Crafting the Data Management Plan

NSF TUES and CAREERS Workshops
Spring and Summer, 2012
Today’s session:

- What is a DMP?
- Guiding questions to answer for your DMP
- Data preservation and strategies for sharing your data
- Sample DMPs
What is a Data Management Plan?

- Brief description of how you will comply with funder’s data sharing policy
- Typically reviewed as part of the grant application
- Requirements vary by funding agency and NSF directorate
Why data management? Why now?

- Conform to the Jan 2011 NSF guideline on the dissemination and sharing of research results
- Enable others to access/use data from funded projects
- Create research networks
- Move information from lab to the community or commerce more quickly
- Transparency/quality/credibility
Other outcomes of DMPs

• Simplify requests for data
• Increase visibility and impact of your research
• Clearly document and provide evidence for your research in conjunction with published results
• Comply with copyright and ethical compliance (ie. HIPAA).
• Preserve data to safeguard from loss
• Support open access
• Facilitate new discoveries
What does the NSF say?

- “Proposals must include a supplementary document of no more than two pages…”
- “…should describe how the proposal will conform to NSF policy on the [dissemination and sharing of research results](http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp)”
- “Fastlane will not permit submission of a proposal that is missing a Data Management Plan.”
- “…may include only the statement that no detailed plan is needed, as long as the statement is accompanied by a clear justification”
- “…will be reviewed as an integral part of the proposal”

DMP templates and guidance

• DMPTool [https://dmp.cdlib.org/](https://dmp.cdlib.org/)
  – Create ready-to-use data management plans for specific funding agencies with step-by-step instructions and guidance for data management plan

• SciDaC [templates](https://scidac.ornl.gov/)
  – Questions to consider

• Directorate [guidance](https://dmp.cdlib.org/)

Data management guide

http://guides.ucf.edu/data
What to include

• **Types of data**, samples, and other materials to be produced in the course of the project. Retention period?

• **Data and metadata standards** for format and content

• **Access and sharing policies** including provisions for appropriate protection of privacy or other rights

• **Re-use policies** and provisions re-distribution and the production of derivatives

• **Archiving plans** for data, samples, and other research products, and for preservation of access to them.
DMP Examples

Data Management Plan. 1. Products of the Research. The data generated from this project primarily includes the structural characterization of new compositions of matter, the kinetic data and the microscopy information (Aims 1-3). All of these will be translated into numerical/digital data for dissemination in the form of publications and presentations. For example, the polyamine derivatives and reagents will be purified and characterized by \(^1H\) and \(^13C\) NMR, mass spectrometry, and elemental analysis. In addition, scans of the NMR spectra for each new compound will be stored as pdf files. These will be stored on a lab server for inclusion as future Supporting Information in publications. In this manner, the lab students can view the spectral data in addition to reading its numeric translation (e.g., δ 2.25 (s, 3H)) described in each student’s digital notebook. The raw kinetic data obtained during the silica condensation experiments in Aim 3 will be collected in an Excel spreadsheet and stored for future review, if needed. Interested parties will also have access to all the relevant experimental data upon publication via the Supporting Information section posted online by the publisher.

The newly synthesized compounds are also entered into a structure-searchable database created via the ChemBioFinder software package (CambridgeSoft). This database is available to the PI’s lab personnel to help them find each synthesized compound in the lab freezers and refrigerators. Labeled plastic boxes contain water-resistant labeled vials, which house each compound as a pure solid or liquid. Note: a complete chemical inventory of the laboratory is stored as an Excel spreadsheet on the central lab server and is updated upon the purchase of new chemicals, etc. In this regard, the PI has complete inventories of all commercial and custom chemicals within the laboratory space.

A structure-activity relationship relating platform geometries (dihedral angles which orient the amide carbonyl groups), rate of assembly and final morphology will be developed to provide a predictive model for future siaffin mimic designs.

2. Data Format. The data will be stored in each researcher’s digital notebook folder housed on the PI’s central lab server as primarily Microsoft Word (.doc) and Excel (.xls) files and Adobe Acrobat (.pdf) files. Microscopy images are stored as .tiff and .jpg formats. In this regard the data is directly compiled into a digital format which facilitates its dissemination. Since the PI's group publishes full papers with complete characterization data and extensive Supporting Information included, interested parties can easily obtain this information online after publication.

3. Access to Data. The data will be stored in each researcher’s digital notebook folder housed on the PI’s central lab server as primarily Microsoft Word (.doc) and Excel (.xls) files and Adobe Acrobat (.pdf) files. Microscopy images are stored as .tiff and .jpg formats. In this regard the data is directly compiled into a digital format which facilitates its dissemination. Since the PI's group publishes full papers with complete characterization data and extensive Supporting Information included, interested parties can easily obtain this information online after publication. The newly synthesized compounds are also entered into a structure-searchable database created via the ChemBioFinder software package (CambridgeSoft). This database is available to the PI’s lab personnel to help them find each synthesized compound in the lab freezers and refrigerators. Labeled plastic boxes contain water-resistant labeled vials, which house each compound as a pure solid or liquid. Note: a complete chemical inventory of the laboratory is stored as an Excel spreadsheet on the central lab server and is updated upon the purchase of new chemicals, etc. In this regard, the PI has complete inventories of all commercial and custom chemicals within the laboratory space.

A structure-activity relationship relating platform geometries (dihedral angles which orient the amide carbonyl groups), rate of assembly and final morphology will be developed to provide a predictive model for future siaffin mimic designs.

2. Data Format. The data will be stored in each researcher’s digital notebook folder housed on the PI’s central lab server as primarily Microsoft Word (.doc) and Excel (.xls) files and Adobe Acrobat (.pdf) files. Microscopy images are stored as .tiff and .jpg formats. In this regard the data is directly compiled into a digital format which facilitates its dissemination. Since the PI's group publishes full papers with complete characterization data and extensive Supporting Information included, interested parties can easily obtain this information online after publication.

3. Access to Data. The PI and UCF maintain three UCF websites which detail the PI’s research interests in each of his affiliated departments/schools. The websites are associated with the UCF Department of Chemistry (courtesy appointment), Burnett School of Biomedical Sciences (member of the PhD program) and College of Medicine’s (COM) Department of Medical Education (full time faculty member). These sites provide brief summaries of the PI’s area of expertise as well as links to his publications. In terms of future publications and associated Supporting Information, the research team plans to provide experimental details and data in the form of scanned NMR data (pdf files) and raw kinetic data in the form of tables or graphs. Data is not released prior to disclosure to UCF or publication. The PI’s policy has been to rapidly publish as soon as the project is complete and to post links to his latest papers after they appear online from the publisher.

b) Data Sharing. As a biochemistry group developing new technologies, the PI is required by UCF to disclose all new compositions of matter and processes to the UCF Office of Research and Commercialization (UCF-ORC) prior to publication or public disclosure in order to protect the University’s (and PI’s) intellectual property rights to the data generated. These disclosures are reviewed by the UCF-ORC and often patent applications are filed on behalf of UCF. Prior to sending new compounds or probes to other researchers worldwide, UCF-ORC and the interested party complete a Material Transfer Agreement (MTA). If UCF-ORC elects to pursue a patent on the new compound or technology, then a Non-Disclosure Agreement (NDA) or Confidentiality Agreement (CA) are completed prior to shipment to the interested party. These agreements fully delineate the expectations of both parties and the legal responsibility of the PI, UCF and its new partner regarding publications, patents and disclosures. The primary data is stored on the password-protected lab server for additional security. Summaries of the work will be published in COMMUNIQUE, a COM electronic magazine for central Florida outreach.

4. Policies for Re-Use, Re-Distribution and Production of Derivatives. The policy regarding the use of data provided via general access is that no data will be disclosed to the public prior to disclosure and approval for disclosure by UCF-ORC. This is to protect the intellectual property of UCF and the PI. In this regard, there are no posted disclaimers or conditions placed on the posted PI’s website data, since the posted information describes already-published information. There are no restrictions placed on the publicly disclosed data in terms of derivatives.

5. Archiving of Data. In terms of the PI’s laboratory, all research information is housed on a central lab server (Black Armor) located in the student office. This involves a RAID backup system using a redundant hard drive system (1 TB). A third copy of the data is backed-up on a remote hard drive located in the adjacent PI’s office. In this regard, the PI has three copies of the lab’s information backed up nightly. The backup drive in the PI’s office is an important additional repository to help protect the information from a localized lab fire or minor water leaks, damage etc. The lab backup systems will be updated by the PI as computer technology/standards change and the system becomes obsolete or non-functional. The PI has the lead responsibility for maintaining the lab database. Since the backup involves automated processes using the Nero software, there is limited maintenance required beyond routine checks of backup scheduling and purging of old incremental archives, etc.

Lastly, the ultimate archive of the PI’s data is in the form of each publication and its Supporting Information section online. In organic chemistry one typically looks in the Supporting Information associated with the pertinent publication for experimental details. Indeed, having the experimental data linked to the publication at the publisher’s website (e.g., pubs.acs.org) makes for a more efficient search process. In this regard, the PI elects to continue to have a full disclosure of experimental details upon publication in the form of a detailed Supporting Information component to his publications. The PI’s significant publication record of full papers in journals, which have a Supporting Information component (e.g., J. Med. Chem. and J. Org. Chem.), is consistent with this choice. In summary, the PI has a well-organized laboratory space and a clear plan for data acquisition, data archiving and dissemination. The combination of rapid full disclosure in the form of full papers with extensive Supporting Information included makes this an efficient plan for data management.
DMP Examples

Sample Data Management Plan: Physical Sciences and Engineering
From Rice University

1. Products of the Research
The data obtained during the proposed project will consist of measurement records of electronic and optical properties of the nanodevices, as described in the main body of the proposal. These records will consist chiefly of measurements of current as a function of voltage, conduction as a function of time, and Raman spectra as a function of wavenumber obtained via custom software. These data will be recorded via computerized data acquisition software, with essential metadata present either as header in the relevant electronic files, or included along with the indexed laboratory notebook narrative.

These data will provide an experimental look at the detailed dissipative processes at work in nanoscale junctions, as delineated in the main body of the proposal. As such, they will be of interest to the nanoelectronics community, as well as to the condensed matter physics and physical chemistry communities.

2. Data format
The electronic transport and optical spectra will be computer files generally in the form of tab-delimited numbers with header information. These computer files will be accompanied by dated laboratory notebooks, as well as by numerical data files analyzed using MATLAB™ and commercial plotting software, such as OriginLab™. A copy of the custom Raman spectrometer data acquisition software also will be included, since it is essential for interpreting the raw data files.

3. Access to data, data sharing practices and policies
The electronic data will be preserved in multiple on-site backups in the form of DVDs and RAID hard drive storage. Copies of the electronic data will be preserved off-site at Rice University’s storage facility. Original laboratory notebooks will be secured by the PI in his campus office or laboratory. If requested, access to the data will be provided via contact with the PI. Data will, in principle, be available for access and sharing as soon as is reasonably possible, and not longer than two years after the acquisition of the data. The data will be preserved for at least three years beyond the award period, as required by NSF guidelines.

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

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

Data Management Plan Example
From UCSD

The data associated with the research project will be systematically managed. The team has multiple backup servers to protect our research findings, and publicly available internet resources to share our findings. All aspects of the research will be carefully tracked, stored, and published. The work detailed in the preceding proposal can be anticipated to produce three broad categories of data: computer software, physical devices, and optical characterization measurements. The computer software category includes not only the optimization subroutines that will be produced, but also the scripting written to control the laboratory equipment used in the characterization of the fabricated photonic circuitry. The physical devices category includes the unit cell, as well as the prototype optical circuitry generated during its development. The optical characterization category includes spectral measurements of the fabricated devices, as well as the associated metadata.

The algorithm development progression will be logged through both handwritten research notebooks as well as digitally generated documents. To ensure the safety of the data, we will use the UCSD research group’s existing server to periodically backup the materials. A Structured Query Language (SQL) database will be created to track the digital documents. The completed design toolbox will be made available by enacting the Release to Public policies of University of California, San Diego. We plan to package our algorithm in a MATLAB toolbox as well as a self-contained software complete with an user interface.

All of the computer code produced during the project will be written using the latest version of MATLAB. Code will be developed using volume shadow copy technology, which will allow the recovery of prior iterations for quality control. The spectral measurement data will be written to MATLAB cell arrays that also include the measurement data and time, duration, and a unique identification number. Subroutines will be produced that allow the effortless extraction of this data both graphically, and in raw matrix form.

The results of the research performed under this proposal will be disseminated primarily through publication in research journals and conference presentations. All of the computer software and optical characterization measurements will be available to interested parties upon request, and will be transmitted electronically via e-mail.

All electronic data generated by proposal research will be redundantly archived. Locally, the laboratory has a secure server on which all information is stored. The server hard drives are set up in a RAID that is capable of full recovery even in the case of multiple simultaneous disk failure. Additionally, the server drives are backed up on an independent server operated by the electrical engineering department. This will allow full recovery of data in the event of catastrophic failure of the local laboratory server. Physical samples will be labeled and stored in a designated area of the measurement laboratory. All of these systems will be in place for the 3 year minimum prescribed by the guidelines.
DMP resources

- Data Management (or DMP) Research Guide: http://guides.ucf.edu/data
- DMP Tool: https://dmp.cdlib.org/
- Databib: http://databib.org/

Meet with a librarian for an in-depth one on one consultation specific to your research topic.
Questions?

• Lee Dotson, Digital Initiatives Librarian, Lee.Dotson@ucf.edu
• Penny Beile, Interim Head, Reference Services, pbeile@ucf.edu
• Selma Jaskowski, Assistant Director, Library Systems & Technology, selmaj@ucf.edu
• Rich Gause, Government Information Librarian, rich.gause@ucf.edu
• Athena Hoeppner, Electronic Services Librarian, athena@ucf.edu