Origin and Development of Johns Hopkins Data Management Services

David S. Fearon
Johns Hopkins University

Follow this and additional works at: http://escholarship.umassmed.edu/escience_symposium

Part of the Scholarly Communication Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License.
Origin and Development of Johns Hopkins Data Management Services

David S. Fearon

NE Area Librarian e-Science Symposium
April 9, 2015

datamanagement@jhu.edu
http://dmp.data.jhu.edu/
Origin story

- Launched July 2011
- Impetus: NSF Data Management Plan requirement
- Nascent data archive from the Data Conservancy
- Matured to adolescence as a new area of research library service
## Services & Staffing:

### What we do

- Assistance With Preparing Data Management Plans & Data Management Consulting
- Data Management Training
- Workshops
- Archiving and Sharing of Research Data

### Staffing

<table>
<thead>
<tr>
<th>Staffing</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 data management consultants (4th in July)</td>
</tr>
<tr>
<td>½ Data Curation Specialist</td>
</tr>
<tr>
<td>Systems administrator</td>
</tr>
<tr>
<td>Archive developer (shared w/ Data Conservancy)</td>
</tr>
</tbody>
</table>

---

**JHU Data Archive**
More on our origins

Data Conservancy → Entrepreneurial Library Program → JHU Data Management Services

Funded by schools → Consulting Service → JHU Data Archive

Fee to researcher

JHU Data Archive

Consulting Service

Funded by schools
ARL SPEC Survey: Research Data Management Services

ARL SPEC Kit 334 (July 2013)

Johns Hopkins Sheridan Libraries
Data Management Services

University of Virginia Library
Data Management Consultant Group

Available for download at ARL.org
## e-Research support by ARL libraries (Spring 2013)

### Survey of 73 research libraries:

<table>
<thead>
<tr>
<th>Service</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research data services</td>
<td>100%</td>
</tr>
<tr>
<td>Data management services</td>
<td>68%</td>
</tr>
<tr>
<td>Inst. Repositories</td>
<td>41% (30)</td>
</tr>
<tr>
<td>Research Data archives</td>
<td>7% (5)</td>
</tr>
<tr>
<td>Internal budgets</td>
<td>84%</td>
</tr>
<tr>
<td>Grants</td>
<td>24%</td>
</tr>
<tr>
<td>(5) Charge researcher</td>
<td>14%</td>
</tr>
</tbody>
</table>

### Staffing Service Structure

- **Mixed departments**
  - 15% Single position
  - 11% Single department

### Position Titles / Roles

- Subject Librarian
- Digital/Data Librarian
- Metadata Librarian
- Systems / IT

---

**JHU DATA MANAGEMENT SERVICES**

[Johns Hopkins Sheridan Libraries Logo]
Data Management Consultation

- In-office consultations with researchers
- Evaluate sharing/archiving options
- Give feedback on DMP drafts

Data Management Planning Tools
- Questionnaire for developing data management plans
- NIH data sharing guide
- Guide for reviewing data management plans
Questionnaire to Help with the Creation of a Data Management Plan

JHU Data Management Services of the Sheridan Libraries; datamanagement@jhu.edu

How to use this document
Use of this questionnaire will address NSF’s guidelines for what to include in your data management plan. Contact a JHU data management consultant at datamanagement@jhu.edu; they will come to your office to discuss your data management needs and, using this questionnaire, help you create an effective data management plan.

You may wish to fill out this questionnaire in part or in whole before meeting with a JHU data management consultant. See endnotes for additional guidance and instructions throughout the questionnaire (view by placing mouse over the red numbers in text). You may not need to address all questions under a heading. However, any omissions should be justified in your plan.

Note that your data management plan must adhere to proposal guidelines (2-page maximum and 1” margins for NSF).

Data Products and Standards
In the table below list data inputs and outputs (i.e., existing datasets you will use and intermediary/final datasets you will generate) in your proposed study, adding additional rows as needed. Include physical samples and collections to be used in the course of your research. This table and its elements can be included in your final plan and referenced in the rest of the questionnaire.

In filling out the table, consider your data workflow; that is, list your data inputs and outputs in the order they will be created and used, from physical samples and raw, unanalyzed datasets to intermediary datasets, and finally to analyzed datasets such as those visualized in peer-reviewed publications.

<table>
<thead>
<tr>
<th>Datasets (Input and/or Output)</th>
<th>Format(s)</th>
<th>Est Size/Amount</th>
<th>Shared?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ethical and Legal Compliance
1. Is access or storage of any of the datasets regulated by policy or law (e.g., classified data, specific handling
Grant Reviewer’s Guide for DMPs

Reviewer’s Worksheet for NSF Data Management Plans

The table & checklists cover NSF’s requested components of the proposal’s data management plan. A ‘*’ indicates details found in more thorough plans, and a quick measure of quality when checked. See pg.2 for more examples and guidelines.

<table>
<thead>
<tr>
<th>Research product</th>
<th>Source</th>
<th>Format</th>
<th>Size</th>
<th>Preserved (how?)</th>
<th>Shared (how?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.g., Tables, images, computer code, curriculum items, physical samples</td>
<td>Data repository, Instrument, interviews, PI’s prior project</td>
<td>JPG, MATLAB, Excel table, device’s format</td>
<td>&gt;1TB, 20K files</td>
<td>Discarded, PI retains, data archive</td>
<td>By request, website, repository</td>
</tr>
</tbody>
</table>

Data Sharing

☐ Is data publically accessible?
☐ When will data be shared?
☐ Who administers?
☐ Describes audience to benefit.

Preparation of data for sharing:
☐ Uses their research field’s metadata standards
☐ AND/OR creates description sufficient for re-use
☐ Metadata or supplementary files explaining: content/ file structure/ procedures/ codebook or variable-level detail
☐ Metadata associated with digital files
☐ Will convert files to non-proprietary formats

Data sharing policy:
☐ Gives conditions for re-use
☐ Accounts for:
☐ privacy (personal identifiers)/security issues
☐ intellectual property (copyrights, patents)

Data management during project:
☐ Storage: has a backup plan
☐ Location & media used:
☐ 2+ copies with 1 off-site
☐ Specifies who is responsible
☐ Data security / access controls
☐ Has conventions for naming & organizing files
☐ Version control
☐ Collaboration coordination

Data retention after the project:
☐ Where is data preserved?
☐ How long?
☐ Who administers?
☐ Gives reasons for preserving data (especially raw data)
Data Management Consultation

• Consultations for NSF DMPs started strong in 2011-2012, less demand for these services now
• Help with funder data management requirements continues
• NIH & OSTP funder requirements on the horizon
• Non-DMP consultations have increased
Support for Research Data Lifecycle

Data Management Planning

Data Management Consulting, archiving & sharing focused

JHU Data Archive, Evaluate archiving and sharing options

Data Organization

Storage/backup

Access/security

Metadata

Idea / proposal
acquire data
data management during the project
dissemination
archiving
preservation
Data re-use / discovery
“Concierge” service for the JHU Research Data Life Cycle

Data Management Planning

Research Conduct

General Counsel

Tech Transfer

Manage Data in JHU Data Archive

Research Admin

IRB(s)

Subject Librarians

Data/GIS Librarians

Preparing Data for Sharing & Archiving

Central IT

IT within Schools

Data manager group

Institutes with focus on security, clinical data, etc.

Institutional Research

Institutional Repository

Biostats Center

HPCs

JHU Research Data Life Cycle

Idea / proposal

Acquire data

Data management during the project

Dissemination

Preservation

Archiving

Data re-use / discovery

Manage Data in JHU Data Archive

Data Management Planning

Research Conduct

General Counsel

Tech Transfer

Manage Data in JHU Data Archive

Institutional Research

Institutional Repository

Biostats Center

HPCs

Preparing Data for Sharing & Archiving

Central IT

IT within Schools

Data manager group

Institutes with focus on security, clinical data, etc.
Data Management Training Workshops

• Arguably our most successful service
• 942 participants so far
• 4 one-hour session topics, 2 in development

Preparation Data Management Plans

Removing Human Subject Identifiers

Data Management Best Practices

Sharing Data in Spreadsheets
Preparing (NSF) Data Management Plans

• Give positives of data management plans vs. simple compliance

• Discuss data management concepts and plan sections (following our Questionnaire)

• Discussion around data management as it pertains to researchers in audience

<table>
<thead>
<tr>
<th>Data Sharing</th>
<th>Data management</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Is data publically accessible?</td>
<td>☐ Storage: has a backup plan</td>
</tr>
<tr>
<td>☐ When will data be shared?</td>
<td>☐ Location &amp; media used:</td>
</tr>
<tr>
<td>☐ Who administers?</td>
<td>☐ Specifies who is responsible</td>
</tr>
<tr>
<td>☐ ♦ Describes audience to benefit.</td>
<td>☐ Data security / access controls</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Management and Sharing Best Practices Topics

- Data Organization
- Storage and Backup
- Security and Access Controls
- Documentation and Metadata

- Part of JHU Responsible Conduct of Research training
Removing Identifiers from Human Subject Data

- How to locate & protect personal identifiers
- How to prepare de-identified datasets for sharing

Disclaimer: We are providing advice; IRB is the final authority on this subject
Goals of Spreadsheet Data Session

• Increase possibility for re-use of data in spreadsheets by others (and researchers themselves) in the future

• Reduce chance of error when using spreadsheets in your research.

### What do these variable codes mean?

### Does this patient code refer to PII?

### What do these variable codes mean?

### BAD! - ONE VALUE TO A CELL

### MISSING VALUE?
**Topic in development:**

- NIH data management plan preparation
- Preparing Data for Archiving & Sharing

---

**Organizing data for deposit**

<table>
<thead>
<tr>
<th>Figure heading</th>
<th>Sec</th>
<th>Folder name</th>
<th>Content description</th>
<th>Content notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 1. Development of HA functional surfaces.</td>
<td>1A</td>
<td>NIH 3t3 on HA and FN surfaces</td>
<td>IA Development of HA functional surfaces</td>
<td>Data or images for this figure are not archived in this collection.</td>
</tr>
<tr>
<td>Fig. 2. Cancer cell interaction with HA</td>
<td>2A</td>
<td>Figure 1</td>
<td>Florescence images HA on APTMS patches</td>
<td>Single file 2-5-09 (10x)HA note:</td>
</tr>
<tr>
<td></td>
<td>2B</td>
<td>2Bii infl memb CD44</td>
<td>1C Pseudocolor florescence imageJ analysis</td>
<td>HTB2640xCD4</td>
</tr>
</tbody>
</table>

- Reverse transcription polymerase chain reaction (RT-PCR) for quantities of ECM products and markers known in vascular differentiation in hESCs, SMLCs, and v-SMCs
- Fibronectin and collagen quantities
  - Fig. 3a and sup. 3a
  - Data Type: Raw Excel, visual representation is bar graph
- Statistical Analysis using Graph Pad
  - Figure 3a

---

- KDR, PDGFR-B, Neuropilin, SMA, Angiopoietin 1, VE-Cad, and FLT-1
- Fig. sup 2A-6
- Data Type: Raw Excel, Bar
Data Archiving Services

Archiving research data

- Planned data repository using software from the Data Conservancy project (dataconservancy.org)
- Online Access with preservation system

• Implemented Harvard’s Dataverse software as access platform (hosted by JHU Sheridan Library)
• Data Conversancy building a preservation system with APIs to interfaces

https://archive.data.jhu.edu

JHU DATA MANAGEMENT SERVICES
This project contains resulting publications from the National Science Foundation grant 1124648 and 0835549. Contains studies by Dr. Micheal Bevan and co-authors.

Size Dependent Thermodynamics and Kinetics in Electric Field Mediated Colloidal Crystal Assembly
by Edwards, Tara D.; Beltran-Villegas, Daniel J.; Bevan, Michael A.
Description: From the study abstract We compare measurements and models of the system size dependent assembly of quasi-two dimensional (2D) colloidal crystals within interfacial quadrupole electrodes. Perturbation theory (PT) and Monte Carlo (MC) simul... Continue [+]

https://archive.data.jhu.edu
<table>
<thead>
<tr>
<th>Folder Name</th>
<th>File Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Thermodynamic sys size microscopy experiments</td>
<td>1_Thermodynamic_sys_size_microscopy_experiments.zip</td>
<td>Video Microscopy showing experimental results (c.f. sec 1. supplemental info &amp; Fig. 1) testing thermodynamic system size effects. Includes Fig 1 movies in Windows Media video format.</td>
</tr>
<tr>
<td>1.1 MatLab code for microscopy experiments data output and movies</td>
<td>1A_MatLab_code_for_experiments_dataoutput_movies.zip</td>
<td>MatLab codes for producing video and experimental results in download File 1, for testing thermodynamic system size effects, and video for File 3. (c.f. sec 1, supplemental info &amp; Fig. 1)</td>
</tr>
<tr>
<td>1.2 TIFF files created by MatLab codes of file 1A (Optional 25GB original files)</td>
<td>1B_TIFF_output_for_1A.zip</td>
<td>Contains raw TIFF files associated with download file 1A generated by 'MatLab codes for experiments' code set, and assemble by 'MatLab code for movies' Optional download for recreation.</td>
</tr>
<tr>
<td>2 Perturbation Theory files</td>
<td>2_Perturbation_theory_files.zip</td>
<td>Perturbation Theory (PT) file (MS Excel format) with previously measured electric field potentials used to generate predictions of minimum voltage.</td>
</tr>
</tbody>
</table>

Download files ordered by publication sections & figures

Additional descriptive metadata at the file level
The JHU Data Archive

- Accepts data from all disciplines and formats
- All data publicly-accessible
- Data embargos by restricting data downloads
- Currently does not support restricted data (e.g., subject identifiers, security concerns)

- Consultants manage all data deposit and metadata entry. (No self-deposit by researchers)
- Guidance on preparing data for archiving
- Archiving is fee-based
## Archiving Service Models

<table>
<thead>
<tr>
<th></th>
<th>Large Data Collections Single Grant</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fee</strong></td>
<td>2% Total Direct Costs on a grant</td>
</tr>
<tr>
<td><strong>Data Size</strong></td>
<td>2 TB</td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>Consultant support throughout project</td>
</tr>
<tr>
<td></td>
<td>Preserve researcher data for 5 years</td>
</tr>
<tr>
<td></td>
<td>Renewable beyond 5 years (mainly for cost of storage)</td>
</tr>
</tbody>
</table>
Development of 2nd deposit model

- 2% cost model less attractive for large grants with smaller amounts of data to share
- Evaluated 3 pilot projects and initial smaller collection deposits

Analysis of data deposit activities

- Appraise and Select 42%
- Create or receive 36%
- Ingest 20%
- Preservation ~30 Hours Total

• 2% cost model less attractive for large grants with smaller amounts of data to share
• Evaluated 3 pilot projects and initial smaller collection deposits
Archiving Service Models

<table>
<thead>
<tr>
<th></th>
<th>Large Data Collections</th>
<th>Small Data Collections</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fee</strong></td>
<td>2% Total Direct Costs</td>
<td>$1,600</td>
</tr>
<tr>
<td></td>
<td>of grant</td>
<td></td>
</tr>
<tr>
<td><strong>Data Size</strong></td>
<td>2 TB</td>
<td>20 GB</td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>Consultant support</td>
<td>Supporting the data</td>
</tr>
<tr>
<td></td>
<td>throughout project</td>
<td>deposit process</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Preserve researcher data for 5 years</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Renewable beyond 5 years (mainly for cost of storage)</td>
<td></td>
</tr>
</tbody>
</table>

- 37 active and pending projects
- 13 collections online
- Open Access Data Fund in FY2015
JHU Data Archive service workflow

• Agreement to archive at proposal time (for large collections) or when data is ready (small collections)

• Researchers sign a Deposit Agreement

• Consultants meet with researchers to discuss deposit plan, early in project (ideally) or before deposit

• For Large Collection plan, consultants offer collection preparation advice during the project.
JHU Data Archive service workflow

• Ahead of deposit, researchers get deposit form to provide project descriptions & metadata for Dataverse collection
• Researchers do the file organization and prep (e.g., removing identifiers or proprietary info)
JHU Data Archive service workflow

- **In-person data transfer**: consultants get copy of data directly from researcher’s computer
- Transferred using Data Conservancy’s Package Tools
- Extracts and “bags” files with MD5 Checksum for copy integrity checks
- Interface for adding metadata to package

http://dataconservancy.org/software/downloads2/
JHU Data Archive service workflow

• At office: original data deposited in preservation system
• Consultants do minor data organization & augmentation of staging version of data
• Enter cataloging metadata in DVN, upload data
• Researchers review collection before release
• Final data is archived using DSC Package Tools
Consulting & Archiving service project management with CRM Database
Consulting & Archiving service project management with CRM Database

- Associate consulting, archiving and training Projects with Contacts and a workflow of Events
- Do-it-yourself, not recommended!
  Check out Microsoft Dynamics CRM or others
JHU DMS into the future

- Entering 4th year, adolescence if not full maturity.
- Becoming well integrated into University, finding new needs and service areas.
- Hiring 4th consultant (health sciences background desired).
- Multiple avenues for service: Can adapt if funder responses to OSTP Data Sharing policy are less of a kick-start than hoped for.
Questions?

datamanagement@jhu.edu
http://dmp.data.jhu.edu/