This is a good level of detail to include.

"The main goal of this project is to conduct simulations to better understand the thermosphere and ionosphere. Therefore, the data that will be produced from this project are simulations. The model that we utilize produces 3D data covering from 100 km to 600 km altitude with roughly 50 grid spacings. In the latitude and longitudinal directions, the spacing is typically 2.5 x 2.5 degrees."

A common error in this section is to lapse into a recap of the project summary, as illustrated by this example. Stick to describing the types of data to be generated, touching on methods only when necessary to explain what (or how much) data you will be creating.

"The nature of the data or other materials produced under this NSF-sponsored project will include data characteristics such as observational, experimental, reference, derived, simulated and/or other. The data types referenced could include data generated by computer, data collected from sensors or instruments, images, video files, reports, and/or other."

This paragraph reads as if it was lifted from funder documentation and tells the reviewer nothing about the nature of the data to be generated under this proposal.

2. How will the data be created or captured?

Be as specific as possible. What stages will the data pass through? Raw, processed, analyzed, published? What methods will you use to get to each stage? What tools (including software and instruments) will you use at each stage? Which personnel will be involved at each stage? Include details about local backup and disaster recovery procedures here, or in Section Vb.

"The data, samples, and materials expected to be produced will consist of laboratory notebooks, raw data files from experiments, experimental analysis data files, simulation data, microscopy images, optical images, LabView acquisition programs, and quantum dot superlattice nanowire thermoelectric samples.... each of these data is described below:

A. Laboratory notebooks: The graduate student and PI will record by hand any observations, procedures, and ideas generated during the course of the research.

B. Experimental raw data files: These files will consist of ASCII text that represents data directly collected from the various electrical instruments used to measure the thermoelectric properties of the superlattice nanowire thermoelectric devices.

ASCII is a character encoding, not a format. Additional information about the contents of the files and what they represent is desirable (you can cover this in Section III).

C. Experimental analysis data files: These files will consist of spreadsheets and plots of the raw data mentioned in Part A. The data in these files will have been manipulated to yield meaningful and quantitative values for the device efficiency and ZT. The analysis will be performed using best practice and acceptable methods for calculating device efficiency and ZT.

If you plan to generate spreadsheet files, consider exporting tab-delimited or CSV formats instead of, or in addition to, Excel workbook files. See also: https://www.dataone.org/content/preserve-information-keep-your-raw-data-raw

D. Simulation data: These data will represent the results from commercially available simulation and modeling software to model the quantum confinement.

If you know which software packages you will be using it would be good to list them here (including version numbers if possible). Consider exporting to common interchange formats if they exist (and be sure to describe them in Section III).

E. Microscopy images: Images of the proposed silicon nanostructures will be generated by scanning electron microscopy (SEM), transmission electron microscopy (TEM) at high resolution to quantify wire diameter and roughness, and atomic force microscopy (AFM).

F. Optical images: Images of the nanostructured devices will be collected using an optical microscope at various magnification settings.

G. LabView acquisition programs: Various programs will be used to interface a personal computer to the multiple electrical instruments used to measure the device efficiency and ZT. The programs will allow the experimenter to collect in real-time electrical data corresponding to temperature, thermoelectric voltage, electrical resistance, etc.

H. Superlattice nanowire samples: The nanostructured samples will consist of silicon quantum dot superlattice nanowires. The experimenter will use these samples to measure device efficiencies and ZT."

"The measurement data are generated by various laboratory electronic measurement instruments. Furthermore photographs of the fabricated circuits will be a part of the data generated. The circuit simulation data are produced by computers and are in the form of graphs and tables describing various performance aspects of the circuits being analyzed."

This answer is too vague to give the reviewer a sense of what data will be created.

3. If you will be using existing data, state that fact and where you got it. What is the relationship between the data you are collecting or generating and the existing data?

"The field data collection will augment existing data sets without creating issues of redundancy. The following *existing* data sets will be used:

Topographic data: data will be obtained through the geodata web portal (*http://gos2/geodata.gov/*); the data will be used to describe topography of all case study basins.

Soil texture data: available from the SSURGO database of the Natural Resources Conservation Service (*http://soils.usda.gov/survey/geography/ssurgo/*). These data will be used to infer soil water retention and hydraulic conductivity relationships using pedotransfer functions. . ."

This response clearly states the nature and provenance of the existing data sets to be used, and explains how they relate to the new data sets that are to be produced in the course of the proposed project.

"Our proposed work does not collect any new observations; we only use existing observations and those in the process of being collected through previously funded NSF projects."

Even if the project won't produce new observations, a re-analysis of existing data will likely result in new datasets that fall under the ENG definition of data.

4. Which data sets will be archived (preserved for the long term) and made available, and which will not? On what basis will data be selected for preservation and sharing?

For data you do not plan to archive, why not? Be sure to review the ENG guidelines if you are unsure which data need to be archived.

"All data will be stored on the PI's server... This includes all aforementioned data levels: raw data, prepared data, and analyzed data, as well as the metadata associated with each data stream. Of these, the analyzed data, which is equivalent to data that is published in scientific journals, and the metadata will be made publicly accessible."

See Section V for an explanation of why storing data on your own server may not be the best solution to fulfilling long-term archiving requirements.

"In the event that we generate data that we do not publish or post online as supplementary data in conjunction with a manuscript, these data will also be shared at the time of completion of the study as described in the proposal. We envision that these data could include computer files of photomask designs, microfabrication recipes, and custom-developed MATLAB programs to analyze biomolecule motion."

"Many simulations will not be analyzed in detail if they do not illustrate the main conclusions reached by the investigators. Hence it is unlikely that they will be included in published work. However it is of course possible that such data runs might contain other results that escape notice during the lifetime of the project and could be useful to other researchers at a later date. Hence the results of numerical data runs will produce large data sets which merit storage."

"For large scans, e.g., over part of an actual mummy, raw data may be produced at a rate of as much as 6000 time-domain waveforms per minute for scans that take tens of minutes. These scans can produce tens of GB of data during a scan, although their real utility is as reduced, graphical images. In this case, only data reduced to B-scan or C-scan images will be made generally accessible as TIFF files."

Each of these four responses illustrates a piece of a good answer to this question. The first example makes a distinction between raw and analyzed data and states which will be subject to long-term archival storage requirements. The second mentions specific types of supplementary data that will be shared. The third describes criteria (which could be made more explicit) for determining whether to archive additional data that are not included in published results. The fourth excludes certain raw datasets from sharing on the basis of their size and lack of utility.

III. Data Formats and Metadata

"The DMP should describe the specific data formats, media, and dissemination approaches that will be used to make data available to others, including any metadata."

http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

IIIa. Data Formats

Summary: *Describe the formats (file types) your data will be in. Proprietary formats are more difficult to preserve, as the software and hardware that reads them quickly becomes obsolete. Data should be stored in stable, non-proprietary file formats, preferably those based on open and published standards, whenever possible. If your research will generate files in proprietary formats, consider converting those files into formats based on open standards for sharing and archival purposes.*

1. Which file formats will you use for your data and why?

"Verilog, SPICE, and MATLAB files generated will be processed and submitted to FTP servers as .mat files with TXT documentation. The data will be distributed in several widely used formats, including ASCII, tab-delimited (for use with Excel), and MAT format. Instructional material and relevant technical reports will be provided as PDF. Digital video data files generated will be processed and submitted to the FTP servers in MPEG-4 (.mp4) and .avi formats. Variables will use a standardized naming convention consisting of a prefix, root, suffix system."

"The data will be in standard, non-proprietary file formats to facilitate both data sharing and long-term preservation. The simulation code will be developed in C and provided to the public in source code format for non-commercial use under GNU General Public License (GPL). The numerical data will be in TXT format so that they are readable by any text editing software. The data can be visualized by using our existing MATLAB code to generate JPEG pictures or MPEG animations. The matlab visualization code will be deposited together with the data. Data and documents in these formats (C, TXT, JPEG, MPEG) will likely be standard for a long time due to their wide use. In the situation that these formats become obsolete in the future, the documents will be reformatted."

"Plasma image data will be RGB colored JPG or TIFF format with resolution determined by the camera. Video data will be RGB colored AVI format."

"Images from the scanning electron microscopes (SEMs) and focused ion beam workstations (FIBs) are saved in tagged image file format (TIFF), which is readily readable by a wide variety of imaging and processing applications."

Try to keep in mind who the potential users of your data are and what their needs might be. For example, if you know your data will only be of interest to a small group of specialists, all of whom use a certain commercial software package, then it might be acceptable to share the data in a proprietary format that's only readable by that package (provided the software isn't expected to become incompatible during the planned lifespan of the data). However, if you want your data to be useful to a wider audience then you may need to consider translating them into a more widely accessible format.

A good resource for identifying formats is [6]. The Library of Congress [7] also offers information on sustainable digital formats.

Best practices for producing PDF files:
http://hdl.handle.net/2027.42/58005

Best practices for producing ZIP and tar files:
http://deepblue.lib.umich.edu/handle/2027.42/50495

Best practices for producing datasets:
http://deepblue.lib.umich.edu/handle/2027.42/40246

Best practices for producing image files:
http://deepblue.lib.umich.edu/handle/2027.42/40247

Best practices for producing digital video files:
http://deepblue.lib.umich.edu/handle/2027.42/83222

Best practices for producing digital audio files:
http://deepblue.lib.umich.edu/handle/2027.42/40248

These examples all illustrate a preference for non-proprietary data formats based on open standards.

"The data format includes digital data recorded by computers and instruments and metadata recorded in lab notebooks and reports."

"Data will be stored in the following formats: Microsoft Excel, PowerPoint, Microsoft Word and Acrobat PDF."

"The format of the electronic data will be specific to the format used by the particular software in which it was created. For data generated from instruments, the output will often be in a proprietary ASCII format, in some cases non-proprietary text format will be available."

This answer is too vague to be informative.

Best practices for Microsoft Office files: http://deepblue.lib.umich.edu/handle/2027.42/40245

If proprietary formats must be used, provide a means for translating them to standard formats or justify your decision not to do so.

2. Which standards will you use and why have you chosen them? (E.g. they are accepted domain-local standards or receive widespread use in your research field.) If your data are generated in a non-standard format, how will you convert them to a more accessible format? Where and how will you make the conversion code available, if you plan to write it?

See DataONE Best Practices, *Document and store data using stable file formats:*

https://www.dataone.org/content/document-and-store-data-using-stable-file-formats

"The output files will be in various binary formats that can be directly read by commercially available visualization software (TecPlot and IDL), and we will also produce data in standard NetCDF and HDF5 formats. All these data formats contain metadata that describes the simulation grid and simulation time of the sequence, the variables and their physical units."

"Whenever possible, standard formats of data will be used, e.g.: images: TIFF, BMP, JPG."

Naming the commercial software packages you plan to use is a good idea, but try to include the version numbers, if known.

3. If there are no applicable standard formats, how will you format your data so that other researchers can make use of them (keeping in mind that others may not have access to the software or instrumentation you used to generate them)?

"Digital Micrograph, although very common, uses a proprietary format and so copies of DM data files are stored in TIFF format from within DM or with the aid of a converter algorithm available as a macro in ImageJ.... AFM data from EMAL's Bruker Dimension Icon is recorded in a proprietary format but also can be exported as ASCII, or the files may be analyzed direct with Image SXM or the commercial SPIP software, for which EMAL has a license."

Converting proprietary image file formats to TIFF is a good use of standardized formats, though ideally the conversion code should be made available as well. ASCII is too general to designate as a file format (it's a character encoding); try to say something more specific about how the ASCII files will be formatted.

4. Who on your team will have the responsibility of ensuring that data standards are properly applied and data properly formatted? What procedures will be in place to ensure that this is done consistently throughout the duration of the project?

See DataONE Best Practices, *Develop a quality assurance and quality control plan:*

https://www.dataone.org/content/develop-quality-assurance-and-quality-control-plan

IIIb. Metadata

Summary: *A metadata record is a file that captures all details about a data set that another researcher would need to make use of the data set in a separate or related line of inquiry. Metadata captures the who, what, when, where, why and how of the data you produce.*

When data curators talk about metadata they are normally referring to a machine-readable description that comes in a standardized format, often defined by an XML schema. Many scientific communities have developed these schemas for the consistent description of datasets. Examples include Ecological Metadata Language, Minimum Information About a Proteomics Experiment (MIAPE), and the Data Documentation Initiative (DDI). While human-readable descriptions of your data are better than nothing (and are more useful when they are created using standard formats such as .txt instead of proprietary document formats), a standardized, machine-readable metadata file is preferable to a human-readable description.

See DataONE Best Practices, *Identify and use relevant metadata standards:*

https://www.dataone.org/content/identify-and-use-relevant-metadata-standards

5. What contextual details (metadata) are needed to make your data meaningful?

This question asks you to describe the metadata elements you will use. The next question asks about metadata formats.

"Metadata will include time, date, and location of measurement, object measured, THz equipment used, and personnel present."

"Data will be identified by the date it was gathered, the name of the investigator who obtained it, and a short descriptor. A detailed description of the data and experimental conditions will be included in an accompanying text file and/or laboratory notebook description."

"Metadata will include SWMF version, event simulated, time of simulation, model parameters, type of plot (3D, 2D slice). For the synthetic observational files this includes spacecraft type (STEREO, SoHO, SDO) and position in its orbit, and instrument type (HI, C3, COR3, EIT, EUVI, AIA, in-situ, etc.)."

"For all of the numerical simulations, representative input files will be kept along with records of the suites of ground motions used and any scaling performed on the ground motions. Figures with the archetype structure's dimensions, nodes, and elements labeled will also be stored with the numerical results for a more clear understanding about where and how the data was obtained. For all of the experimental tests, descriptions and photographs of the test setup will be stored with the experimental data along with schematics specifying dimensions and location of instrumentation. A spreadsheet will be maintained for all experimental tests run specifying the date, time, and parameters for the test along with any comment made during the running of the test."

These are all good answers due to the level of detail included.

"In addition to data, metadata concerning how the data was generated (software, hardware, dates, measurement protocols, etc.) will be maintained and disseminated on request."

One primary purpose of metadata is to enable others to discover your data. Providing it on request defeats this purpose. This answer also needs to go into more detail about the metadata elements to be generated.

6. What form or format will the metadata describing your data take? Which metadata standards will you use? If there is no applicable standard, how will you describe your data in a way that will make them accessible to others?

"The Dublin Core will be used as the standard for metadata. The metadata set mainly consists of fifteen elements, including title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. These elements have been ratified as both national (i.e., ANSI/NISO Standard Z39.85) and international standards (i.e., ISO Standard 15836). Further, they describe resources such as text, video, audio, and data files. These standard formats will be used in our study."

This is a good use of metadata standards, but remember that one size does not fit all. Dublin Core is a very general standard; you should prefer a standard that is specific to the type of data you plan to generate.

"For each code made available, a user's manual will be provided with instructions for compiling the source codes, installing and running the codes, formulating input data streams, and visualizing the output. Documentation will be in PDF format."

Documentation is a form of metadata, though it doesn't substitute for machine-readable metadata. A text file is preferable to PDF because more programs can read it.

"Embedded comments in English will be used to provide sufficient metadata to interpret the meaning of the text files. We have chosen this metadata format because it is flexible and enables later interpretation."

Unless no applicable metadata standards exist, this should be considered a last resort.

"Metadata will be stored in Microsoft Word documents."

Is Microsoft Word really necessary here? A simpler format would be more accessible.

7. How will metadata files be generated for each of the data sets you produce? Who will do the work of data description and how will the costs be borne?

Include any costs associated with data description in the budget justification section of your proposal, but briefly describe them here.

"Metadata will be created as soon as the data is collected or produced, thus allowing efficient management and rapid sharing of the data with others."

This is an ideal approach, but the answer needs to go into more detail about procedures for metadata creation.

8. Who on your team will be responsible for ensuring that metadata standards are followed and are correctly applied to the corresponding data sets?

This may be the responsibility of someone else, such as a curator at a disciplinary data repository (see Section V), but most repositories require metadata to be submitted along with data files.

See DataONE Best Practices, *Confirm a match between data and their description in metadata:*

https://www.dataone.org/content/confirm-match-between-data-and-their-description-metadata

IV. Data Sharing, Access & Rights

"The DMP should clearly articulate how 'sharing of primary data' is to be implemented. It should describe the dissemination approaches that will be used to make data available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. Research centers and major partnerships with industry or other user communities must also address how data are to be shared and managed with partners, center members, and other major stakeholders. Publication delay policies (if applicable) must be clearly stated. Investigators are expected to submit significant findings for publication quickly that are consistent with the publication delay obligations of key partners, such as industrial members of a research center."

http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

IVa. Data Sharing and Access

"Investigators are expected to share with other researchers, at no more than incremental cost and within reasonable time, the primary data, samples, physical collections, and other supporting materials created or gathered in the course of work under NSF grants."

http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

Summary: *Explain, in as much detail as possible, how and when your data will be made available to people outside your research group. Keep in mind the potential users of the data as well as any norms for data sharing that exist within their communities. For instance, if you are generating data that is of potential interest to researchers who are heavy users of a disciplinary data repository then it is probably not sufficient to make your data available by email request to the PI. On the other hand, if that is the standard method of data sharing in the relevant user communities then it may be an acceptable solution.*

By default, "access" means unmediated public access to your data (possibly with an embargo period). The rest of this section is largely concerned with reasons why you might need to limit that access.

See also DataONE Best Practices, *Sharing data: Legal and policy considerations:*

https://www.dataone.org/content/sharing-data-legal-and-policy-considerations

1. Who is likely to be interested in your data? (E.g. other researchers in your field, researchers in other fields, the general public?) For how long after the conclusion of the project are your data likely to be of interest to these people?

"The data will be of interest to the algal biofuels community, as well as to the more general chemical engineering and renewable energy communities."

"These data will provide a detailed experimental look at the mechanical regulation of mesenchymal stem cell osteogenesis. The data will further delineate the functional role of the cytoskeleton-focal adhesion-extracellular matrix signaling axis in the mechanoresponsive mesenchymal stem cell osteogenesis, as described in the main body of the proposal. As such, they will be of interest to the tissue engineering and regenerative medicine communities."

"No, or very little, derivative use of the data is expected since the research is aimed at developing machining process models and the data to be gathered is directed toward this end."

A belief that others won't be interested in your data is an insufficient reason for declining to share it.

2. How and when will you make the data available? Describe resources needed to do this, such as equipment, software and staff. Does your intended method of dissemination match the data sharing norms practiced by the groups you described in response to the previous question?

Remember that the smaller the effort that is required to obtain your data, the more people are likely to use it (and to cite your research).

"Due to the huge size of the data, access will be through GridFTP via Teragrid, a secure, fast and optimized data transfer protocol for high-bandwidth networks."

"The PIs will make copies of data available to co-investigators, students, and others by request within 45 days from receipt of the request unless a longer period is necessary for protection of intellectual property.... To facilitate the sharing of data, project data will be made publicly available via a university-controlled repository, Deep Blue."

A good start, but where will the server be? You can refer to what you've written in Section V when discussing your plans for data archiving and preservation of access.

If planning to use Deep Blue to archive and disseminate your data you should first become familiar with its policies [9].

"We plan to archive and make available by request data that are used to produce published results. We will either use email or HTTP to provide access, depending on the contents of the request."

"Any data that is required to be publicly accessible is placed in a generally accessible read-only directory. Computer security rules established by the University of Michigan, however, still require the use of an SFTP client to access the public data."

"Members of the research project team will be able to share data using the University of Michigan secure CTools website.... Access can be granted for any University of Michigan personnel and, if necessary, external access can be granted on a case by case basis."

"Data will be shared by mailing of a CD with the data to the requestor..."

"Significant findings from data recorded during the proposed project will be promptly submitted for journal publication. Thus, the most important data will be freely available to all, either as part of journal articles or as supplementary material that is available at the journals' websites."

A "by request only" sharing policy requires justification, which could be as simple as a statement that this is standard practice in the communities that have an interest in the data.

Be sure that your use of FTP doesn't limit access to your data. Is the FTP site discoverable on the open web?

If users must register for a CTools account, this will seriously limit the use of the data.

Mailing physical media should be avoided if at all possible; this is a serious barrier to use.

See summary for Section Vb. Most journals are restricted to subscribers, so supplementary material is certainly not available to all. Publication ≠ data sharing.

3. For how long will the PI(s) retain the right to use the data before opening it up for wider use? Explain the reasons for any embargo period.

"Data will be uploaded into the NEES Project Warehouse for this NEESR project and will conform to the NEES Data Sharing and Archiving Policies. These policies also define data confidentiality during the research, but require that data be available to the broader community no later than 12 months after completion of an experiment."

Some common options for data release:

- Immediately after data are gathered;
- After processing, normalization and/or correction;
- After an analysis has been completed;
- Immediately before publication;
- Immediately after the publication of derived results;
- Immediately after funding for the project has expired;
- Within one year after funding has expired.

For projects that are expected to generate a variety of datasets, a timetable describing your planned milestones for data archiving and release is an excellent way to answer this question. See Appendix D for an example.

4. What steps would another researcher need to go through to obtain access to your data? Will any restrictions be placed on the data? For example, will users be required to sign an agreement before using the data? If so, explain the nature of the agreement (link to the full agreement if possible).

"Our software is already and will remain publicly available at the Center for Space Environment Modeling website. Scientists and researchers can obtain the software and documentation after filling in a simple form and user agreement and sending back a signed hard copy to us. The registration and user agreement are legal requirements by the University of Michigan, and this procedure is similar to those used by other similar projects."

This is a good start, but the plan should include more information about the user agreement, preferably with a link to the form.

"External users of our simulation software must sign a registration and user agreement form before obtaining the code. The user agreement includes a detailed description of the limitations on the redistribution and use of the software. In particular, external users cannot redistribute the software, documentation or coding details. The user is also required to limit publications to the topics indicated in the registration form. The detailed text is available at..."

The agreement mentioned here sounds needlessly restrictive, but it has the virtue of being explicit, which is desirable in this case.

"Our policy is that codes and data resulting from this project will be distributed at no cost, but a signed license will be required. There will be two types of license: user/group license and site license. A user/group license will specify that no user or site will redistribute the code or data to a third party outside the group in original or modified form without written permission of the PI from whom it was obtained, except as explicitly permitted in the license.... A user/group license entitles the licensee (a person) and his/her research group and collaborators to use the code or data and share it for use within a single research group. Publications resulting from using the code or data will reference the source as provided in the license; the recommended citation will be provided. A site license is the same as a user/group license except that it enables the licensee (company, organization, or computing center) to install the code or data and allow access to the executable code or data to any number of users at that company, organization, or computing center."

This example spells out the steps another party would need to go through in order to obtain the data.

"Data requestors can only access to the data provided a data use agreement is first executed, in which, limited responsibility from data provided will be described."

This answer is too vague to be informative.

5. Are there any potential proprietary, security, ethical or privacy issues that might require limitations on which of the data to be generated can be shared (for example, human subject data, geolocations of sensitive ecological or cultural resources)? If so, how will these issues be addressed?

See DataONE Best Practices, *Identify data sensitivity:*

https://www.dataone.org/content/identify-data-sensitivity

"Before data is stored, it will be stripped of all institutional and individual identifiers to ensure confidentiality by staff of the Center following procedures developed by the researchers."

"Audio files of interviews will be stored on a password-protected secure server during the study and for two years after, and destroyed subsequently."

"Exceptions to shared data include proprietary DTE GIS utility information (for security reasons) and software code of commercial interest to the project's GOALI partners or identified licensees. Both exceptions are permitted by the ENG DMP policy.... The research team will however develop a set of 3D GIS datasets for distribution the public. These datasets will represent non-existent buried infrastructure and will only be useful for the evaluation of the other research products."

These answers all go into detail about the steps that will be taken to deal with security and privacy issues. The third is notable for providing a means to share a type of data that would normally not be shareable for security reasons.

"There is no ethical and privacy issue. The dataset will not be covered by any copyright."

These statements stand in need of demonstration. Consult with Tech Transfer or DRDA if you are uncertain about privacy or IP issues.

6. If applicable, what procedures will be followed in order to comply with IRB obligations? Be as specific as possible.

"Informed consent statements will use language that will not prohibit sharing of the data with the research community. The following language (or a close variant) will be used in consent statements: The information in this study will be used in ways that will not reveal who you are. You will not be personally identified in reports or publications on the study or in any data files shared with other researchers. Your participation in the study is confidential."

This is ICPSR boilerplate, but it's good language to use in cases where informed consent procedures are truly applicable. If you do use boilerplate text in your plan, be certain that it's relevant to the types of data you plan to create.

"The applied anonymization techniques will conform to the standards set by our Institutional Review Board (IRB) for published data of this type."

More detail about the nature of these standards would be appropriate here.

7. Do you plan on publishing findings that rely on your data? If so, do you anticipate any restrictions on data sharing as a result of publisher policies?

Remember that publishers' agreements are often negotiable, and that sharing your data increases the impact of your research. You may need to review the policies of publishers in your field in order to answer this question.

IVb. Rights and Conditions for Re-Use

"Unless otherwise provided in the grant, all legal rights to tangible property collected or created during NSF-assisted research remain with the grantee or investigators as determined by the policies of the organization... Such incentives do not, however, reduce the responsibility that investigators and organizations have as members of the scientific and engineering community, to make results, data and collections available to other researchers."

http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

Summary: *This set of questions deals with legal issues surrounding the sharing of research data. We strongly encourage you to consult with DRDA [10] and the Office of Technology Transfer [11] if you have any special licensing requirements or intellectual property concerns. The library's Copyright Office is also available for consultations regarding the licensing of software and research data (see Appendix A for contacts).*

These issues are best addressed through the use of a formal legal license to be provided along with your software or data. See Appendix B for an overview of software licensing issues, and Appendix C for instructions on applying a license.

8. Will any group members be claiming intellectual property rights to the data, and if so how will this affect access and reuse? For instance, are there plans to patent any products of the research project that might restrict the sharing of research data as defined by NSF and ENG? Are there other stakeholders (such as academic or corporate partners) who need to be consulted before the data are made available?

Consult with DRDA and/or the Office of Tech Transfer [10] [11] if your project is expected to have any special IP requirements.

"In the event that discoveries or inventions are made in direct connection with the data, access to the data will be granted upon request once appropriate invention disclosures and/or provisional patent filings are made. Data will, in principle, be available for access and sharing not later than two years after the acquisition of the data."

This response clearly states the policy associated with data related to IP claims and provides a timeframe for its release.

"Key data relevant to the discovery will be preserved until all issues of intellectual property are resolved."

"Dissemination of data shall be consistent with decisions regarding the management of intellectual property pertaining to the project."

These answers don't provide any specific information about the project's IP policies.

9. Think about potential uses for your data and what conditions, if any, you will attach to those uses:
Will you permit the re-use of your data?
Will you permit the re-distribution of your data?
Will you permit the creation and publication of derivatives based on your data?
Will you permit others to use the data for commercial purposes?

"After uploading the data into the NEES Project Warehouse and allowing public access, all data will be available for re-use and re-distribution with proper acknowledgement of their originators."

"Researchers and practitioners in diverse fields will be able to readily reuse and redistribute shared data. Terms of use will include the prohibition of commercial commercial use of the work – modifications of the work will be allowed with the proper citations."

"The simulation code will be developed in C and provided to the public in source code format for non-commercial use under GNU General Public License (GPL)."

The use of a defined license is ideal here.

"Re-use or re-distribution of data is allowed with the permission of the PI."

"We anticipate no problems or restrictions on the re-use of our data for the purposes of additional analyses that another researcher might wish to perform. We will impose no limitations on the re-use of the data by a legitimate interested party."

These two examples are laudably permissive, but without an explicit license it is unlikely that the data will receive much use.

"All datasets created during the course of this project and deposited will be licensed under the following Creative Commons license agreement for Attribution Non-Commercial Share-Alike. The full agreement is available at …"

Be careful when selecting a license, as not all are appropriate for datasets. For more information see the guide from JISC [12].

10. If you plan to impose conditions on any of the above uses, is there a preferred means of complying with those conditions? For instance, if attribution is to be a condition, is there a preferred method of attribution?

For data citation best practices see DataCite [13]. Also see DataONE Best Practices, *Provide a citation and document provenance for your dataset:*

https://www.dataone.org/content/provide-citation-and-document-provenance-your-dataset

V. Data Archiving and Preservation

"Long-term large-scale digital archiving requires systems, organizational structures, policies and business models that are robust enough to withstand i) technological progress, ii) failures, iii) changing standards, iv) changes in institutional missions, and v) interruptions in management and funding."

From NSF 04-592, "Digital Archiving and Long-Term Preservation:"

http://www.nsf.gov/pubs/2004/nsf04592/nsf04592.htm

Summary: *The next two sections deal with long-term data archiving, which is distinct from short-term backup and disaster recovery strategies. The former may employ storage solutions, such as lab servers or removable media, that are neither reliable in the long term nor discoverable by outside parties. Once data are ready to be archived and shared they will most likely need to be transferred to a repository or data center with a commitment to long-term curation. Consider both backup and archival strategies as part of your data management planning process, but avoid confusing the two.*

Because ENG makes no explicit distinction between short- and long-term storage in its guidance documents, we have combined these two topics into one section (Vb). Short-term backup and disaster recovery procedures can be described there or in Section II, Question 2 ("How will the data be created or captured?"). Nevertheless, these two topics need to be considered separately to avoid the inappropriate use of storage.

Va. Period of Data Retention

"The DMP should describe the period of data retention. Minimum data retention of research data is three years after conclusion of the award or three years after public release, whichever is later. Public release of data should be at the earliest reasonable time. A reasonable standard of timeliness is to make the data accessible immediately after publication, where submission for publication is also expected to be timely. Exceptions requiring longer retention periods may occur when data supports patents, when questions arise from inquiries or investigations with respect to research, or when a student is involved, requiring data to be retained a timely period after the degree is awarded. Research data that support patents should be retained for the entire term of the patent. Longer retention periods may also be necessary when data represents a large collection that is widely useful to the research community. For example, special circumstances arise from the collection and analysis of large, longitudinal data sets that may require retention for more than three years. Project data-retention and data-sharing policies should account for these needs."

http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

Summary: *This section asks how long your data will be archived after the conclusion of the award. The minimum period is three years as specified by ENG, but you must also take into account any more stringent requirements imposed by the specific solicitation and the PIs' home institutions. Consider who is likely to be interested in using the data and estimate a useful lifespan based on the current pace of research in their field(s). For instance, if methods and instrumentation are evolving rapidly then your data may not be useful to your own research community after five years, but if the state of the art is relatively stable you may need to preserve the data for much longer.*

1. Which of the data you plan to generate will have long-term value to others? For how long?

If you plan to generate more than one dataset or type of dataset, they may need to be archived for different periods of time.

2. How long will data be kept beyond the life of the project?

ENG specifies 3 years, minimum.

"Data will be maintained on our groups' public server for a minimum of three years after the conclusion of the award or public release, whichever comes later. From experience, we expect this period to extend to eight years."

The explicit timeframe is a plus here, but see the next section for an explanation of why a server maintained by your research group may not be the best solution for long-term data storage.

Vb. Archiving and Preservation Strategy

"The DMP should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. In collaborative proposals or proposals involving sub-awards, the lead PI is responsible for assuring data storage and access."

http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

Summary: *This section asks you to provide a strategy for preserving your research data in the long term. Your first choice for long-term data preservation should be a disciplinary repository serving a relevant area of research. If no such repository exists, consider using one of the institutional data archiving options referenced in this section, or use those provided by collaborators' institutions. Perpetual archiving in a curated disciplinary or institutional repository is the preferred solution for long-term data preservation; rolling your own solution should be considered a last resort. If there are no applicable repositories, describe how you will keep the data accessible for its expected useful lifespan.*

Storing data on a lab server, removable media, NAS, etc. is acceptable for temporary storage or backup purposes but does not constitute a reliable long-term archiving solution. Also, be aware that most journals limit access to subscribers and are susceptible to journal or publisher failure. Don't imagine that data archiving requirements are being fulfilled simply because you have published results in a journal or other venue that makes supplementary data available to subscribers.

3. What is your long-term strategy for maintaining, curating, and archiving your data? How will you ensure access beyond the life of the project?

"For archiving, the data along with any related publications will be deposited in Libra, the UVA archival system, with an appropriate licensing statement. DOIs will be attached to all data stored from this project. Since the current preservation plan for Libra is indefinite data storage, preservation of access is assured."

"Materials to be publicly shared will be stored with the Deep Blue repository, a service of the UM Libraries that provides deposit access and preservation services. Deposited items will be assigned a persistent URL that will be registered with the Handle System for assigning, managing, and resolving persistent identifiers ('handles') for digital objects and other Internet resources."

Both of these examples invoke institutional repositories as a long-term data archiving solution. The mention of attached licenses and persistent identifiers is an added strength. See this guide from the Australian National Data Service for more information on persistent identifiers:

http://ands.org.au/guides/persistent-identifiers-working.html

"All data will be available at request immediately after the scientific results are published and will be stored at least another five years on magnetic and optical storage devices (hard disks, CDs, DVDs). Optical storage devices will serve the purpose of the Disaster Recovery Plan."

Optical storage media are not sufficiently reliable for long-term archival storage.

"The data and related publications will be stored in the University of Michigan controlled NAS storage unit to enable future access and sharing. It is anticipated that this data will be stored for a minimum of three years from the date of completion."

Be sure that your NAS unit supports HTTP access before planning to use it for data sharing, and consider using a curated data repository instead if at all possible.

"All important data and descriptions of samples will be published, and relevant data will be included in supplementary information attached to the publications which are accessible to the readers of the journals."

Publishers are not in the business of data archiving and have little incentive to care for your data in the long term. Most journals are accessible only to subscribers.

4. Which archive, repository, or database have you identified as the best place to deposit your data? If there are no appropriate disciplinary repositories, what tools will you use to preserve and disseminate your data and what resources will you need to make use of those tools?

Find a disciplinary data repository: [15], [16], [17].

"Upon finalization, the calving histories we derive from satellite data will be submitted to the National Snow and Ice Data Center (NSDIC) for permanent archival, and will also be converted into .kmz files for display in Google Earth, as part of a broader effort to catalog our results and inform the public about coastal change in the region of Antarctica."

This is an excellent example of choosing an appropriate disciplinary data repository.

"The infrastructure provided by NSDIC is not yet either designed or suitable for distribution of software that evolves and changes over time. Instead, we will release and distribute GPL versions of data analysis and numerical modeling software as tarballs directly from our respective institutional webpages."

Software is one example of data that may not be suited for archival in a repository. A good plan should address how such data will be preserved, although the specific solution mentioned here may be less than ideal.

"At the end of the project, original laboratory notebooks will be secured by the PI in his campus office, and computer files will be stored in the form of hard-drive storage as well as on the University of Michigan CTools websites."

None of the three storage solutions listed here is appropriate for long-term archival storage.

5. What procedures does the repository have in place for preservation and backup? Are there any security measures that need to be taken when storing and distributing the data (e.g. permissions management, restrictions on use?) Who will manage these security procedures and how?

If you are delegating long-term curation to a data archive or institutional repository you may simply want to reference their security procedures and disaster recovery plan here. See also the LSA Data Classification Flowsheet, Disaster Recovery Plan Guide, and Deep Blue Details (for archival storage): [18], [19], [20].

See also DataONE Best Practices, *Create and document a data backup policy* and *Ensure integrity and accessibility when making backups of data:*

https://www.dataone.org/content/create-and-document-data-backup-policy

https://www.dataone.org/content/ensure-integrity-and-accessibility-when-making-backups-data

"We use a mirrored CVS server to store the current as well as all earlier versions of the simulation software.... The CVS repositories are backed up regularly.... Our simulation software is checked out from the CVS server every night, and we run a large test suite on several machines. The results of these tests are sent back to a web site that we check every morning, so we can discover minor or major problems with the software or hardware within a day. The machines running the nightly tests are distributed in the department of AOSS, across the campus of the University of Michigan, and also include a machine in California."

"Data generated at the University of Michigan will be stored in a repository called Deep Blue. Deep Blue offers high-level preservation, security, and compatibility for data stored in a variety of standard file formats, and assures high visibility in most commonly used search engines."

"The data will be stored on the hard disks of computers at the PIs' labs as well as on the PIs' personal computers.... All data will be stored on at least two computers which are regularly backed up using external hard drives as well as optical storage devices (CDs and DVDs). The latter will serve the purpose of the Disaster Recovery plan."

Be sure to keep in mind the distinction between short- and long-term storage here. These may be adequate backup solutions, but they would not serve the purpose of long-term archival storage.

6. What procedures does the repository have in place for forward migration of storage technologies, to avoid obsolescence?

"All data will be available at request immediately after the scientific results are published and will be stored at least another 5 years on magnetic and optical storage devices (hard disks, CDs, DVDs). Optical storage devices will serve the purpose of the Disaster Recovery plan. In case new storage devices become available in this period, the data will be transferred to these."

7. What data preparation, description, or cleaning procedures will be necessary to prepare data for archiving and sharing? (E.g. quality or consistency checks, de-identification, insuring compliance with IRB requirements, obtaining consent from project members or other stakeholders?)

"The data to be collected will not require transformation, cleaning, or anonymization."

If this is true, explain why.

8. What metadata and other documentation will be submitted alongside the data (or created after deposit) to ensure the data can be found and used by others?

You can refer here to the metadata considerations you discussed in Section IIIb. Just be sure to say something about *how* the metadata will be attached to the data you plan to deposit.

"The data and designs will include metadata describing the experimental setup, methodology, data types, and other relevant metadata."

This belongs in Section IIIb, and needs to be fleshed out further.

9. Will any other related information be deposited (e.g. publications, software, reports)?

"The data will include a README file with links to any reports or papers generated using it."

Who will maintain the links when new papers are published?

10. How much will it cost to preserve and disseminate the data and how will these costs be covered?

Include any costs associated with data preservation in the budget justification section of your proposal, but briefly describe them here.

References

1. National Science Foundation. Award and Administration Guide: Section VI.D.4.b - Dissemination and Sharing of Research Results. NSF 11-1 January 2011.
http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

2. National Science Foundation. Grant Proposal Guide: Chapter II - Proposal Preparation Instructions, NSF 11-1 January 2011.
http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

3. National Science Foundation, Directorate for Engineering. Data Management for NSF Engineering Directorate Proposals and Awards.
http://nsf.gov/eng/general/ENG_DMP_Policy.pdf

4. University of Michigan Library. NSF Data Management Plans.
http://www.lib.umich.edu/node/24509/

5. National Science Foundation. Data Management & Sharing Frequently Asked Questions (FAQs) - updated November 30, 2010.
http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp

6. Wotsit.org: The Programmer's File and Data Resource.
http://www.wotsit.org/

7. Library of Congress. Sustainability of Digital Formats: Planning for Library of Congress Collections.
http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml

8. National Archives. Frequently Asked Questions (FAQs) About Selecting Sustainable Formats for Electronic Records.
http://www.archives.gov/records-mgmt/initiatives/sustainable-faq.html

9. University of Michigan Library. Deep Blue FAQ.
http://deepblue.lib.umich.edu/about/deepbluefaq.jsp

10. University of Michigan Division of Research Development and Administration. Data Sharing Resource Center.
http://www.drda.umich.edu/datasharing/

11. University of Michigan Office of Technology Transfer. U-M Policies Overview.
http://www.techtransfer.umich.edu/resources/policies_overview.php

12. Naomi Korn and Professor Charles Oppenheim. Licensing Open Data: A Practical Guide. Version 2.0, June 2011.
http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

13. DataCite. DataCite: Helping you to find, access, and reuse research data.
http://datacite.org

14. National Science Foundation. NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. September 2005.
http://www.nsf.gov/pubs/2005/nsb0540/start.jsp

15. DataCite. Repositories.
http://datacite.org/repolist

16. Open Access Repository. Data repositories.
http://oad.simmons.edu/oadwiki/Data_repositories

17. Purdue University Libraries. Other Data repositories.
http://d2c2.lib.purdue.edu/OtherRepositories.php

18. University of Michigan College of Literature, Science, and the Arts. Data Classification Flowsheet. Version 5.
http://webapps.lsa.umich.edu/lsait/admin/Data_classification_flowsheet_v5.doc

19. University of Michigan College of Literature, Science, and the Arts. Disaster Recovery Plan Guide.
http://webapps.lsa.umich.edu/lsait/admin/LSA-DRP-Guide.pdf

20. University of Michigan College of Literature, Science, and the Arts. LSA.Security's Deep Blue Details (For Archival Purposes).
http://webapps.lsa.umich.edu/lsait/admin/Deep_Blue.html

21. Open Knowledge Foundation. Panton Principles: Principles for Open Data in Science.
http://pantonprinciples.org/

22. Open Knowledge Foundation. Open Definition: Conformant Data Licenses.
http://www.opendefinition.org/licenses/#Data

23. The University of Michigan. Standard Practice Guide 601.12: Institutional Data Resource Management Policy. Revised 10/4/2008.
http://spg.umich.edu/pdf/601.12.pdf

24. The University of Michigan. Data Administration Guidelines for Institutional Data Resources. October 2008.
http://www.mais.umich.edu/access/download/daguide.pdf

25. The University of Michigan. Data Stewards/Data Managers List.
http://www.mais.umich.edu/access/download/data_stewards.doc

26. University of Michigan Information and Technology Services. Storage Services Side-by-Side.
http://www.itcs.umich.edu/storage/compare.php

27. University of Michigan College of Engineering, CAEN Advanced Computing. Flux.
http://cac.engin.umich.edu/resources/systems/flux/index.html

28. DataONE. DataONEpedia.
https://www.dataone.org/dataonepedia

29. James Brunt. How to Write a Data Management Plan for a National Science Foundation (NSF) Proposal. February 18, 2011.
http://intranet2.lternet.edu/node/3248

30. Sarah Jones. How to develop a data management and sharing plan.
http://www.dcc.ac.uk/resources/how-guides/develop-data-plan

Appendix A: DMP Contacts at the University of Michigan

Topic	Name	Email
Data sharing and compliance	Alex Kanous (DRDA, Data Sharing Resource Center)	akanous@umich.edu
Copyright; software and data licensing	Doug Hockstad (Tech Transfer) or U-M Library Copyright Office	dhocksta@umich.edu copyright@umich.edu
Storage options; data formats	Paul Killey	paul@umich.edu
Disciplinary repositories; data and metadata standards; additional DMP information	Jacob Glenn (U-M Library) or Your subject librarian	jkglenn@umich.edu http://www.lib.umich.edu/subject-specialists
Deep Blue (U-M's institutional repository)	Jim Ottaviani	deepblue@umich.edu

Appendix B: Software Licensing

Please note: *The guidelines below are offered for illustrative purposes only; they are not endorsed by the University of Michigan. Contact the Office of Technology Transfer or the library's Copyright Office for help with software and data licensing issues.*

Copyright and software – The Basics

In the United States most software is subject to copyright law. This applies to both the human readable ("source code") and the machine readable ("object code") versions. Thus, it is important to think about copyright and licensing when releasing or sharing a new software project or a project that may include code.

Copyright

At the University of Michigan, except in certain defined situations (please see The University of Michigan Copyright Policy for detail), faculty members hold the rights for any work they create. Thus, the responsibility falls upon that faculty member to determine whether and how any software that they have created and intend to release will be licensed. A rights holder is not necessarily required to place a license on released software, but depending on its nature or the objective of the rights holder a license is often recommended. A variety of licensing options are addressed below.

Remember, the party that actually holds the copyright for a work can vary due to a number of reasons, as outlined in the University copyright policy. These can include contractual terms of a grant or grants, substantiveness of University resources used in the creation of the work, or whether the work was created in response to a direct assignment or duty of employment.

The usual copyright rules apply to works of joint authorship (between two or more individuals) where the copyright is held by all equally. In those situations, it is best to discuss these issues early on in the process so that all copyright holders are satisfied with the decision.

Copyright Registration

Just like any other copyrightable work the author (or copyright holder) has the option of registering the work with the United States Copyright Office. There are two main methods of registration: one for when the source code does not contain trade secrets and one when it does. For more information on the process of registration of computer programs with the US Copyright Office, see Circular 61. Registration is not required to benefit from copyright protection in the US.

Licensing Options

Proprietary or Free/Open Source

One of the initial questions that must be answered when deciding how to license a software project is whether or not the code is to be Open Source/Free Software versus proprietary software. The difference can be summarized as simply whether or not you want others to be able to see and share the code. If the choice is made to not share the source code with others then registration with the US Copyright Office (see above) and notice of copyright is sufficient. It is recommended to put a copyright notice, such as "© 2010 by Jane Doe," on any splash screen, about dialog, or physical media that contains the software.

If the intent is to realize revenue through the distribution of the software, a proprietary licensing scheme is recommended and the University of Michigan Office of Technology Transfer should be consulted for assistance.

The license chosen will often depend on the social norms of the research community the rights holder operates in. If standard licenses exist within that community it is often recommended that such licenses be used unless the rights holder has particular objections or other preferences.

Copyleft or Permissive

At this point there is another choice: "copyleft" or "permissive." By copyleft we mean a license that ensures not only that the original program is made available under a Free/Open license, but also that any derivatives, modifications, and/or forks are as well. Copyleft, in this context, is also sometimes referred to as "reciprocal" or "viral."

An example:

Sarah is an academic who writes a piece of software, **MapGenerator**, and releases it to the world under the terms of the GNU General Public License version 3.0 (GPLv3). Later, **Sam** incorporates **MapGenerator** (with or without modification) into their project, **MapsForGeologists**, and released it to the world in a proprietary format (ie: the source code is not viewable). As soon as any person receives **MapsForGeologists** from **Sam**, that person is allowed to request the full source code of **MapsForGeologists** per the terms of the GPL. If **Sam** does not disclose the full source code then **Sarah** (the copyright holder of **MapGenerator**) can bring suit against **Sam** for violation of the license. This very issue (concerning the Artistic License instead of the GPL) was recently litigated in the case *Jacobson v Katzer*.

If, however, the license on the original software project was released under a permissive license, for example the Apache or MIT license, then the actions of **Sam** would be allowed (permissible).

Recommended Licenses

The recommended permissive software licenses are:

- Apache 2.0
    - This license allows others to reuse the software source code in any way including incorporating the source into proprietary projects.
    - It does require that any copyright notices stay in tact.
    - It also provides licenses to any patents held by the software authors to the public under the same terms.

The recommended copyleft software licenses are:

- GNU General Public License version 3.0
    - Latest version of the GPL which ensures that the original program and any derivatives of the original, including when other programs link to it, are released under the same license. Includes new provisions around patent and "tivoization."
    - A quick guide to the GPL: http://www.gnu.org/licenses/quick-guide-gplv3.html
- GNU Affero General Public License version 3.0
    - A one-paragraph modification to the GPL which requires the source code to any program that a user interacts with over a network be made available to those users.
    - Why the Affero GPL: http://www.gnu.org/licenses/why-affero-gpl.html
- GNU Lesser General Public License version 3.0
    - The less viral version of the GPL which allows other programs to programmatically link to it without forcing themselves being GPL. The license still ensures that any modifications of the LGPL'd software are also LGPL'd.

Applying a License

After a license has been chosen for a project, the following few steps should be taken to apply the license to the work:

1. At the top of every source file in the project, include a copyright statement as a comment. An example: "Copyright 2010, Regents of the University of Michigan" if there was only one author and one release of the project.
2. If there were multiple authors and multiple releases of the project add a copyright statement for each author who has contributed to a specific file and the year in which they did.
3. At the top of every source file in the project include, below the copyright statement, a short statement describing which license the code is released under. For example:

```
#   Copyright [yyyy] [name of copyright holder]
#
#   Licensed under the Apache License, Version 2.0 (the "License");
#   you may not use this file except in compliance with the License.
#   You may obtain a copy of the License at
#
#       http://www.apache.org/licenses/LICENSE-2.0
#
#   Unless required by applicable law or agreed to in writing, software
#   distributed under the License is distributed on an "AS IS" BASIS,
#   WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
#   See the License for the specific language governing permissions and
#   limitations under the License.
```

4. In the root folder of the project include a COPYING or LICENSE file that is a copy of the entire license in plaintext.

External resources

The decision as to the appropriate license for a particular work is not only often a practical choice, but, especially in the case of copyleft or permissive licenses, also philosophical. As such, the MLibrary Copyright Office does not recommend one type over the other. However, there are some useful guides online that can help one think through the implications and outcomes of the choice. These are just a few:

- "Choosing an Open Source License" – Chapter 10 of Open Source Licensing Software Freedom and Intellectual Property Law by Lawrence Rosen.
- "Choosing an Open Source License" – by Ian Lance Taylor
- "Comparison of free software licenses" on Wikipedia
    - Outlines the similarities and differences between the majority of in-use free/open source software licenses.

More Information

If you have any questions regarding the licensing of software feel free to contact the Library Copyright Office at copyright@umich.edu.

Appendix C: Applying a License to your Software

Step 0: Pick a license and insure that all code (and other) contributions can be made available under the terms of that license. See <url of license guide> for more details and suggestions.

Step 1: Create a LICENSE file in the base directory of the project source code. Copy the entire text of the license to that file in plain text. Plain text versions of all free and open source licenses can be found at http://opensource.org/licenses/alphabetical.

Step 2: If you have a README file, reference the chosen license, for example:
"Foobar is licensed under the terms of the Apache 2.0 license. See the LICENSE file for more details."

Step 3: It is a best practice to add a copyright/license notice within the header of every source file in the project. In the case of the Apache 2.0 license, you can add the following block of text before the other comments describing the function of the file:

```
#  Copyright [yyyy] [name of copyright holder]
#
#  Licensed under the Apache License, Version 2.0 (the "License");
#  you may not use this file except in compliance with the License.
#  You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
#  Unless required by applicable law or agreed to in writing, software
#  distributed under the License is distributed on an "AS IS" BASIS,
#  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
#  See the License for the specific language governing permissions and
#  limitations under the License.
```

For the GNU GPL:

```
#   This file is part of Foobar.
#
#   Foobar is free software: you can redistribute it and/or modify
#   it under the terms of the GNU General Public License as published by
#   the Free Software Foundation, either version 3 of the License, or
#   (at your option) any later version.
#
#   Foobar is distributed in the hope that it will be useful,
#   but WITHOUT ANY WARRANTY; without even the implied warranty of
#   MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
#   GNU General Public License for more details.
#
#   You should have received a copy of the GNU General Public License
#   along with Foobar.  If not, see <http://www.gnu.org/licenses/>.
```

NOTE: When using the Lesser GPL, insert the word "Lesser" before "General" in all three places. When using the GNU AGPL, insert the word "Affero" before "General" in all three places.

If you want to use the Creative Commons Zero Waiver for your source code, see this useful blog post from Creative Commons: http://creativecommons.org/weblog/entry/27094

Appendix D: A Data Sharing and Archiving Timetable

Archiving Schedule with Milestone Dates

Data Resource	Date of Milestone			
	Production	Archiving		Curation and Public Release
		Repository	NEEShub Project Warehouse	
Unprocessed+ Cornell test 1 Cornell test 2 Cornell test 3	 June/July 2012 June/July 2013 June/July 2014	 July 2012 July 2013 July 2014	 January 2013 January 2014 January 2015	
Structured with metadata and documentation* Cornell test 1 Cornell test 2 Cornell test 3			 Feb/March 2013 Feb/March 2014 Feb/March 2015	After publication, no later than: June 2013 June 2014 June 2015
Encapsulation and Learning Modules			Spring 2014 Spring 2015	September 2015

+ Will be stored and made available to participants as recorded.

* Structured data will be delivered to the repository within 6 months of the end of each experiment.