Slide 1: Hi and welcome to data and donuts. I'm Tobin Magle the Data Management Specialist at the Morgan Library. Today we're going to be talking about data archiving and sharing.

Slide 2: This is the final session in the data and donuts series, which corresponds to the end of one turn of the research data life cycle: how to preserve and share your research data at the end of a project.

Slide 3: In this session, we'll discuss the things you need to do to

- Preserve your research data
- Describe your data for findability and reuse
- and share your data with other researchers while adhering to the FAIR principles

Slide 4: To put this topic into context, we're going to start with a video that illustrates why this topic is important to your research, as a data producer and a data consumer.

http://www.youtube.com/watch?v=N2zK3sAtr-4

watch the video linked to on this slide, and then return to this presentation.

Slide 5: Let's jump right in with an exercise:

Exercise 1:

- make a list of the bad data management decisions you saw in the video
- then, keep an eye on this list during the presentation to see what the researcher could have done better.

Slide 6: Let's start with preserving your data. There are three main topics to address when you're getting ready to archive your data:

1. Where will the data be stored?
2. What format should you store them in? and
3. How will you describe your data so that it makes sense to you or new users in the future.

Slide 7: When thinking about preserving digital data, you have to separate short term from long term storage

- During the course of the project, data changes frequently, and it's almost entirely up to the researcher to take care of things. See the Reproducible research lesson to learn about tools to manage this process.
- However, storing data long term after the project is over is a little different. Because the data are not being altered and don't need to be accessed frequently, it's possible to outsource the responsibility. For example, you could store it (and share it) in a trusted repository.

Slide 8: But before we talk about where to share it, let's go over some basics of good data preservation, such as

1. Back ups
2. Archival formats and
3. Metadata

Slide 9: When backing up your data

- First, you want to have 3 copies of the data.
- One copy should be in a geographically distinct location to guard against loss due to events like natural disasters.
    - At CSU, we just observed the 20th anniversary of a flood that caused massive damage on our campus during which a lot of resources were lost.
    - Other data librarians have told me stories of data centers destroyed by hurricanes or tornadoes
    - However other mundane things like, spilling coffee on your laptop, is also a danger.
- An example of this type of robust backup system is storing the data on your computer's hard drive, an external hard drive and in the cloud.

Slide 10: Now let's talk about data formats.

- We recommend using non-proprietary formats
- These formats will increase the chances that you'll be able to use the data into the future
- Proprietary formats change at the whim of the company that produces them, as you probably know if you've ever tried to open an old Microsoft Excel spreadsheet.

It's also important use data standards that are common in your field. For example, if you're working with DNA sequence data, you'd probably save it in FASTA format, rather than a Word document.

- One place to look at data standards that meet both of these guidelines is FAIRsharing.org, which provides support for sharing data according to the FAIR principles. We'll be referring back to this resource and the FAIR principles throughout the presentation.
- Let's take a look at standards in FAIRSharing.org.

Demo 1:

- Go to fairsharing.org
- Select standards
- Select Model/Format
- Select image under domains
- See standards like png and jpg, and point out DBs that use them

Slide 11: Here's a list of common proprietary formats and alternatives. You can see that there are formats for things like word docs, powerpoints, images, movies and sound. But probably the most common format I see when curating data are excel files.

Slide 12: Using Excel sheets during the research process is easy and convenient, but it's even more convenient if you follow some simple guidelines that will make your data easier to preserve later.

- First, only include 1 table per sheet
- Also, make these sheets csv friendly by not including formatting, images and formulas, or at least making them so that they're easily removable.
- Finally, it's important to make your data tidy, which means that every column is a variable and every row is an observation. See the data organization lesson for more information on this topic.

Slide 13: One way to get excel files into archival formats is Excel Archival tool, created at the University of Minnesota. This Visual Basic Script takes an excel file as input, and separates the content into 3 nonproprietary formats

- .csv for each tab
- A text file containing formulas
- An HTML visualization of what the original spreadsheet looked like

Let's look at the output from excel archival tool

Demo 2:

- Open excel doc
- Go over the folder of excel archival tool output
    - View csvs
    - Look at text file with formulas
    - Open HTML viz

Slide 14:

Exercise 2: Think of one type of data that you produce

- What format is it in?
- Is the format proprietary? If so, what alternative format could you use?
- Is it the most commonly used data format for that type of data you're producing?

Slide 15: Now that we have our data in good, archival file formats, let's talk about how to describe the data.

Slide 16: We call content that describes data metadata, which literally means data about data

1. Metadata encompasses all relevant information for discovery, recreation and reuse of your data. Metadata comes in two general types
2. Descriptive metadata are what data users need to know to use your data.
3. Discovery metadata is the metadata that repositories use to make your data discoverable via search services like google scholar.

Slide 17: Let's look at descriptive metadata first.

- Descriptive metadata provides context and tells you everything you need to know to interpret and reuse the data responsibly
- Two general types of descriptive metadata are Readme files and codebooks.

Slide 18: Readme files give a broad overview of what the data are, how they were collected and list the files that are included in the dataset.

- For example, the readme file on the right describes weather data measured in the Fort Collins area from the late 1800s to the present.
- It describes what variables were measured and the exact geographic location of the measurements in lat, long for each year.
- The advantage of the readme file is that it's human readable and can accommodate any dataset, however, the lack of structure reduces its machine readability making it insufficient for discovery.

Slide 19: Another type of descriptive metadata is a code book.

- Codebooks are more structured in comparison to readme files.
- There's typically one codebook for one type of sheet. For example, if you have a bunch of sheets that use the same variables, you can use one codebooks.
- They typically have a row for each variable in the data sheet, and
- columns for
    - variable names,
    - variable definitions including units,
    - and what the possible values are, including any special formating and missing or coded values.

Slide 20: While readmes and code books can be used for any dataset, some fields and data types have discipline-specific metadata standards,

- which specify required and optional elements,
- and a specific format to put them in
- however, these standards are not available in every field
- You can look for metadata standards at the Digital curation centre or fairsharing sites.

Slide 21: Exercise 3:

- Think of the data you chose to work with in Exercise 2
- What information would you include a README file? A codebook?
- Is there a metadata standard you could use?

Slide 22: Those are the basics you need to preserve your research data for future use. Now let's start to think about sharing our research data in the context of the FAIR principles.

Slide 23: The FAIR principles were developed to provide general guidance about effectively sharing research data. The FAIR principles stipulate that the data (or at least the metadata) must be

1. Findable by search and through a unique ID
2. Accessible in a repository
3. Interoperable with data of the same type and
4. Reusable by having proper metadata and a license for reuse

Slide 24: We're going to step backwards through the FAIR principles to Illustrate the fact that most of the work required to be FAIR is completed by archiving the data for yourself. When we think about reusability

- We need metadata to put the data into context, which we've discussed
- File formats that can be opened, which we've also discussed
- And only one new element to think about: a use license

Slide 25:

- Licenses allow you to state your provisions for other people to use your data, typically things like citing either the dataset itself or an associated manuscript
- Creative commons licensing is a good starting point in general, but in the United States, data isn't copyrightable, so it's probably best to apply a CC-0 license to put your data in the public domain

Slide 26: Now let's talk about interoperability. This provision allows your data to be combined with other similar data for metaanalyses. This is where community standards come into play

- For this we need comparable metadata and file formats, which we've already addressed.
- One element that we haven't discussed are Controlled vocabularies, which are specific languages used to describe data

Slide 27: Controlled vocabularies are essentially official names for things, like a list of airport codes.

- Ontologies are a type of controlled vocabulary that contains relationships between entities.
- A good example is taxonomy, where many species have multiple names at one point, but have been unified into one hierarchical system.
- For example, we have the official species name Homo sapiens for humans but another entity called mammals. We could connect these two with a link that tells us that Homo sapiens are also mammals

Slide 28: Exercise 4

- Think about the data you've been using in these exercises
- Is there a standard ontology you could use to describe the data?

Slide 29: Now let's talk about accessibility. This is where we're getting into some new territory with

- Repositories, which hold data and/or metadata
- And Persistent identifiers, which give your data a unique, citable label to make it easy to locate

Slide 30: First let's talk about repositories. These databases provide

- A place to put data and metadata
- A unique ID for each dataset

- A way to find the data
- And specific discovery metadata requirements

Slide 31: There are many discipline specific repositories out there. A good place to look for these databases are fairsharing.org, which we've been looking at for standards, and re3data, which contains only repositories

Slide 32: A CSU Specific option is to use our digital repository, which already contains over 100 datasets.

- It uses Dublin core metadata for discovery,
- but supports other descriptive metadata types (like readmes and codebooks), which are provided in a zip file with the data.
- Also, it's free unless a single dataset is over 1 TB in size.

Slide 33: Another feature that affects accessibility are stable identifiers.

- These are an improvement over URLs because they are permanent,
- but are often still linkable through a web browser.
- Common unique identifiers that fit both of these criteria are
    - Digital object identifiers (DOIs) and
    - handles, which we use in CSU's digital repository.

Slide 34: If you want to put your data in our repository, it's not (yet) and automatic process.

- Contact me if you're interested
- I can help you get your data and functional metadata for good shape for submission and write the Dublin Core metadata for discovery.
- Then you can agree to our deposit agreement,
- and we can upload and give you a handle for your data! DOIs are also available upon request.
- If you need it for a manuscript, we can give you a handle ahead of time to put in the paper and put the data in later.

Slide 35: Finally, let's discuss findability. The only factor that we haven't discussed is discovery metadata.

Slide 36: Discovery metadata is intrinsic to the repository system that you're using. As such, each repository has a defined standard that it uses to populate the metadata and make your data findable.

Slide 37: We use the Dublin core discovery metadata standard in the CSU digital repository because it can be applied to anything.

- There are 15 core elements, or fields to fill out describing your data
- These elements have qualifiers that make the terms more specific.
- For example there's a dc term for contributor, that has a qualifier called author, which is a specific type of contributor.

- If you put your data into CSU's repository, you don't need to worry about writing the discovery metadata yourself. We'll help you do it.
- Let's look at an example dataset that's already in the repository.

Demo 3:
- Go to https://dspace.library.colostate.edu/handle/10217/182733
- This is data submitted as supplemental information for a manuscript
- The abstract for the paper appears as the item description
- There's a list of contributors, a publication year, and a link to where the data live within the repository, in this case, it's the data collection
- The Unique ID is the handle on the left beneath the thumbnail
- Show fill item record.
    - Authors, Date, abstract, unique ID as seen on main page
    - Extra: sponsorship, data type, language, publisher
- If we want the dataset, you can click on the view/open link
    - Unzip the file
    - Look at the structure, point out codebook and readme

Slide 38: Here's our final exercise;

Exercise 5: How FAIR is your data?
- Is it findable?
- Is it accessible?
- Is it interoperable?
- Is it resuable?

Slide 38: So in summary,
- that most of the work that it takes to share your data are things you should be doing to preserve your data for your own personal use.
- Sharing your data in a repository makes your job even easier by outsourcing the preservation responsibilities.
- Finally, this is all really complicated and context dependent, so don't be afraid to ask for help!

Slide 39: Thanks for listening. Make sure to contact me at tobin.magle@colostate.edu with any DM questions, or see our DM services site to see what we offer. There are also links to information about file formats, excel archive tool, Dublin core, and our repository.