Best Practices for Research Data Management

Jean-Paul Courneya, Bioinformationist
jpcourneya@hshsl.umaryland.edu

Amy Yarnell, Data Services Librarian
ayarnell@hshsl.umaryland.edu

April 16, 2020
Learning Objectives

• Recognize the benefits and requirements related to the management and sharing of research data
• Apply best practices for data planning and management
• Understand options for storing and preserving research data
• Identify repositories and determine best sharing option for data
Why Data Management?
✓ Funding agency data management and sharing requirements
✓ Publisher data sharing policies
✓ NIH rigor and reproducibility
Publisher requirements

…must specify that data are deposited publicly and list the name(s) of repositories along with digital object identifiers or accession numbers”

“All data necessary to understand, assess, and extend the conclusions of the manuscript must be available

http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml
Why Data Management?

• Transparency
• Re-use and innovation
• Reproducibility and Replicability
• Open Science
HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?
Most scientists have experienced failure to reproduce results.

<table>
<thead>
<tr>
<th>Field</th>
<th>Others</th>
<th>My own</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physics and engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earth and environment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Always/often contribute</th>
<th>Sometimes contribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selective reporting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pressure to publish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low statistical power or poor analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not replicated enough in original lab</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insufficient oversight/mentoring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods, code unavailable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor experimental design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw data not available from original lab</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fraud</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insufficient peer review</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problems with reproduction efforts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical expertise required for reproduction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variability of standard reagents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad luck</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Have you ever...?

- Conducted research with a team?
- Put data on a flash drive that got lost or broken?
- Collected data that you had trouble understanding later?
Why Data Management?- What’s in it for me?

- Organization
- Comprehensibility
- Efficiency
- Quality
- Access
Lost data!

Retraction Watch

NEJM paper on sleep apnea retracted when original data can’t be found

The authors of a paper in the New England Journal of Medicine are retracting it, after being unable to find data supporting a table that required corrections.

Multiple errors in table
Did not alter conclusions in article

**BUT**, could not locate primary data

http://retractionwatch.com/2013/10/30/nejm-paper-on-sleep-apnea-retracted-when-original-data-cant-be-found/
Data management best practices
Data Management Planning
Data Lifecycle

Plan for Data

Share Data → Collect Data

Preserve Data → Clean Data

Store Data → Analyze Data

Analyze Data → Plan for Data
Planning!

Begin with a clear, well-thought out hypothesis. Your research question will guide your data collection plan.

What data will you need to collect?

What do you plan to do with the data? Why are you collecting it? What do you need to measure?

How will you collect the data? Who will enter the data?

Are there ethical considerations that will complicate data collection?
Data Management Plans (DMPs)

A formal plan that:

- describes the data your research will produce
- describes how your data will be handled during and after your project
- is being required by more and more funders
- is typically less than three pages
Data Management Plans

https://dmptool.org

Welcome to the DMPTool
Create data management plans that meet institutional and funder requirements.

**DMPTool by the Numbers**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>🧨 Users</td>
<td>43,821</td>
</tr>
<tr>
<td>🗂 Plans</td>
<td>40,887</td>
</tr>
<tr>
<td>🎯 Participating Institutions</td>
<td>263</td>
</tr>
</tbody>
</table>

**Top Templates**

- Digital Curation Centre
- NIH-GEN: Generic
- NSF-SBE: Social, Behavioral, Economic Sciences
- NSF-DSE: Computer and Information Science and Engineering
- NSF-BIO: Biological Sciences

---

DMPTool News

DMP services unitel!

Go to the blog

RSS

16
Data Management Plans

https://dmptool.org

DMPTool by the Numbers

Top Templates
Data Management Plans

https://dmptool.org

Data sharing plan (0 / 1)

Investigators seeking $500,000 or more in direct costs in any year should include a description of how final research data will be shared, or explain why data sharing is not possible.

Data format

- Clearly note what format(s) your data will be in, e.g., plain text (.txt), comma-separated values (.csv), geo-referenced TIFF (.tiff, .tif).
- Explain why you have chosen certain formats. Decisions may be based on staff expertise, a preference for open formats, the standards accepted by data centers, or widespread usage within a given community.
- Using standardized, interchangeable, or open formats ensures the long-term usability of data; these are recommended for sharing and archiving.
- See DataONE Best Practices for file formats.

Metadata & documentation
Who is responsible for data management?

Everyone!

(but everyone means no one without assigning responsibility)
Assign a person to be responsible for ensuring quality control:

✓ File naming conventions adhered to
✓ Minimum documentation
✓ Version controls followed
✓ Data backed up
Data Collection
Your variables

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SID</td>
<td>wgt</td>
<td>smoking</td>
<td>name</td>
<td>sam</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>49</td>
<td>Y</td>
<td>Smith</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>252</td>
<td>2 packs</td>
<td>Sam Jones</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>28</td>
<td>N</td>
<td>Read, Kevin</td>
<td>A21</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>157</td>
<td>Never</td>
<td>Emma Banks</td>
<td>January</td>
</tr>
</tbody>
</table>
Coding data will reduce inconsistency in data entry

0 = no high school

1 = some high school

2 = graduated high school

3 = some college

4 = graduated college

Code missing data! Participants might not remember exact dates or might not want to disclose information.

8888 = participant cannot remember (date of appendectomy)

9999 = participant will not disclose (past drug use)
Variables – Best Practices

Do not calculate variables

<table>
<thead>
<tr>
<th>USE THESE</th>
<th>NOT THIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic BP, Diastolic BP</td>
<td>Hypertension— yes/no</td>
</tr>
<tr>
<td>Height, weight</td>
<td>BMI</td>
</tr>
<tr>
<td>Temperature</td>
<td>Fever— yes/no</td>
</tr>
</tbody>
</table>

Avoid oversimplification

Medication – yes/no; dose; duration of treatment
# Data Dictionaries

<table>
<thead>
<tr>
<th>#</th>
<th>Variable / Field Name</th>
<th>Form Name</th>
<th>Field Type</th>
<th>Field Label</th>
<th>Choices, Calculations, OR Slider Labels</th>
<th>Text Validation Min</th>
<th>Text Validation Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>record_id</td>
<td>demographics</td>
<td>text</td>
<td>Record ID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>mrn</td>
<td>demographics</td>
<td>text</td>
<td>MRN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>last_name</td>
<td>demographics</td>
<td>text</td>
<td>Last name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>first_name</td>
<td>demographics</td>
<td>text</td>
<td>First name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>age</td>
<td>demographics</td>
<td>text</td>
<td>Age</td>
<td>1, &lt;55</td>
<td>2, between 55 and 75</td>
<td>3, &gt;75</td>
</tr>
<tr>
<td>6</td>
<td>gender</td>
<td>demographics</td>
<td>radio</td>
<td>Gender</td>
<td>1, Male</td>
<td>2, Female</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>race</td>
<td>demographics</td>
<td>radio</td>
<td>Race/Ethnicity</td>
<td>1, White</td>
<td>2, Black</td>
<td>3, Asian</td>
</tr>
<tr>
<td>8</td>
<td>describe_other</td>
<td>demographics</td>
<td>text</td>
<td>Describe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>education</td>
<td>demographics</td>
<td>radio</td>
<td>Highest Level of Education Completed</td>
<td>1, &lt; highschool diploma</td>
<td>2, highschool diploma</td>
<td>3, associate degree</td>
</tr>
<tr>
<td>10</td>
<td>yes</td>
<td>demographics</td>
<td>radio</td>
<td>Working</td>
<td>1, Yes</td>
<td>2, No</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>age</td>
<td>demographics</td>
<td>text</td>
<td>Occupation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>income</td>
<td>demographics</td>
<td>radio</td>
<td>Household Income</td>
<td>1, Household income &lt;30,000/year</td>
<td>2, Household income between 30-50,000/year</td>
<td>3, Household income 50-75,000/year</td>
</tr>
<tr>
<td>13</td>
<td>htn</td>
<td>medical_history</td>
<td>radio</td>
<td>Hypertension</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>14</td>
<td>hld</td>
<td>medical_history</td>
<td>radio</td>
<td>Hyperlipidemia</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>15</td>
<td>dm</td>
<td>medical_history</td>
<td>radio</td>
<td>Diabetes</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>16</td>
<td>current_smoker</td>
<td>medical_history</td>
<td>radio</td>
<td>Current Smoker</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>17</td>
<td>former_smoker</td>
<td>medical_history</td>
<td>radio</td>
<td>Former Smoker</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>18</td>
<td>smoking_start_date</td>
<td>medical_history</td>
<td>text</td>
<td>Smoking start date</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>smoking_quit_date</td>
<td>medical_history</td>
<td>text</td>
<td>Smoking Quit Date</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>depression</td>
<td>medical_history</td>
<td>radio</td>
<td>Depression</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>21</td>
<td>anxiety</td>
<td>medical_history</td>
<td>radio</td>
<td>Anxiety</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>22</td>
<td>stress_cardiomyopathy</td>
<td>medical_history</td>
<td>radio</td>
<td>Stress Cardiomyopathy</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>23</td>
<td>prior_mi</td>
<td>medical_history</td>
<td>radio</td>
<td>Prior MI</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>24</td>
<td>prior_stroke</td>
<td>medical_history</td>
<td>radio</td>
<td>Prior Stroke</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>25</td>
<td>prior_tia</td>
<td>medical_history</td>
<td>radio</td>
<td>Prior TIA</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>26</td>
<td>prior_hf</td>
<td>medical_history</td>
<td>radio</td>
<td>Prior HF</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>27</td>
<td>etoh_use</td>
<td>medical_history</td>
<td>radio</td>
<td>Alcohol Use</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>28</td>
<td>etoh_use_quantity</td>
<td>medical_history</td>
<td>radio</td>
<td>How much alcohol do you drink in a week</td>
<td>1, 1-3 drinks</td>
<td>2, 4-7 drinks</td>
<td>3, 7-15 drinks</td>
</tr>
<tr>
<td>29</td>
<td>mj_use</td>
<td>medical_history</td>
<td>radio</td>
<td>Marijuana Use</td>
<td>1, Yes</td>
<td>2, No</td>
<td>3, Unknown</td>
</tr>
<tr>
<td>30</td>
<td>age_at-menopause</td>
<td>medical_history</td>
<td>text</td>
<td>Age at Menopause</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Document your variables

• Intuitive / meaningful variable names e.g. study_id
• What do variable names mean?
• What does each variable contain?
• Are there a limited set of possible values?

<table>
<thead>
<tr>
<th>Name</th>
<th>Field Type</th>
<th>Description</th>
<th>Possible values</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>study_id</td>
<td>text</td>
<td>Unique ID of study</td>
<td>8-digit number</td>
<td></td>
</tr>
<tr>
<td>date_enrolled</td>
<td>date</td>
<td>Initial subject enrollment</td>
<td>Date in format YYYY-MM-DD; All dates later than 2011-09-01</td>
<td></td>
</tr>
<tr>
<td>weight</td>
<td>integer</td>
<td>Weight of subject</td>
<td></td>
<td>lbs</td>
</tr>
</tbody>
</table>
File Organization
PRO TIP: NEVER LOOK IN SOMEONE ELSE’S DOCUMENTS FOLDER.

OH MY GOD.

CC BY-NC2.5 https://xkcd.com/1459/
File Names

sam_1262011.tif
File Names

sam_1262011.tif

12 June, 2011?
December 6, 2011?
January 26, 2011?
File Names

sam_1262011.tif

12 June, 2011?
December 6, 2011?
January 26, 2011?

Unambiguous dates, the **ISO standard**: 
- YYYYMMDD *or* YYYY-MM-DD
  - e.g. 20120612 = June 6, 2012
- YYYYMMDDTHH:MM:SS
  - e.g. 20120612T14-03-12 = June 6, 2012 2:03:12 pm
File Names

sam_1262011.tif

Scanning acoustic microscope? 12 June, 2011?
Systolic anterior motion? December 6, 2011?
Sam the postdoc? January 26, 2011?

Unambiguous dates, the ISO standard:

- YYYYMMDD or YYYY-MM-DD
  - e.g. 20120612 = June 6, 2012
- YYYYMMDDTHH:MM:SS
  - e.g. 20120612T14-03-12 = June 6, 2012 2:03:12 pm
1 rat heart

100s of slices

3 post docs

100s of slides

5-7 experiments a week...

100s of huge images

1000s of image files
File names should...

1. Embody their content, including major parameters

AtherRat_ex012_ather_lipitor_128.tif
File names should...

1. Embody their content, including major parameters
   AtherRat_ex012_ather_lipitor_128.tif

2. Have non-cryptic/intuitive names where possible
   AtherRat_SOP_DataValidation_v01.docx
3. Be extensible. “ex001” not “ex1”
4. Be unique, where possible and practical.

Avoid 20 files named “data.xlsx” in different folders
4. Be unique, where possible and practical. Avoid 20 files called “data.xlsx” in different folders
File names should...

5. Do not use special characters – restrict file names to numbers, letters, and underscores
6. Use consistent, documentable rules for naming files

AtherRat_012_056_mb_0423_raw.csv

AtherRat = experiment name

012 = experiment number

056 = sample number

mb = stain used, methylene blue

0423 = 2-digit coordinates of image (4 across, 23 down)

Raw = data stage
In the folder…

<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtherRat_ex012_ather_lipitor_126.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_lipitor_127.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_lipitor_128.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_lipitor_129.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_001.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_002.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_003.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_004.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_005.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_006.tif</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF image</td>
</tr>
</tbody>
</table>
In the folder...

<table>
<thead>
<tr>
<th>Name</th>
<th>Date Modified</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtherRat_ex012_ather_lipitor_126.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_lipitor_127.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_lipitor_128.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_lipitor_129.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_001.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_002.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_003.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_004.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_005.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
<tr>
<td>AtherRat_ex012_ather_notreat_006.tiff</td>
<td>5/9/2014 7:55 PM</td>
<td>TIFF imag</td>
</tr>
</tbody>
</table>
Storage Solutions at UMB

**Office 365** – One Drive

**SOMFiles** – Backup data storage for PIs

Discuss storage options with IT at your school
Storage Options

If you are going to use cloud storage:

- Talk to IT at your school.
- On campus: SharePoint, OneDrive
- Others: Google Cloud Drive, Amazon, Box.
How will you back up your data?

How frequently will data be backed up?

How long will backups be stored?

How much storage space will be needed?

And how can you keep track of different versions of data, especially when backing up to multiple devices?
Save multiple copies…

…and disperse them geographically
Security Considerations

Certain data requires special protection:

• Protected Health Information
• Patents or commercial data
• Data as intellectual property
Security Extra Steps

- Password protect files or folders.
- Lock computers when not in use.
- Have others sign data use agreements.
- At UMB, use Accellion for transfer of secure files.

http://pixabay.com/p-538722/?no_redirect
Preservation
storage ≠ preservation
Preservation

Protects from: Hardware obsolescence
Preservation

Protects from: Software obsolescence
Preservation

Collecting data

Disseminating data
Trusted Open Formats

XML
CSV
PDF
TIFF
However...

Microsoft Excel

Molecular Devices
pClamp Software
Data Formats

Encryption and Compression

https://pixabay.com/en/key-tool-open-lock-security-2114459/
Data Formats

You can’t assume you own your data
Check for:

• Funder policies on data ownership

• Institution policies on data ownership
Providing Access
Why share data?

Open scientific inquiry

Verification and reproducibility

Promotes new research

Creates teaching resources

Increase visibility and impact

Prevents faulty or fraudulent research

Reduces duplication

Encourage collaboration

Better Research!
Data sharing challenges

- Time and effort
- Fear of losing control of data
- Confidential and sensitive information
- Ownership of data
- Lack of incentives
- Inexperience with data management
Providing access to your data

• Access vs. **meaningful** access

• Well-documented data
Repository Types

- Institutional
- Cross Disciplinary
- Discipline Specific
## NIH Data Sharing Repositories

This table lists NIH-supported data repositories that accept submissions of appropriate data from NIH-funded investigators (and others). Also included are resources that aggregate information about biomedical data and information sharing systems. The table can be sorted according by name and by NIH Institute or Center and may be searched using keywords so that you can find repositories more relevant to your data. Links are provided to information about submitting data to and accessing data from the listed repositories. Additional information about the repositories and points-of-contact for further information or inquiries can be found on the websites of the individual repositories.

<table>
<thead>
<tr>
<th>IC</th>
<th>Repository Name</th>
<th>Repository Description</th>
<th>Data Submission Policy</th>
<th>Access to Data</th>
<th>Search:</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI</td>
<td>The Cancer Imaging Archive (TCIA)</td>
<td>The Cancer Imaging Archive (TCIA) is a large archive of medical images of cancer accessible for public download. All images are stored in DICOM file format. The images are organized as &quot;Collections&quot;, typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus.</td>
<td>How to Submit Data to TCIA</td>
<td>How to Access TCIA Data</td>
<td></td>
</tr>
<tr>
<td>NCI (NHGRI, NIGMS)</td>
<td>PeptideAtlas</td>
<td>PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences.</td>
<td>How to Submit Data to PeptideAtlas</td>
<td>How to Access PeptideAtlas Data</td>
<td></td>
</tr>
<tr>
<td>NHGRI</td>
<td>FlyBase: A Drosophila Genomic and Genetic Database</td>
<td>Drosophila Genomic and Genetic database that includes proteomics data, microarrays and Tiling BAC's.</td>
<td>How to Submit Data to FlyBase</td>
<td>How to Access FlyBase Data</td>
<td></td>
</tr>
<tr>
<td>NHGRI</td>
<td>The Zebrafish Model Organism Database (ZFIN)</td>
<td>ZFIN serves as the zebrafish model organism database. It aims to: a) be the community database resource for the laboratory use of zebrafish, b) develop and support integrated zebrafish genetic, genomic and developmental information, c) maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e) facilitate the use of zebrafish as a model for human biology, and f) serve the needs of the research community.</td>
<td>How to Submit Data to ZFIN</td>
<td>How to Access ZFIN Data</td>
<td></td>
</tr>
<tr>
<td>NHGRI</td>
<td>WormBase</td>
<td>WormBase is an international consortium of biologists and computer scientists dedicated to providing the research community with accurate, current, accessible information concerning the genetics, genomics and biology of C. elegans and related nematodes.</td>
<td>How to Submit Data to WormBase</td>
<td>How to Access WormBase Data</td>
<td></td>
</tr>
</tbody>
</table>

Research Data Repositories

http://www.re3data.org/
**Dataset of an Exploratory Review of Clinical Study Reports of Randomised Controlled Trials**

**Internal Dataset** [unpublished]

**Author(s):** Peter Doshi

**Subject Domain**
- Drug Industry
- Randomized Controlled Clinical Trials as Topic

**Keywords**
- clinical study reports/evaluation

**Description**
This dataset is associated with an exploratory evaluation of pharmaceutical industry clinical study reports (CSR) for possible use in evidence synthesis and systematic reviews. 78 CSRs from public sources were selected for data extraction. The report dates ranged from 1991 through 2011, inclusive, and represented 50 randomized controlled trials of 14 pharmaceuticals. The primary outcome measures included presence and length of essential elements of trial design and reporting and compression factor (ratio of page length for CSRs compared to its published counterpart in a scientific journal). The dataset is comprised of an audited table of extracted and derived variables. Data were extracted on MS Word extraction tables, migrated to MS Excel, and audited (double-checked). The Excel file contains multiple “sheets” (worksheets) the contents of which are described in an accompanying readme file. The uncorrected (original) and corrected extraction sheets as well as audit records are available upon request from Peter Doshi, corresponding author (pdoshi@rx.umaryland.edu).

**Access via Dryad**
Dataset and readme file

**Access Restrictions**
Free to All

**Access Instructions**
Available to download from the Dryad site

**Associated Publications**

**Data Type**
Administrative

**Dataset Format(s)**
Microsoft Excel

**Dataset Size**
96.98 Kb
Data Publishing

Submit to *Scientific Data* in three simple steps:

1. **DESCRIBE**
   Write a detailed description of your dataset. We have templates to help you and a detailed guide to authors.

2. **DEPOSIT**
   See our list of recommended repositories. We will help you find the right place for your data.

3. **SUBMIT**
   Submit online and get the credit you deserve for your data!

Get credit where credit’s due and share your data.

Sample Data Description:

Proteomic profile of embryonic stem cell pluripotent stem cell resource library.

Search GigaScience
Conclusion
How is the library supporting data management?

- Workshops, consultations, and assistance
  - Data management plan guidance
  - Storage and preservation recommendations
  - File naming and organization advice
  - Data citation
  - Locating secondary data for reuse
  - Recommending technology for specific needs

- Data catalog

- Referrals to other UMB experts
Research Data Management Services Introduction

Consult with a member of our data management team for assistance with developing a data management plan, or in locating, describing, storing, or sharing data. Following best practices in research data management can help you, your lab, or center secure grant funding and create data output that becomes part of the scholarly record. The FAIR Data Principles provide a guideline for facilitating Findability, Accessibility, Interoperability, and Re-usability.

Given the importance of managing data, many funding agencies are requiring that plans for managing data be submitted with grant applications. An effective plan to collect, share, reproduce and preserve data may increase the impact of your research.

Read about HS/HSL’s workshops, including Data Management 101 and Creating a Data Management Plan with the DMP Tool.
Questions?
Workshop Attribution

Thanks to the NYU Health Sciences Library’s Data Services Team for developing the template for this workshop

Kevin Read
Data Services Librarian and Lead, Data Discovery

Alisa Surkis
Assistant Director, Research Data and Metrics
Photo references

- Wikipedia.org. “To Share or Not to Share” 2013 http://upload.wikimedia.org/wikipedia/commons/6/65/To_deposit_or_not_to_deposit,_that_is_the_question_-_journal.pbio.1001779.g001.png
- Wikimedia. “Modern warehouse pallent rack storage system” 2000 http://upload.wikimedia.org/wikipedia/commons/a/a2/Modern_warehouse_with_pallet_rack_storage_system.jpg
- Sarah. “Metadata is a love note to the future” 2011 https://www.flickr.com/photos/sarahseverson/6245395188/
- Niklas Wilkstrom. “Sharing is caring” 2010 https://www.flickr.com/photos/niklaswikstrom/5214708665/
- Github. “Repositories images” 2013 https://pages.github.com/images/ghfm@2x.png
- Futureatlas.com. “Citation needed” 2010 https://www.flickr.com/photos/87913776@N00/5129607997/