**Data and Donuts: How to write a data management plan**

**Slide 1:**
Hi, and welcome to Data and Donuts. I’m Mara Sedlins, the data management specialist at Morgan Library. I’m trained as a research scientist, and part of my job involves helping researchers manage their research data. I am taking over this monthly workshop series that my predecessor, Tobin Magle, started in 2016 to raise awareness about research data management on campus. All materials will be made available online and are open for reuse as long as you attribute the source.

But before we get into the details, since this is my first workshop at CSU I’d like to take a moment to introduce myself and give you a little information about my background.

**Slide 2:**
My research training was in social psychology. I got my PhD at the University of Washington, where I studied the way people categorize faces that vary in characteristics related to perceived race, gender, and age. I also did some consulting work at Microsoft Research, applying social psychology research methods to the study of workplace communication technologies like 3D avatars.

After finishing my degree, I decided I wanted to pursue a non-traditional academic path and use my experience to help promote open science and better data management practices in researchers from a range of disciplines. Before coming to CSU, I finished a 2-year Postdoctoral Fellowship in Data Curation in the Social Sciences at Duke University, where part of my job involved teaching workshops similar to this one.

**Slide 3:**
Here is an overview of what I plan to cover today. And I’d like to acknowledge that I’m re-using a lot of the material that Tobin has developed for this workshop in the past – thanks Tobin! – although I’ve added my own perspective with some examples from the social sciences and some materials I’ve used in past workshops.

**Slide 4:**
To start, let’s define some terms:

What do you think of when you think of data management? In other words, what are you expecting to hear about in this workshop?

**Slide 5:**
Here is one definition:

**Data management** is the policies, practices and procedures needed to manage the storage, access and preservation of data produced from a research project. But mostly, it’s about having a strategy for how to keep your data safe.

**Slide 6:**
Even more fundamentally, what is data?

I’d like you to take a moment to think about the data that you work with in your research ... now, please introduce yourself to someone near you and tell each other what kind of research data you each work with.
Slide 7:
As you probably just discovered, research data are diverse!

Slide 8:
Here are some examples of types of research data. Note that:

- Data doesn’t have to be numbers
- It can be digital or analogue, born digital or digitized
- It can be observational, experimental, simulated, or derived/compiled

Slide 9:
A commonly cited definition of research data is the U.S. OMB Circular A-110, which defines it as “the recorded factual material commonly accepted in the research community as necessary to validate research findings.”

This definition hints at the diversity of research data that we’ve illustrated – what counts as data depends on the research community you belong to. But the common element is that research data is whatever you are using as evidence for your research findings.

This means that data are contextual – something that might be considered irrelevant noise by one researcher might be valuable data to another researcher depending on what their research question is.

Note that evidence collected that *doesn’t* support your hypotheses counts as data too.

Slide 10:
There is also one very controversial aspect of data that I’d like to get your feedback on.

(Poll about whether data is/are singular or plural)

Slide 11:
I also want to issue this disclaimer: **Data management** is not the same thing as **data sharing**, however the same principles apply to both.

- I'll often describe data management best practices in the context of allowing someone who has never seen your dataset before being able to use it based on how it’s described in the supplementary material.
- This mindset is useful even if the data are never widely shared because "you are the future user of your own data", not to mention others in your research group.

Slide 12:
So why should you care about data management?

- Well, you already care about your data.
- The image on this slide is a researcher submission from the "Day of data" at Brown University, where researchers were asked to submit post cards with their feelings about data.
- It illustrates that data elicit a range of strong emotions, from fear to happiness. This response is understandable, because data dictate your success or failure as a researcher.
Slide 13:
Another attribute of data is that it tends to get messy

- Most of us just don’t realize this because our messy, disorganized files are locked up in a neat little box called your computer.
- Don’t believe me? How long would it take you to find a photo from five years ago on your computer? Here is a hint. If your image files start with DSC_ or IMG_ and some number following it, it will probably take you a very long time.
- If most people’s digital files were analog, this is exactly what they would look like.

Slide 14:
But taking care of your research data has changed recently because everything is digital. Preserving digital data takes a different set of skills than preserving a physical object like a lab notebook:

- digital objects are a lot easier to lose. A hard drive failure is a lot more common than the types of natural disasters that can wipe out a physical record.
- datasets are huge and complex, making them harder to keep track of
- We have the ability (and some say obligation) to share primary data

Thus we need new skills.

Slide 15:
The availability of research data also tends to degrade over time:

- only about half of the corresponding authors on manuscripts have a valid email address listed.
- Even if the email address works, the author only responds about half the time.
- Even if you get a response, only about three quarters of the authors will even mention the status of the data.
- Even then, only a quarter of the data is accessible by 20 years after publication.

Slide 16:
And so, we are losing vast amounts of research data.

Slide 17:
In the midst of all this, research funding is tight. Despite the number of PhDs increasing every year, research funding is essentially flat (with the exception of NIH biomedical research).

Slide 18:
In order to make funding go further, all federal funding agencies have recently completed plans to make results of federally funded research available to all.

Slide 19: 2003 NIH policy

Slide 20: 2011 NSF DMP requirement

Slide 21: 2013 White House OSTP memo
The number of federal agencies that require data sharing plans has been increasing. SPARC (the Scholarly Publishing and Academic Resources Coalition) has been maintaining a list.

And private funders do too. For example, the Gates foundation has one of the strictest open access policies, requiring results to be immediately available as of Jan 1, 2017.

Increasingly, journals are enforcing data sharing policies too, as a condition of publication.

But even if you're not subject to any of these requirements, data management still has its benefits:

- It improves research reproducibility by providing transparency
- It improves efficiency through organization and data reuse
- It spurs innovation by allowing ideas to be shared more freely

Good data management practices are also good for you because

- You are the future user of your own data, and the you of 5 years ago is really bad at answering email
- Your data get reused and cited along with your papers
- Collaborators get to see your work first hand
- And grant reviewers take DMPs seriously

Data management & the research cycle

When does research data management come up for you?

It's not something that should be left to the end of the research project: it occurs continuously throughout the research cycle.

Generally the research process goes as follows: You start with a hypothesis

Then, you design experiments to assess the hypothesis

Then you collect the data

Analyze the data to produce results
Publish these results in a research article.

And these findings can be used to generate new hypotheses.

This simple cycle is more complicated now that we have the ability and mandate to share our research data. We need to write data management plans even before data collection begins.

And with the large amounts of data collected, cleaning and analysis of the data are bigger jobs that are made easier by automation by scripts.

We have to take special care to archive the data properly and share digital data so that it is usable into the future.

On the bright side, technology also allows us to publish and get credit for not only for research articles, but associated research data and code. This is central to the concept of reproducible research.

All of these outputs can be used to generate new hypotheses.

This session focuses on the first data management activity in the new research lifecycle: writing a data management plan.

Exercise: DMPs deconstructed

A data management plan is a description of how you plan to describe, preserve and share your research data. They are often required for grant applications.

While the specifics vary by funder, successful Data Management plans include:

- A data inventory
- A strategy for describing the data
- A plan for preserving the data
- And a method to access the data.

Always make sure to check your funder’s requirements.

Let’s start with the contents of the data inventory section. This section should include:

- The type of data you will collect
• The file types that will be produced
• How the data may change over the course of the research
• The size and number of the files
• And what other research products will be produced, such as code and templates that might be useful to other researchers

Slide 46: Data inventory example 1

Slide 47: Data inventory example 2

Slide 48:
While building your data inventory, it's important to consider what types of files you store your data in. This is important to ensure that the data are accessible into the future.

• If you can, convert your data into non-proprietary formats like text or .csv files. This concept also applies to more complicated data formats like images, audio and movies, which can be stored as .tiff, mp4 or mp3 files respectively.
• For example, you could store tabular data in Excel, but what happens if Microsoft changes their software so that it doesn't open older file versions? This is a problem with all proprietary formats.
• Also, know what type of software can read your data and put that information in the readme file that we'll talk about later
• For a more comprehensive list of recommended file formats, see the Library of Congress Recommended Formats Statement.

Slide 49:
How much will your data change over the course of the research? Will it continue to grow? Will you be making revisions to it over time? This might suggest the need for a data versioning strategy.

You can take different approaches for file versioning. The most common and simplest way to do that is through numbering. A good rule to allow for greater scalability and to ensure files are displayed in the right order is to use leading zeros.

Slide 50:
• In addition to assigning a scalable number, you can also follow the common practice of appending a v1, v2, v3 for major revisions to files and using decimals for more minor changes. This convention is most often used for software releases.
• You can also use dates as a versioning system for your files if that would be most appropriate.

Slide 51:
If you're using dates in your file naming convention, I highly encourage you to use the standard ISO 8601 format, that begins with the year, followed by the month, followed by the day. This is used internationally and makes it easier to sort your files by date than the bad examples shown here.

Slide 52:
• If you need more sophisticated version control, there is software you can use to track changes across versions and make collaboration easier. The Git platform seems to be the most
commonly used – it underlies GitHub, a popular file sharing site that you’ve probably come across.

- I don’t have experience with these others, but Subversion, Mercurial, and Bitbucket are other popular versioning tools. And certain storage and file sharing platforms, like Box have a built-in versioning system.

**Slide 53:**

**Exercise 1:**

Let's take a moment to write a data inventory. Open a text editor, and describe:

1. The kind of data you'll collect
2. What file type you'll produce
3. Whether you’ll need a version control plan for growing or changing datasets
4. How big these files will be
5. And any other research outputs you will produce

**Slide 54:**

The next aspect of data management that you should consider is how you will describe your data. This description is commonly referred to as metadata, or data about data.

- At the minimum, your metadata should include contact information for the researchers who produced the data, a description of how the data were collected, when and where the data were collected, and the units that are being measured.
- The format can be as simple as a text file.

**Slide 55:**

You can think of metadata as descriptive information at a couple of different levels.

- Study level documentation explains the broader context of the specific study. Some of the elements of your study-level documentation include explicit details about the source of your data whether you are using existing data or are collecting your own.
- For existing data, you would want to list where and when you found the data – using a full citation where possible as that will come in handy when you are writing up your results for publication.
- For data you collected you would have information on your data collection methodology which could include sample design, interview schedule, instrument calibration, key dates, weighting, etc.
- You’ll also want to provide information about how your different data files and documentation interact with each other. You should include information about measures used in your analysis. Finally also having some details on the programs you used to process or manipulate your data, and annotated scripts used in those programs should also be included.
• Your study level documentation answers the contextual questions - where the data came from, how your files were constructed, and how you processed and manipulated the data to transform it from raw into final analysis files.

 Slide 56:
Your data-level documentation will describe the information that other people will need to actually use and understand what is within your data files. This may include your variable names and descriptions of those variables, information on your units of measurement, variable type information, value labels so what do those 1, 2s, and 3s actually mean. If you have any component or derived variables you will want to describe how these variables were constructed and then also keep the original variables used in creating them. So for instance a component variable might be a variable on BMI (or body mass index) that is computed from height and weight variables – you would still keep those variables as well. And finally you will want to define any missing data values. There are different ways you can define missing values – the one thing you definitely do not want to do is just leave fields blank.

 Slide 57:
• There are many different ways you can format your documentation. The most common is the README file which is a very flexible format that can be used at multiple levels of description. For example, you could have a README file for your whole research project, in addition to separate README files for each dataset.

• A tabular codebook is something that is commonly used for numeric data at the data file level – often study level information will be included as the front matter to the codebook. A codebook includes variable type information including detailed variable and value descriptions. Some codebooks will include copies of instruments used to collect the data with them as well. In the social sciences, aggregate information such as frequency of a response are often included in the text of a codebook.

• For studies that do not include a lot of variables, you might use an annotated instrument to provide variable and value information.

• And finally there are formal metadata standards that follow a specific structure based on discipline or file type that help with file interoperability and discovery.

 Slide 58:
As a genomics example, my predecessor, Tobin, created a readme file that describes

• what the data are (mouse miRNA sequences),
• how it was produced (RNA extracted from mouse brains, then sequenced)
• Links to where the data can be found
• And contact information

 Slide 59:
• Here is an example of a tabular codebook implemented in Excel. The codebook includes columns for key attributes of the data, including variable name, type, and length; the parent measure, the question text, and the possible response values. One advantage of creating a codebook in Excel, as opposed to Word, is the ease of adding additional columns as needed. You
may set up a codebook at the beginning of a project that contains all the pertinent information at that time; however, as a project progresses, there may be additional details that come up over time.

- If you use Excel for creating your codebook, you would also want to consider saving the final version in a more preservation friendly format such as a CSV file for long-term preservation.

**Slide 60:**

- Here is an example of an annotated instrument. This document has the original instrument and includes the variable names and identifies the values associated with each response. It uses the color red to differentiate the annotations from the original instrument.
- Once data has been collected, the annotated instrument can also be updated with information about response frequencies if the layout allows sufficient room.

**Slide 61:**

If you want your data to be integrated with other similar data, it's useful to use metadata standards.

- Dublin Core is a generic metadata standard that covers the elements listed on the last slide.
- Many disciplines have field-specific metadata standards, like EML for ecology, MIAME for microarray experiments, or the DDI for social science data.
- If you're not sure what the common metadata standard is for your field, you can use the Digital Curation center or FAIRsharing to find metadata standards.

**Slide 62:**

**Exercise 2:**

Let's take a moment to decide on a metadata standard and add it to the text file with your data inventory. Important questions to ask yourself are:

- What do people need to know to reuse your data? Like I said, even if it will never be shared widely, this is a good head space to be in when deciding how to describe your data.
- Are there any standards that are common in your field?
- What format will the metadata be in?
- What metadata fields will you include?

**Slide 63:**

Now that you know what data you will have and how to describe it, you need to think about how you will provide for its safety over the long term:

- What are you going to do to ensure the data are stored properly and preserved?
- What metadata or other products need to go with it?
- Also note that your data preservation techniques will change over the course of the project, so account for all stages. We'll talk about this a little bit more later.

**Slide 64:**

When thinking about backing up your data, here are some things to consider:
• Storing things in geographically distinct locations can be important in the case of natural disaster. We know CSU has had issues with flooding, so backing up your data on an external hard drive that is next to your computer isn't a good solution for this problem.
• Cloud solutions can be advantageous here. CSU provides researchers 1 TB on Microsoft OneDrive, and allows the use of other systems like Dropbox and Google Drive.
• Is your system automated? A backup system is only as good as the last time you synced, so make sure you're saving early and often.
• Finally, not every solution is secure enough if you're dealing with private information. Generally, things like Dropbox are not certified secure for private data and need to be stored locally.

Slide 65:
When thinking about preservation, you should ask yourself:
• What will you store?
• Who will be in charge? This is especially important when backing up manually
• How long will you store it? - Often determined by granting agency, rule of thumb is 5-10 years
• Where will you store it? - For example, during the project, you'll probably have your data stored on your machine and backed up on a departmental server or the cloud.
• After the project is done, it might be easier to have it stored in a repository or on a less accessible central IT server to make room for new data and so you don't have to worry about remembering backups.

Slide 66:
Exercise 3: Let's take a moment to write out your preservation plan. Open up your document and answer the following questions:
• What will you store?
• Who will be responsible for the data? This can be a specific person, or a position
• How long will you store it?
• Where will you store it at different stages of the project?
• And how will you back it up?

Slide 67:
Now that you know what data you will have, how to describe it, and how to preserve it, you need to figure out how others will be able to access your data.
• This is particularly important to funding agencies because they want the public to see what their taxes are paying for.
• Your data must be easily available: saying "will be made available upon request" doesn't really cut it anymore.
• However, it is accepted to embargo your data for about 12 months after publication or project completion.
• As always, if you're working with private data, you have to consider security when assessing this plan.
Slide 68:
- When sharing your data, it's best to use non-proprietary formats, because not everyone will have access to the same software that you do.
- Including metadata is also essential so others can make sense of your data and use it responsibly.
- As always, you want to make sure the data is stored properly with a backup system in place.
- A good place to do this is a trusted repository.

Slide 69:
Trusted repositories have a dual purpose: storage and sharing.
- Optimally, you would put your data in a discipline-specific repository so others in your field know where to look for it.
- You can use re3data to look for repositories in your field.
- If you do not have a repository that fits your needs, Colorado State has its own digital repository that can store research data.

Slide 70:
To see datasets that are in the repository, you can go to the repository and look in the "data collection".
- It's a flexible system that uses Dublin Core metadata, but other standards can be integrated as needed.
- Also, it's at no cost for less than a TB of data.
- Above that, the cost is $150 per terabyte for 5 years, or $300 per terabyte for longer than 5 years.

Slide 71:
Trusted repositories also provide what are known as stable identifiers. We all know the URLs break, so stable IDs were created to prevent links from breaking over time. You probably have seen DOIs, or digital object identifiers, on journal articles, but they can be applied to any digital object, including data.
To find a digital object, search for the DOI at dx.doi.org, which will direct you to the specific digital object on the web. If the object's URL changes, the DOI is updated, preserving the association. CSU's digital repository provides another stable identifier called a handle by default, but can mint DOIs for data by request.

Slide 72:
Even if you share your data, you don't need to do it without restrictions. You can state your conditions for reuse in a license. A common condition for reuse is that users cite your data and/or associated publication. You can also issue disclaimers about your dataset. However, you must justify your usage limitations. A good starting point for licenses is the creative commons web site, but we won't go into that here because that could take an entire session.

Slide 73:
Exercise 4:
You're almost done with your data management plan! Take a moment to consider your data sharing plan with the following questions:
- Where will people access the data? Does your discipline have a repository?
- What kind of stable identifier do you want?
- What are the conditions for reuse?
- What limitations do you want to place on the data and why?

**Slide 74:**
More resources: the DMPTool

**Slide 75:**
A good place to look for funding agency requirements is the DMPTool, an online tool for data management plans. You can review funder requirements at the link listed here. You can also create an account to use DMP templates and search public DMPs.

**Slide 76:**
As always, if you need help with any of these topics, my contact information is here, or you can consult the DMPTool or the CSU libraries data management web site. Thanks for listening!