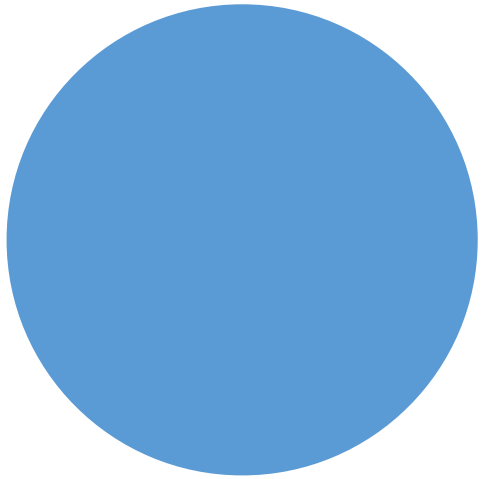
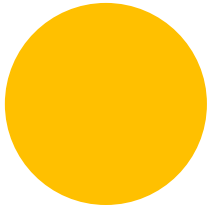



Data Management

What is it and how do you do it?

The purpose of this is not to tell anyone how to conduct their research

What are we talking about?

The purpose is to discuss concepts and best practices related to how you plan, organize, save, and share the data generated from your research

What is Data management?

Activities related to the collection, processing, analysis, preservation, and publication of data to ensure for easy reuse by the original researcher and sharing with the research community.

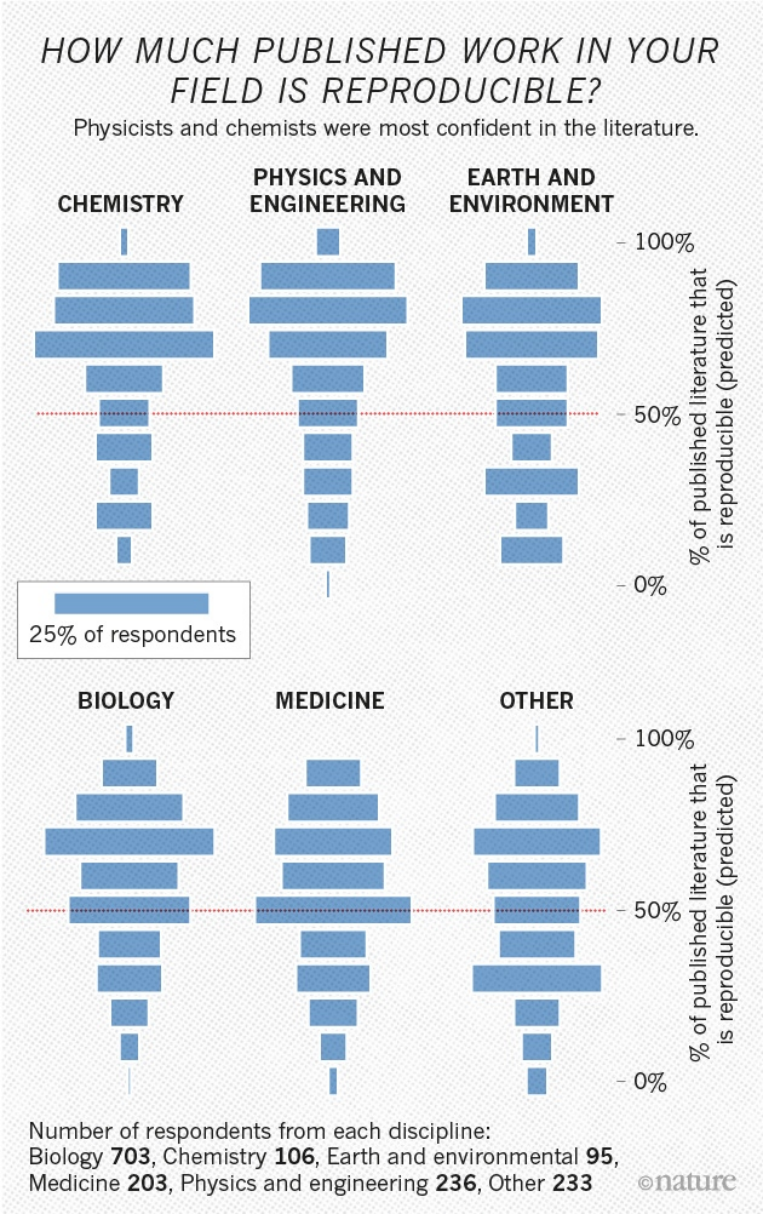
The Data Management Associations book “Data Management Body of Knowledge (DMBOK)” refers to Data Management as:
“The development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles.”

Why is Data Management important?

1. Everyone in the lab generates and/or needs access to data
2. Everyone needs to know where the data is and to avoid duplication of data
3. Funders, Publishers, or Institutions may require you to report on your data
4. Ensure transparency and reproducibility of your findings
5. Enable efficient access to data regardless of how much time has past or how many people have passed through the lab

Data Management aids reproducibility

The Reproducibility Crisis



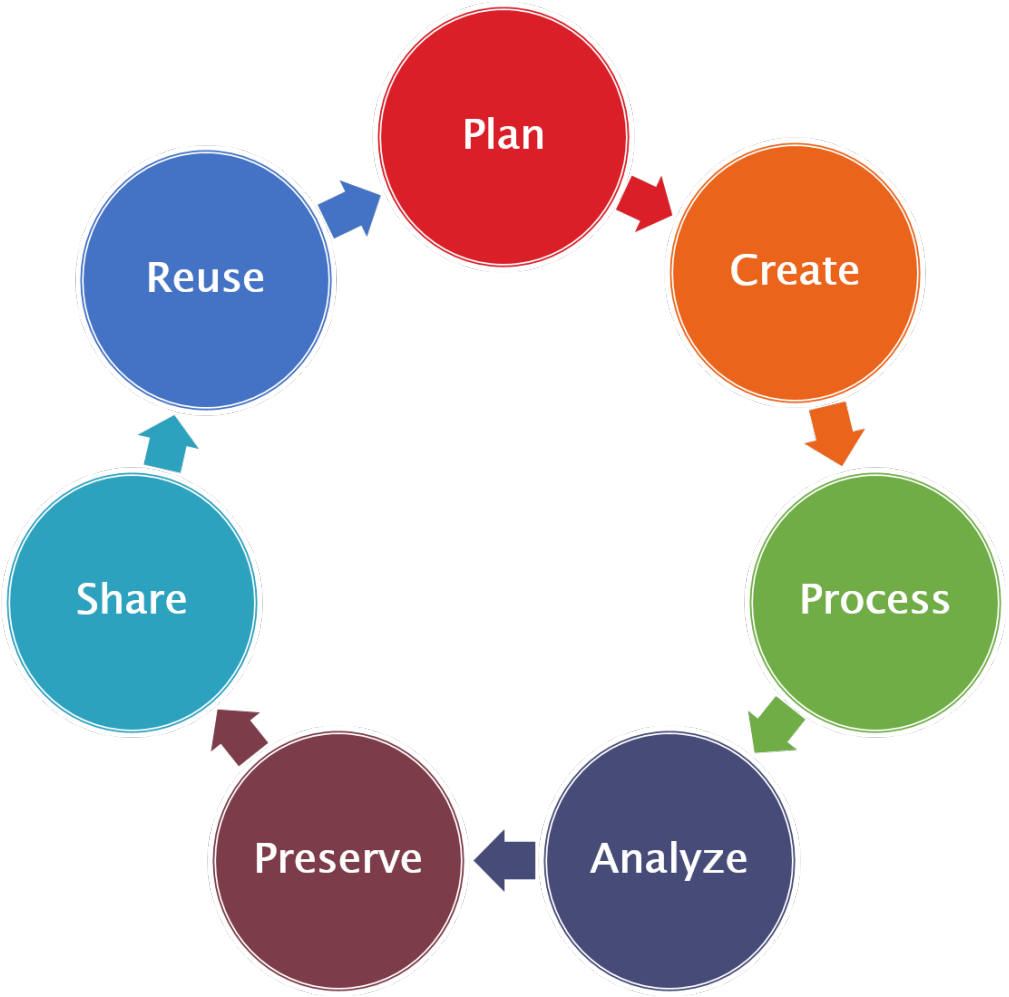
What is really happening to all the data that's being published?

https://www.youtube.com/watch?time_continue=2&v=N2zK3sAtr-4

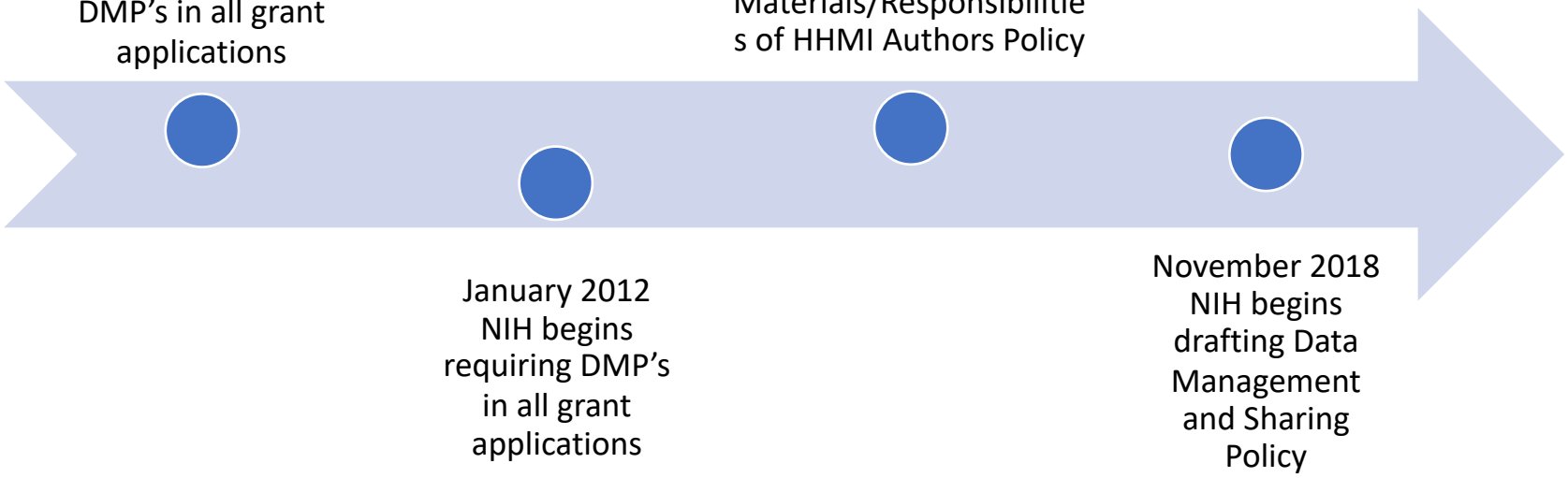
How can we be better?

Use proper Data Management practices and utilize Data Management Plans to make sure any generated data is accessible throughout the Research Data life cycle.

Data Life Cycle



The recent history of Data Management



Where do we start?

Planning

Why is Planning Important?

It is important to approach how you will manage your data in the same way that you approach planning your research.
Just like you consider the questions you are trying to answer, the methods you'll use to answer them and how you will present the information you obtained you have to think about your data.

You have to consider how the data will be

1. Organized
2. Saved
3. Prepared
4. Analyzed
5. Shared

The 30,000ft view

Data Types	Lab Roles	Data Storage	Data Archiving
What data is being generated? What is the format of the data?	Who is responsible for organizing and enforcing the plan?	Where will the data generated be stored? How will it be backed up?	How will the data be preserved? How will it be shared with others?

What does all this planning lead to?

All of your careful planning creates your

DATA MANAGEMENT PLAN

Your Data Management Plan will set out

1. The type of data you are collecting and the formats you are using
2. How the data will be documented so it can be understood by others
3. How the data will be prepared and analyzed
4. How the data will be stored
5. How the data may be shared later

Important concepts to understand

- **Data**
- **Metadata**
- **Organization**
- **Preparation and Analysis**
- **Storage**
- **Archiving/Sharing**

What is Data?

"the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." -Federal Office of Management & Budget Circular A-110

When we consider what constitutes data we have to include not only the obvious, raw and processed data from each study but also the code and other documentation related to the study

What is Metadata?

Metadata is the “**Data about the Data**”

Metadata basics

WHO, WHAT, WHERE, WHEN and HOW

- **Who** created the data?
- **Who** maintains it?

- **What** is the data about?
- **What** is the content of the data?
- **What** is the structure of the data?

- **Where** was it collected (geographic location)?
- **Where** is the data stored

- **When** were the data collected?
- **When** were they published?

- **How** were the data created?
- **How** should the data be analyzed?
- **How** should the data be cited?

Metadata formats

Metadata comes in a variety of formats but the most common are

1. **A text or HTML document**
   This will record all the information necessary and the document could be kept with the data files. A text file could also be used to create a data dictionary. This records information about metadata elements, sub-elements and attributes and provides sample content.

2. **An XML document either linked to the data files, or embedded within it**
   If using an XML document then you are probably using a predefined metadata standard, e.g. Dublin Core. Dublin Core is one of the most common metadata standards, and may meet most of your metadata needs.

How do you prepare your data?

After collection but before analysis it's very likely you'll have to prepare your data so it can easily be examined. Once analyzed you'll also need to prepare it from publication.

Preparation means

1. Cleaning, coding, processing or otherwise readying for analysis
2. Transforming if dealing with PII or PHI
    - Deidentifying and/or Anonymizing

As with everything you need to document your preparation process.
At the data level this includes

1. Variable names
2. Codes
3. Definitions of terms
4. Transformation processes
5. File formats and software

At the Project level this includes

1. Collection context
2. Collection methodology
3. File structure
4. Validation
5. Manipulations

What about analysis?

After your data is prepared you are going to analyze it but there is more to this step than just running your data through a stats package, making visualization and generating comparisons.
You need to be aware of what is required for you or someone else to re-do your analysis.

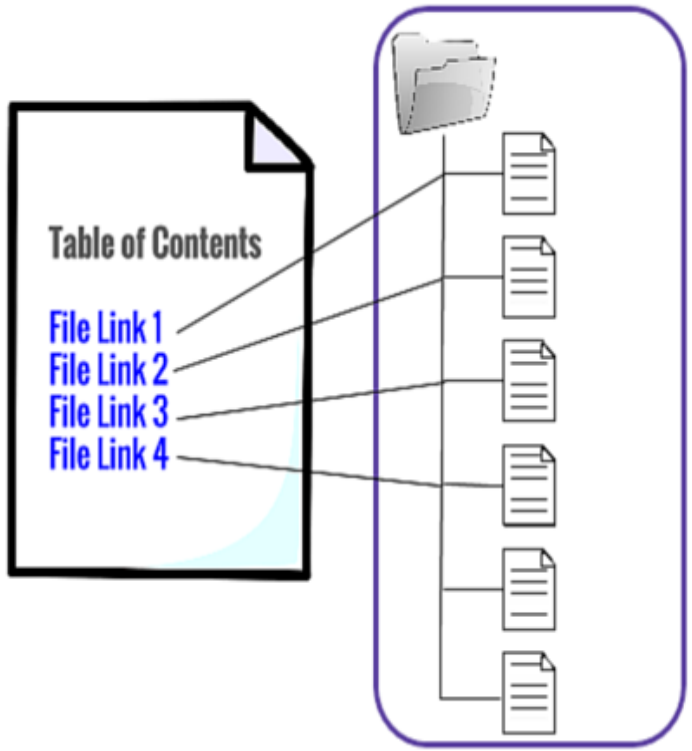
Remember **WHO, WHAT, WHERE, WHEN and HOW** from MetaData?

It's back for the analysis phase

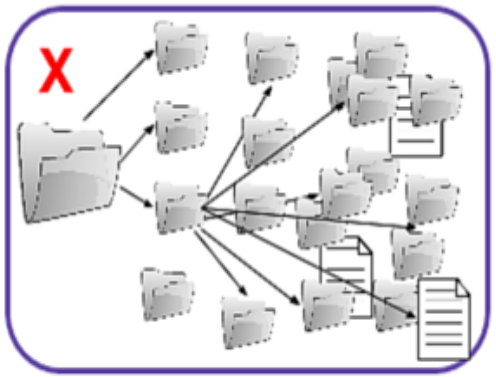
As part of documenting your analysis you also need to remember the following

1. Transparency
   - How and why you conducted your analyses
2. Analytic Outputs
   - Your analyses are like data, treat them as such, organize/save/store them as such

How do you organize your data?



AVOID TANGLED FOLDER NESTS



X /Project/Main/Initial Work/Experiment 1/Good/Results1/Keep These/January/Beginning of Month/Week 1/Saturday/Cycle1/0034tz.tiff

File Organization

Organizational points

Making sure your data is properly organized is just as important as the metadata added to it.

When organizing your data you should consider

1. **Names**
    - Consistent and concise
    - Descriptive
    - Unique to the content
2. **Structure**
    - Consistent across the whole project
    - Easy to navigate
3. **Connections**
    - The links between data should be obvious
        - Versions of the same file or different files related to aims or projects
4. **Documentation**
    - Your file structure should be documented
    - Data dictionaries, metadata and notes need to be part of your organizational thinking

How do you store or save your data?

Saving

Location

- Multiple copies, multiple places

Time

- Storing your data properly saves you time later

Format

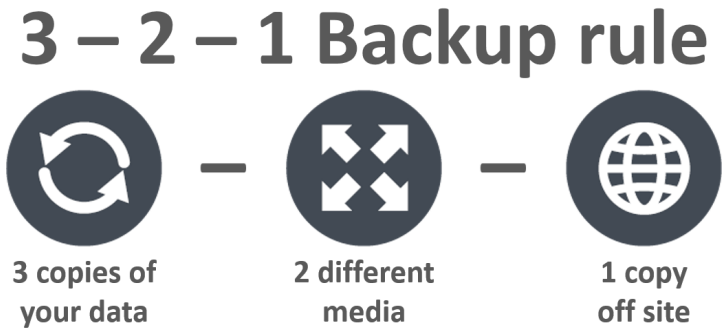
- Ensure the format you save your data in can be accessed later either by using open formats or properly documenting the formats

Where do you store your data?

Storing

Consider the following issues when considering Data Storage

1. Sensitivity of the data
    - Are you working with personal data/PHI
        - Deidentification
        - Secure storage
2. Failure rates of your storage
    - Physical media v cloud
3. The make up of the data
    - How much is sensitive v non sensitive
    - Transfer sensitive data

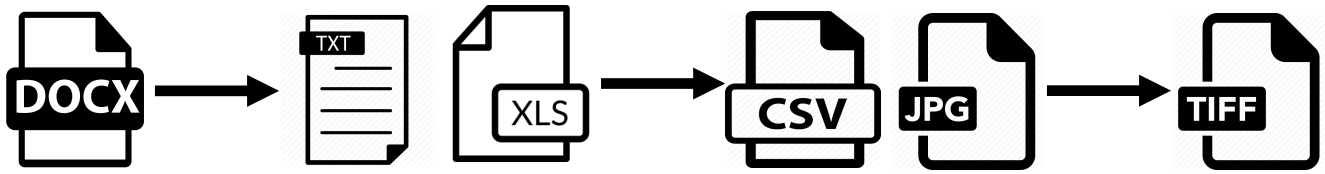


Ensure backups are made regularly
Make use of physical and cloud storage
Check on versioning of your files

How do you deal with Data long term?

Archiving

Preserve all of what you need
Use archival file formats
Make it as hardware/software agnostic as possible



How do you share your data?

Choose a repository that offers DOI's or other permanent identifiers

Make use of those identifiers to track how much your data is downloaded and re-used

Make sure your data is in usable formats

Ensure your data is complete

Make sure your data shared in a space that is appropriate for your field and conforms to the conventions of your field

How do I know what to keep?

If I wanted to use this data in 10 years, what would I need to do with it to make it useful?

What are the benefits of a Data Management Plan?

Increase the exposure your work gets

By using standardized data formats you are likely to get more citations and become more fundable in the long run

Increase the reusability of your work

Proper documentation via appropriate metadata makes it easier for other labs to understand your data and reuse it

Reduce duplication and increase effectiveness

Easy to understand datasets reduce the need to duplicate prior work and increases the effectiveness of follow up work both in your lab and others

Effectively preserve your data

Proper data management makes it easier to preserve your data in the long run

Keep your funders happy

More and more funders want to know where the data their money paid to generate lives and how it can be reused. A data management plan provides those answers

Make your data defensible

Good data management prevents issues around reproducibility and credibility

Support the Open Access/Open Science movement

Be good citizens of science, use proper data management

Data Management in Clinical Research?

The Same but Different

How is Data Management in Clinical Research different?

When conducting clinical research and clinical trials you have to be aware of the following

1. 21 CFR – Regulates electronic records in the US
2. GDPR – Regulates data privacy in the EU
3. HIPAA and the Privacy Rule
4. Data security
5. Appropriately designed Case Report Form
6. Controlled access database where data entry, validation and queries can be monitored

Governmental Regulations

21 CFR Part 11 – Regulates electronic records in the US for the purposes of clinical trials.

- This defines how electronic records and signatures are considered equivalent to paper version

GDPR (General Data Protection Regulation) – Regulates data protection and privacy in the EU

- Users of personal data must have consent to use the data and have appropriate measures in place to protect that data

Governmental Regulations II (because there are always more)

HIPAA and the Privacy Rule – Outlines the conditions under which PHI can be used in research

Applicable entities are permitted to use and disclose protected health information for research with individual authorization, or without individual authorization under limited circumstances outlined in the Privacy Rule.

These circumstances include the following

1. Documented Institutional Review Board (IRB) or Privacy Board Approval
2. Preparatory to Research
3. Research on Protected Health Information of Decedents
4. Limited Data Sets with a Data Use Agreement
5. Research Use/Disclosure With Individual Authorization
6. Accounting for Research Disclosures
7. Transition Provisions

What other best practices need to be followed?

1. Data Security
   Because clinical research often involves PHI Data Security is a foremost consideration.
   It is important to consider the following
   a. Who is the Data Custodian?
   b. What is the level of sensitivity of the data?
   c. How does the data “flow” and how is it transmitted?
   d. Who accesses the data
   e. How is the data stored, backed up and retained?

2. Case Report Form
   A well designed CRF is important for clinical research. A good CRF allows for the data to be collected accurately and without duplication

3. Database
   It is important to choose a database system that conforms to your needs and can capture the data you need for your study

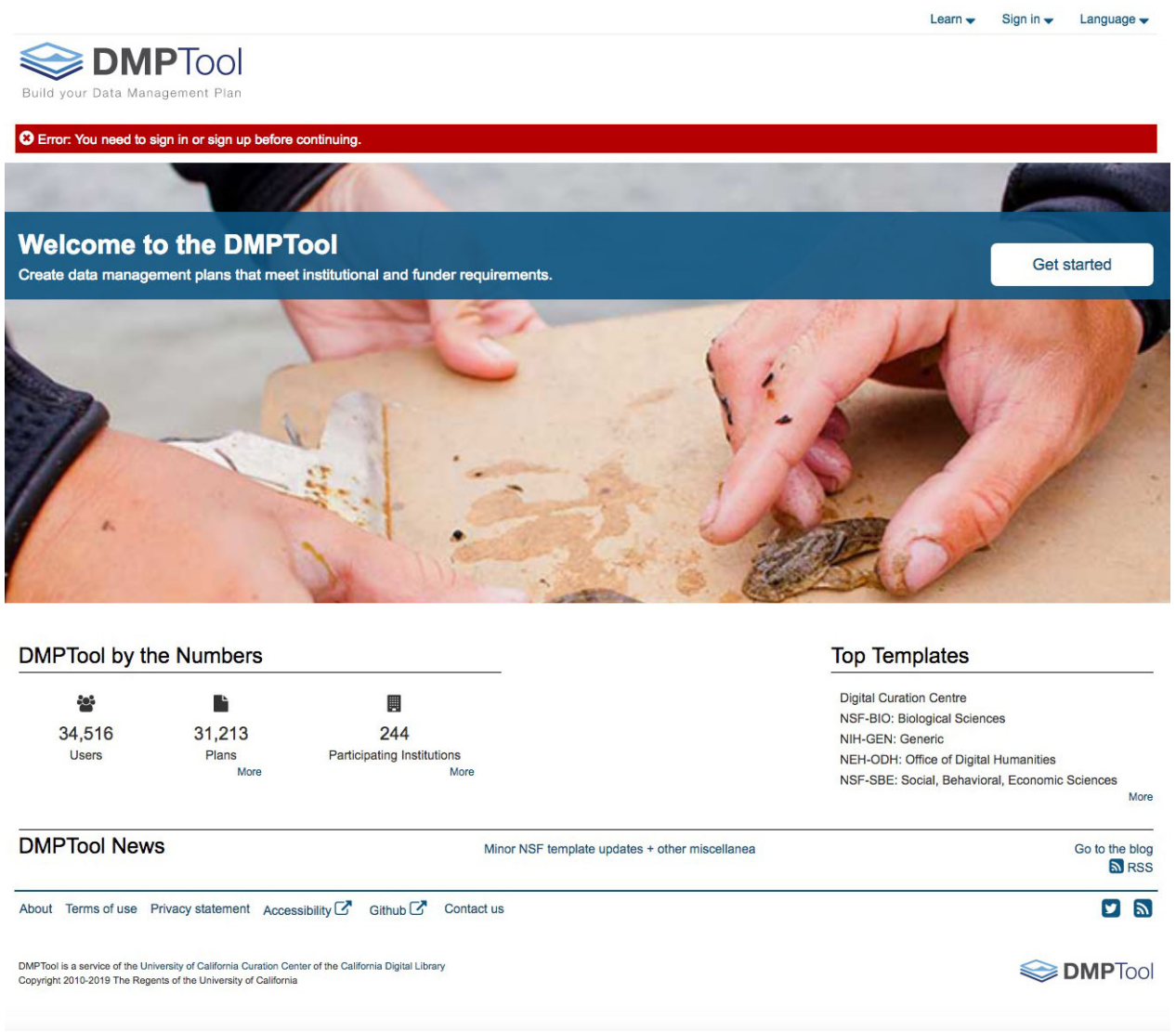
Tools for creating DMP’s

While Data management plans follow a set format and it would be relatively easy to write one from scratch it is better to utilize a Data management plan tool like DMPTool or DMPOnline.
These tools make it much faster, easier and more efficient to create a data management plan, to share it with collaborators, to store it for future reference and to transmit it to funders.

DMPTool

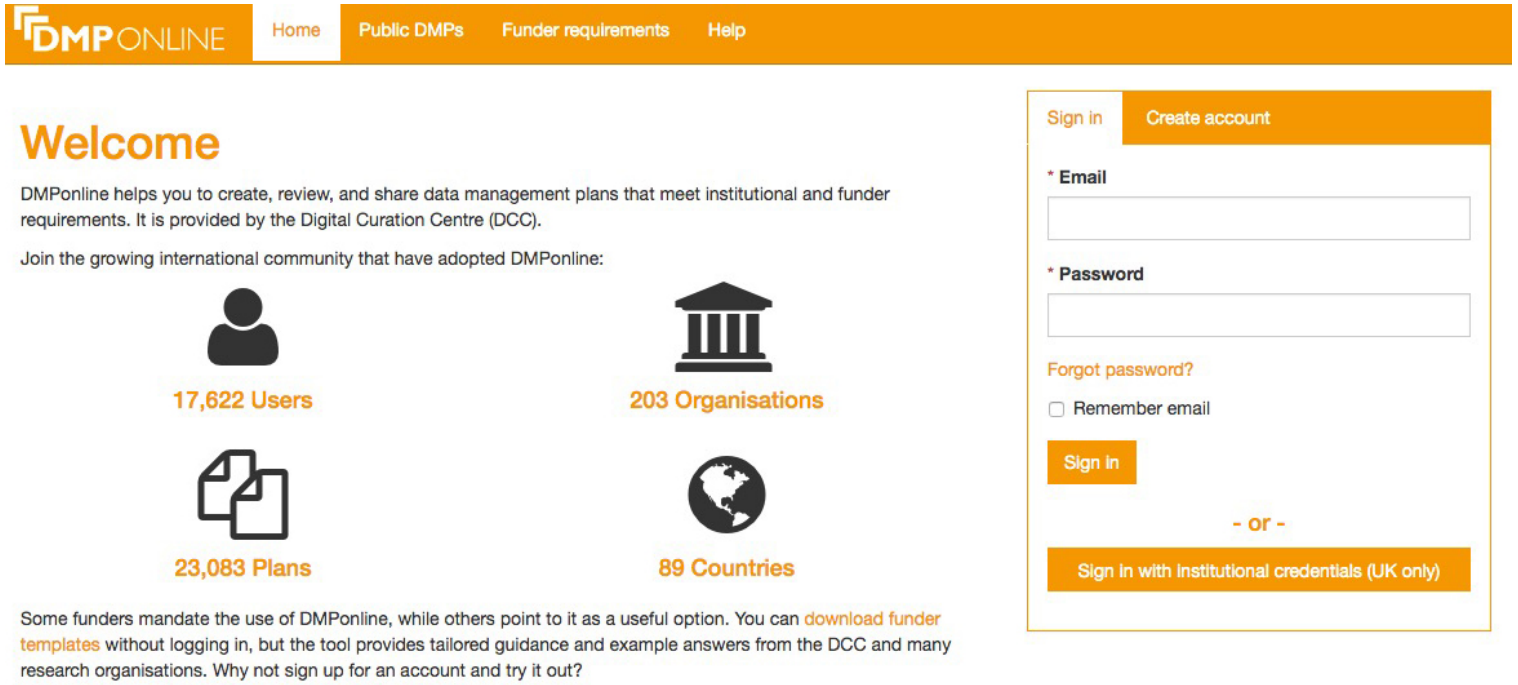
DMPTool is a great resource from University of California Curation Center of the California Digital Library.
It is a web-based tool to help you build your data management plan. It provides step-by-step guidance and information specific to many United States granting agencies and their directorates.

DMPTool



DMPTool can be accessed via creating an individual account and soon will be accessible via your RU login credentials.

DMPOnline



DMPOnline is another resource for building DMP's from the Digital Curation Centre (DCC).
It is a web-based tool to help you build your data management plan. It provides step-by-step guidance and other information.

Mantra

MANTRA provides excellent online training for Data Management.
It is web-based and 100% free and covers all aspects of Data Management.

Contact us

For more information on Data Management please contact the following staff at the Rita and Frits Markus Library

1. Matthew Covey - University Librarian
   mcovey@rockefeller.edu
   x8909
2. Rie Goto – Assistant University Librarian
   rgoto@rockefeller.edu
   x8980
3. Ilaria Ceglia – Science Informationist
   ilariac@rockefeller.edu
   x8944