Data Management Plans For Universities and Institutes of Higher Education and Research

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services,
Texas State University Libraries
Data Management Plans: What are They?

• Plans to Manage a Researcher’s Data Outputs
• Long Term Archiving/Preservation Strategy
• Ingestion of Various Data Types (from Excel to SPSS to more esoteric forms)
• Permalinking Strategy for Online Citation and Access (DOI: Digital Object Identifier, UNF:University Numerical Fingerprint, Linked Data, Interoperability)
• Paratextual Research Material Archiving/Sharing Possibilities (field notes, docs, data, multimedia and programs)
Types of Data Set Project Repositories

1) Project specific
(usually large single faculty/faculty team projects)

2) Discipline specific
(i.e. Academic Specialization, Purdue Nanohub, Engineering etc.)

3) Institutional Repository
(either institution wide or consortial)
Why are Data Management Plans Required?

Increasingly, major grant agencies require data management plans as mandatory part of the grant proposal process.

(NIH 2003; NSF 2011; NEH, 2013 USDA)
Data Sharing

Currently, 80% of researchers do not share their data

Why Are Data Management Plans Required

- Leverage and make available faculty, departmental and institutional research
- Allow publication of negative data (less research replication)

_Wordle of the National Science Foundation’s Award and Administration Guide. Chapter VI.D.4, Mandatory 2011_
Data Management Plans

- Part of Evolving Science, Social Science and Humanities Research Process (Accuracy, efficiency, sharing)

Wordle of the data management policy of the Office of Digital Humanities, National Endowment for the Humanities, 2013
Research Data Management Software

• Online Repositories/Infrastructures created to manage a researcher’s data (sharing, archiving, preservation, metadata)
• May be hosted or installed on a university’s server
• Each software contains different ranges of management/collaborative options
• Open source and proprietary options
Current DMP Platforms (2014)

<table>
<thead>
<tr>
<th>Inst. Repository w/ Data (top 5)</th>
<th>Data-specific Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dspace</td>
<td>Dataverse</td>
</tr>
<tr>
<td>Fedora</td>
<td>Chronopolis</td>
</tr>
<tr>
<td>BePress Digital Commons</td>
<td>HubZero (customized)</td>
</tr>
<tr>
<td>Hydra</td>
<td>DataConservancy</td>
</tr>
<tr>
<td>Drupal</td>
<td>Custom repository</td>
</tr>
</tbody>
</table>

Dataverse: Harvard’s Research Data Open Source Solution

Software framework that enables institutions to host research data repositories

Preservation and archival Infrastructure: allows sharing, control, persistent data citation, publishing and management

Social Sciences Beginnings (IQSS)
Data Science (site)
http://thedata.org
Dataverse Open Source Download (Github), Software Background
Dataverse Network Architecture

**Why the Dataverse Network?**
(silent video overview)

**Open Journal Systems**
**Dataverse Integration** (2014)

**Research Study Data**
Data Set Files
Metadata (Data Describing the data)
Paratextual Research Material
(Methodology, Field Notes etc.)
Graph Data Files
Dataverse Metadata Example

Harvard Dataverse Network

REPLICATION DATA FOR: A MULTIVARIATE MODEL OF STRATEGIC ASSET ALLOCATION

hdl:10.02.1/QBXRSLBQJUNF.3.ZnYhHk2e2vTJAWhBDpPKA---

CATALOGING INFORMATION

Data Citation

If you use these data, please add the following citation to your scholarly references. Why cite?


Citation Format: Print

Original Publication


Publications


Data Citation Details

Title
Replication data for: A Multivariate Model of Strategic Asset Allocation

Study Global ID
hdl:10.02.1/QBXRSLBQJ

Authors
John Y. Campbell (Harvard University); Yeung L. Chan; and Luis Viceira

Producer
John Y. Campbell

Production Date
2003

Funding Agency
National Science Foundation; Hong Kong RGC Competitive Earmarked Research Grant (HKUST 6965/01H); Division of Research of the Business School
# Dataverse Metadata Example

## Citation Metadata

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Data from “Social determinants of unmet hospitalisation need amongst the poor in Andhra Pradesh, India: A cross-sectional study.”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author</strong></td>
<td><strong>Name</strong> Nagulapalli, Srikant <strong>Affiliation</strong> Andhra University</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>The dataset is of a health survey amongst the 21.5 million poor families of the Indian state of Andhra Pradesh conducted during April and May 2013. The dataset captures individual characteristics and household characteristics of the past 365 days and was used to analyse the unmet hospitalisation need in the Indian State of Andhra Pradesh. Data was collected by 2022 trained field staff of Aarogyasri Health Care Trust (AHCT) of Government of Andhra Pradesh using a questionnaire modelled after that used for the health surveys by National Sample Survey Organisation of India.</td>
</tr>
<tr>
<td><strong>Subject</strong></td>
<td>Medicine, Health &amp; Life Sciences</td>
</tr>
<tr>
<td><strong>Keyword</strong></td>
<td>unmet hospitalisation need</td>
</tr>
<tr>
<td><strong>Production Date</strong></td>
<td>2013-06-01</td>
</tr>
<tr>
<td><strong>Depositor</strong></td>
<td>Privileged, admin</td>
</tr>
<tr>
<td><strong>Deposit Date</strong></td>
<td>2013-08-03</td>
</tr>
</tbody>
</table>
PURR and Hubzero: Purdue’s Data Management System

1.) Create Data Management Plans
2) Collaborate with other Researchers
3) Publish Data Sets (Purdue can publish a DOI: Digital Object Identifier for Data Sets)
   Useful For Citation
4) Archive Data Sets

Boilerplate text for data management proposals available

Purr is part of Hubzero platform for scientific collaboration (Originally Nanohub)

- Purr: Purdue University Research Repository (video)
- Purr Site (Proprietary to University)
- Purr Background
Hubzero: Open Source Platform for Scientific Collaboration

- [https://hubzero.org/](https://hubzero.org/)
- [Getting Started, Downloadable](https://hubzero.org/) and [Hosted Options](https://hubzero.org/)
- [Hubzero Video, Hubzero2](https://hubzero.org/)

Research Collaboration and Data Management Solution

**Research Data Types**
- Spreadsheets
- Instrument or Sensor Readings
- Software Source Code
- Surveys
- Interview Transcripts
- Images and Audiovisual Files
Hubzero/Purr Customization

Start Your Research Project

Create a Data Management Plan
Learn about the detailed requirements for your data management plan (DMP). Funding agency requirements are very specific and our DMP resources can help you to clear up any confusion. Get Started ›

Upload Research Data to Your Project
Create a project to upload and share your data with collaborators using our step-by-step form to guide you through the process. Invite collaborators from other institutions to join your project. Create a Project ›

Publish your Dataset
Package, describe, and publish your dataset with a Datacite DOI. Publishing will ensure your dataset is citable, reusable, and archived for the long-term. See Published Datasets ›
Localization of potential regulatory signals in the human genome

By Minou Bina, Phillip J Wyss

Purdue University


Abstract
In the human genome, the expression of protein-coding genes is in part regulated by specific DNA sequence elements known as proximal and basal promoters. These elements are relatively short and usually correspond to transcription start site (TSS) determining the genomic DNA sequences that control the expression of human genes. However, it has been difficult to identify these regions via various experimental strategies including chromatin immunoprecipitation assays (ChIPs). To develop a computational approach to identify these promoter regions would provide a rich source of data for designing tools to predict the position of regulatory signals. Based on these considerations, we have designed a scheme to create density plots of sequences that appear frequently upstream of TSSs.
Repository where users make their research available in citable, shareable and discoverable manner

Figures, datasets, media, papers, posters presentations and file sets can be disseminated in a way that the current scholarly publishing Model does not allow

Open Source Platform for Sharing Research

Figshare (video)

Figshare for Institutions (Video)
Figshare Features
(Cloud Based/Proprietary)

1GB of private space
- taggable and easily filtered, your research data is better managed and easy to locate

Unlimited public space
- upload to your heart's content, the more - the better

Publish negative data
- all published research is citable

Upload all formats

Quick & simple upload

Cloud based
One Size Does Not Fit All Data Project Needs

Types of Data Projects (Sizes)

1) Normal range
Files/Data Fit on Server/Cloud, may be uploaded, Dataverse, Purr)

2) Large Projects
(Data may require specialized university IT Support, i.e. terabyte/petabyte tape drives)

3) Very Large
(Projects require consortial possibilities, national models, Chronopolis, Texas Advanced Computer Center, DEEPN, Duracloud)
DMP Documentation/Policy Tool

Overview Video

Customizable Plan Outline Tool Resource Links Supports All Major Funders

https://dmptool.org/
California Digital Library
Further Links/References

- California Digital Library DMT Tool: [https://dmptool.org/](https://dmptool.org/)
- Dataverse (Data Science Site). [http://datascience.iq.harvard.edu/dataverse](http://datascience.iq.harvard.edu/dataverse)
- Purr. (Purdue Institutional Data Repository). [https://purr.purdue.edu/](https://purr.purdue.edu/)
- Hubzero. [https://hubzero.org/](https://hubzero.org/)
- “Research Data Management”. pp. 6-7 and pp 24 – 45.
Comments/Questions