The Data Management Plan and Project Repository

A Data Management Plan will benefit from having a project repository, (a kind of manual for the entire project). With a project repository, all the active and temporary members of the project have a single place to refer to for protocols, expectations and methods, especially with regard to the resulting data.

In the setup of a project repository, an initial decision needs to be made on where to store this information: a wiki? Course Management site (ie: Moodle; which could include a scheduler for organizing access to the equipment and an embedded calendar for specific events), a web site? Or less-desirably; a simple networked folder of files, google group, google docs folder.

The project repository should include details on all aspects of the project. It is a repository which will store materials and documentation as they are accumulated related to the development of the project. This will help ensure continuity over time as well as across new and existing collaborators. All lead members of the group must work to keep the repository accurate and up to date. Below are some categories which could/should be covered in a project repository:

> Project Resources (Equipment): Records relating to equipment history, repairs, questions, manuals, history of who used it, doing what & when. Scheduling for use of the equipment, training and other documentation on how to use the specialty equipment (specific scientific devices, desktop computer/s, mini-computer/s, attached storage devices, [etc?]) Where & what are the warranties for hardware? What happens to the equipment at its end-of-life?
>
> Project Resources (Human): It is also helpful to have the explicit role expectations for each member of the project laid out in the project repository. It could also define what to do if/when someone needs to stop working on the project or is added to the project while underway; What file and/or server rights or accounts need to be managed and in which ways? This should include anyone from student workers through co-PIs.
>
> Who is responsible for managing the equipment repair process? Be clear about the expected response time for help/support/fixes and make sure these people or offices know that you may require their assistance.
>
> Project Resource Scheduling: Be sure to allow time for routine, one time or surprise events such as; Repairs, maintenance, testing, student training for researchers, student training for course-related work, creating the related documentation for training, and basic scheduling of the equipment itself. It's very helpful to create a plan or method for many different aspects of the project in advance of needing them. This is particularly useful with considering how to minimize down-time for unscheduled repairs. Having easy access to manuals and warranty information can make a great deal of difference

to the speed of repair. This may mean asking for the information when the item is purchased and documenting the process directly in the project repository - before it is needed.

Research Data (Lifecycle) protocols: How will records be managed and who will be responsible regarding what has happened to the the meta-data and the data throughout the project? Consider if, where and how the resulting (presumably the most useful) data sets will be documented and archived? Make sure its noted who is responsible for this decision. This should either be a specific person by name or if it's too far in advance, identified by role (eg: the PI is responsible for choosing the appropriate national archival location for the conclusion of this research project.) (*See additional DMP section below.*)

Project Related Outcomes: Grant reports, citations for any resulting publications or presentations, future research ideas. Provide a specific citation for using data from this project. {See, for example; Micah Altman and Gary King, 2007, "A Proposed Standard for the Scholarly Citation of Quantitative Data", D-Lib Magazine, Vol. 13, No. 3/4 (March/ April). doi:10.1045/march2007-altman copy at http://gking.harvard.edu/files/abs/cite-abs.shtml}

Intellectual Property: Since projecs may include external collaborators and temporary (interns or post-docs) researches, it is a good idea to include a brief section clarifying who owns what aspects of the data, information or equipment. Include a creative commons license whenever possible. (http://creativecommons.org/) In addition, verify whether any aspect of the project deals with copyright material belonging to someone else.

Funding Arrangements: Include a copy of the grant application and final approval. Be explicit about who is responsible for which aspects/outcomes/obligations as related to the funding sources. This is a good place to keep track of the budget/s as well.

Include, for example, NSF expectations or guidelines regarding funding acknowledgements in related publications, what the expectations are regarding sharing the research data, and rules for what the funding may be used for.

With a project repository comes an infrastructure for the data documentation portion of the project. It is this sections which may require much more attention than is usually anticipated.

The Data Management Plan (DMP) typically refers to the following general categories:

Research and Data Protocols

These are the explicit processes, protocols, or standard operating procedures for the management of data and meta-data with these devices and among the participants.

The data protocols should include a glossary of relevant terms for what is meant by each elementary component of the project in this context. In particular: research data, meta-data, backup, archive, sharing, cleaning, analysis (etc.) While this may seem remarkably elementary, it is common when working across disciplines for people to have different unexamined semantic norms for basic components. Clarity in these areas is essential to good communication. It also models good communication for junior (and often temporary) research assistants by making explicit the framework into which the research data should be managed.

It should be explicitly defined as who is responsible for authoring the protocol section. In some cases, co-PI sign-off may be indicated. These protocols should specify the data collection, *meta-data design, storage and access protocols, in addition to explicitly outlining the data life-cycle expectations of data resulting from each of the 3 layers of access: PI created research data, student-created research data, and data used in course-related work.

*Data in an of themselves are typically useless. They require meta-data for interpretation. The quality of the meta-data help to define the overall quality of the project outcomes and the longevity of the resulting data.

Data register
This is a kind of catalog of the files, their locations, managers and owners. To keep from acquiring and managing excessive files, it is helpful to the progress of research to identify important files in a kind of register or just a simple table. Here are some suggested categories for a data register:

- Description of the data files or records (ie: their format - whether in reference to individual variables or in reference to a set of files.)
- Quantity and disk space (if recording a category of files)
- Location
- Date stored
- Restricted/confidential?
- “Delete after” date or event
- Intended use
- Person responsible for the file/s

Data life-cycle issues:
What is the anticipated state for each stage of the various data components? These data components are likely to include all of the following, though the process may be slightly different for data resulting from the work from each category of researcher (PIs, Students, Students in courses):

> Data capture: Exactly where does it go? in which format/s? for which participant? (will it/ could it be different for each participant?)
>
> Data processing & Maintenance: Identify the naming conventions, backup & retention policies (see also; preservation), the process by which meta-data are created (or

captured), identify what's known of what must be done to process the raw data for analysis. Be overt about data file format expectations to assure interoperability of files among all participants and their systems.

Data preservation: An ongoing aspect of data management is data preservation. The balance of security with ease-of-use must be weighed for each stage that the data go through. This is about data backups and data archives. A *backup* is a copy of the most recent work. An *archive* is a snapshot of the data at a specific point in time. The archive version is often considered the final version of a data set.

- How do you make sure the data are safe from damage (unintentional and intentional)? What are the possible consequences at which tolerance level for each phase?
- How soon and to what degree do backups need to be conducted? As soon as it's captured? When it's named and transferred to its initial storage location? At each stage until the data are in an analyzable format?
- What does it mean to have it "backed up"? (eg: In the same computer but a duplicate file? To an external storage device and the internal computer drive? To a raided drive location only?)
- At what point may the backup data be considered redundant enough to overwrite?
- Should the original raw format be archived in addition to a processed version?
- Data may need to be transformed out of obsolete formats to ensure persistence. This may also be true for some backed-up files as well as copies in the archive. Decide whether archived previous versions need to be retained and for how long. (Note that a similar procedure should exist for meta-data schema updates.)
- Make a plan for identifying the end-of-life for the project and its data and meta-data.
- Pay special attention to any data requiring privacy controls. For instance, any meta-data which identify participants with their resulting associated experimental data recordings. Where are sensitive to be stored? When appropriate, to what degree of security should they be destroyed?

A part of preservation is assuring that non-useful data are destroyed. Define protocols for how data are identified for destruction, evaluated and finally destroyed or sanitised for re-use (include retention periods and how they are enforced.)

Decide whether or at what point these data may be appropriate for a local research data repository and/or whether they should be donated to a national repository such as ICPSR (http://www.icpsr.org). Note that grant guidelines often stipulate that data resulting from funded research must be made generally accessible, donating your research data to an archive is an excellent way to comply with these rules without incurring additional long-term data management responsibilities.

Roles and Responsibilities: Identify the probable roles for PI, co-PIs, IT systems &

instruments specialist, research data specialist, post-doc or intern, student researchers, students enrolled in courses. Make sure that anyone named but not actively in the project is aware of the project and has access to read relevant portions of the research repository.

The Data Management Plan must be updated to reflect who is responsible for all transitions, such as new data ownership, additional PIs and who is responsible for keeping track of the various copies. (eg: Will a new PI inherit previous experimental data?)

Additional points to consider:

- How will particularly rich data files (either for research or teaching purposes) be identified and retained? Apart from the overall data resources or in line with the rest but in some way specially identified?
- If any publications resulted from analysis of these data, then it is appropriate to assume that the articles & their related data files should be preserved in an institutional repository (either at your institution or at a national archive such as http://www.icpsr.org.)
- A portion of the project funding should be set aside to simply manage the project and its data. In particular, plan for the funding of the transition project data to archival quality research records. This will include substantial project and key data element documentation as well as preservation data formats.

Technical and Infrastructure Requirements

Verify your plans with your appropriate information technology staff. This may relate to network infrastructure access, additional hardware, potential electrical power needs, additions to a shared backup service and expectations for equipment end-of-life treatment. For instance, consider the following:

- What will happen to the equipment (eg: the computers) when you are through with them?
- Who is responsible for keeping the hardware running?
- Are you expecting to have any special (ie; non-ITS managed) systems connected to the institutional networks? If not, to what degree shall the equipment software or firmware be kept up-to-date? If a firmware or software update is required but the equipment is not on a network attached to the internet, how will that be managed?
- Include, if possible, even a rough schedule for an infrastructure or technical review of the equipment. This may be in conjunction with planned maintenance.