2023-01-17

Digital Scholarship 101: Managing your data

Woodbrook, Rachel; Carruthers, Matthew

https://dx.doi.org/10.7302/6558
https://hdl.handle.net/2027.42/175009

Digital Scholarship 101: Managing your data

January 19, 2023

Matt Carruthers
Rachel Woodbrook

Caitlin &/or Joe

Caitlin &/or Joe?

Turn on live transcript

[**Add link to chat**:
https://docs.google.com/presentation/d/1kBIQGaxFHrFWPLVSbm1mDe_sD0sUptzd9tSEy7dgqqc/edit?usp=sharing]

DS 101 Workshop Series

http://myumi.ch/ds-101

Understanding Accessibility for Digital Projects (February 7)

Advocating for Your Digital Project (March 8)

Caitlin &/or Joe

Facilitator Introductions

Matt Carruthers
(He/Him/His)
Metadata Engagement Librarian

Rachel Woodbrook
(She/Her/Hers)
Data Curation Librarian

Rachel & Matt

Names, pronouns, titles...interesting fact?
Recording

Attendee Introductions

(2-3 min.)
Rachel

Name, pronoun (if you want), one thing you are hoping for from this session - in chat

Goals/Agenda

- Introductions
- Goals
- Data & data management planning
- Activity/Break
- Overview of data management topics
- Questions/Discussion

Kristin Briney
@KristinBriney

Data management is hard and everyone is bad at it. This includes data managers more often that we care to admit. Data management is basically fighting against chaos.

(5 min.)
Rachel

Here's our agenda for the next hour or two.

We've done introductions already; in terms of goals, we want this workshop to be as useful as possible to those of you here today. We expect that there is a wide range of needs and expectations across the group, and data management isn't one-size-fits-all--it's contextual, based one what you are working with, and what you are trying to accomplish. We will talk about a lot of different possible considerations in the next hour or so. One takeaway we hope you get from this session is that **it's fine to start by choosing one or two areas you feel are most important that will make your life easier or your work better, rather than trying to tackle everything at once.** In order to build new habits and skills, it's important to start somewhere manageable.

So we'll start by taking a step back and talking about data--what it is and how it is situated in reality--and then about some of the benefits and considerations for data planning regardless of the type of project you're

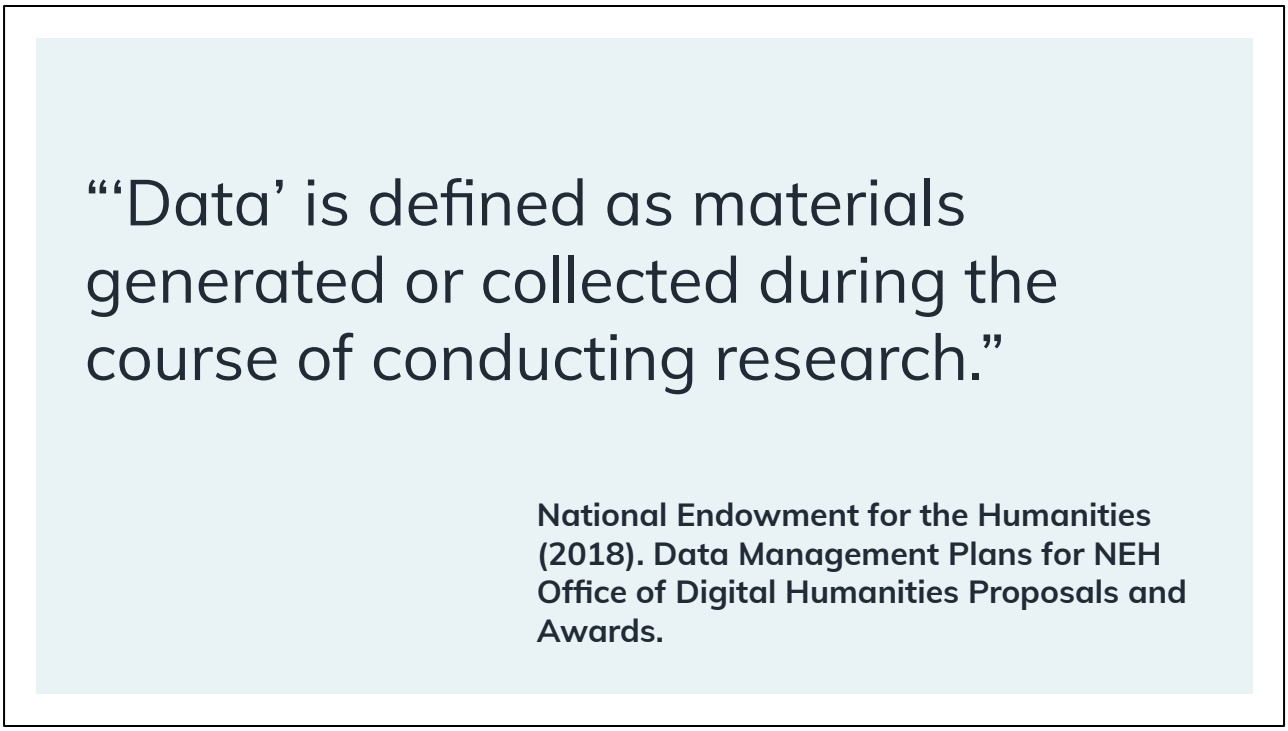
doing.

Next we'll give you some time and a few tools to start thinking about how some of the topics we're covering today relate to your own work, and what you might need to focus your planning effort, wherever you are in your process. ***How many of you here are in the workshop because you have a specific project in mind that you know will be using data?*** *Raise your hand if you are able*

We'll follow that with a more detailed but still high-level overview of the main concepts behind data management, with some specific examples along the way. If particular steps or sections are especially relevant for you, let us know and we can dig a little deeper in those areas.

We've planned for time at the end for questions and discussion, but we're flexible depending on how you would like to interact. There are a few too many of us to just unmute, but please feel free to raise your hand or ask questions in the chat as we talk, we'll try to keep an eye on it and address things as we go when that makes sense.

Are there any questions or concerns you'd like to address before we start?



(2 min.)
Rachel

So--let’s talk about data! Defining data may seem basic, but we don’t want to assume everyone is starting on the same page. We probably each have our own working definition, and I’d be willing to bet not all of our ideas cover the same things.

Just as data practices are dependent on context, so is the understanding of what data is (or are). For instance, if you’re applying for funding, it’s especially important to know how the funder defines data. Effective communication is also important, and it’s worth thinking about whether you need to intentionally build a shared understanding around vocabulary with colleagues or your audience, especially if you’re doing interdisciplinary research. (For example, people may have different understandings of what is meant by “corpus”, or “research consent”).

On this slide is one (very broad) definition from the NEH’s 2018 guidance on data management plans. *[Read slide].* Notice that this could cover almost anything! In some ways it has more to do with where the

material originates than what it is--and this breadth makes sense given the range of possible data types in the humanities.

(Unfortunately this may not be current and there isn't an updated version available, but it's still a good illustration for our purposes.)

"[...] Examples of humanities data could include citations, software code, algorithms, digital tools, documentation, databases, geospatial coordinates (for example, from archaeological digs), reports, and articles. **Excluded, however, are things such as preliminary analyses, drafts of papers, plans for future research, peer-review assessments, communications with colleagues, materials that must remain confidential until they are published, and information whose release would result in an invasion of personal privacy** (for example, information that could be used to identify a particular person who was one of the subjects of a research study)."

National Endowment for the Humanities (2018). Data Management Plans for NEH Office of Digital Humanities Proposals and Awards.

(2 min.)
Rachel

This is a continuation of the same definition. I'm not going to read it all, but the point I wanted to make is that they go on to specify examples of what is and isn't data--and interestingly, they *exclude* personally identifiable information from their definition, which most people would agree counts as data. However in this case, they are defining data broadly in terms of format, but also very specifically in terms of possible impact, and in terms of the actions they are expecting to be taken (namely, data sharing).

“[...] research data are the recorded factual materials commonly accepted in the scientific community as necessary to validate research findings.”

U.S. Department of Education. “Plan and Policy Development Guidance for Public Access.”

(2 min.)
Rachel

Here is a narrower definition of data that the U.S. Department of Education (and a number of other federal agencies) use. *[Read slide]*, oriented toward data specifically as recording of facts for the purposes of validation. This definition includes use as part of what defines data--it’s not just about formats or file types, but what is made *possible* by data. (An interesting point we won’t have time to discuss is that this definition is drawn from the Electronic Code of Federal Regulations on Intangible property. How might this context might inform the focus of the definition?)

"Our data bodies are **discrete parts of our whole selves** that are **collected, stored** in databases, the cloud, and other spaces of digitally networked flows, and **used to make decisions or determinations** about us. They are a **manifestation of our relationships** with our communities and institutions, including institutions of privilege, oppression, and domination."

Petty, Tawana; Saba, Mariella; Lewis, Tamika; Pena Gangadharan, Seeta; Eubanks, Virginia. (2018). Our Data Bodies: Reclaiming our Data.

(3 min.)
Rachel

And finally, here is a definition that emphasizes data as *necessarily* representational and symbolic. *[Read slide]*. In this sense, data don't have meaning divorced from context; we use them to tell stories about how things are and how they should be. They are manifestations of relationships (here personal data is being discussed, but the same is true of much research data as well).

Any questions or thoughts so far?

Special considerations

- Governing bodies and laws, e.g.:
  - HIPAA (Health data)
  - FERPA (Educational data)
  - IRB (Human subjects data)

- Example Practices
  - Data collection: e.g., Consent
  - Data storage; e.g., UM Sensitive Data Guide to IT Services
  - Data sharing: Data use agreements / restricted access
  - Data sharing: De-identification, anonymization

(3 min.)
Rachel

There are also certain data types that require special consideration that spans the entire life of a project, due to requirements at the federal, state, or institutional level, and you will need to plan especially carefully about how to handle these data appropriately from the start. These include things like health data (HIPAA) or other Personally Identifiable Information; education data (FERPA); and data associated with a project deemed to be human subjects data (IRB). (These categories are not exhaustive nor mutually exclusive, but each has different implications and requirements).

If any of your data fall into a specific protected category, this can have implications for handling, storing, and sharing the data throughout your project. We don't have time to delve too deeply into these considerations, but it is definitely worth identifying resources you may have at your disposal (including experts you may be able to consult with) to support you in answering questions around how to address concerns appropriately. Some of the tools and considerations we'll mention

throughout the workshop address these considerations as well, but for example you may need to think about…*[Read slide]*

General data ethics

- Context: Who do you trust? Whose voice(s) do you need to hear?
- Be aware of existing power structures
- Understand the challenges
  - De-identification, anonymization
  - Sub-field considerations (ecological data; social media data; indigenous data)
- Accessibility
- Use supportive tools (E.g.,
  - Data Ethics Canvas,
  - Principles for Advancing Equitable Data Practice
  - Why Am I Always Being Researched?)

(5 min.)
Rachel

I hope some of the discussion so far has helped make the argument for why ethics are a big part of *all* data decisions. I briefly mentioned some types of data that are governed by specific bodies or laws; however, we also want to emphasize that ethics decisions can't (and shouldn't) be outsourced. It is not enough to ask what "can" be done; personal judgement is also required, and good choices are contextual.

Although there are general best practices that can apply across disciplines and types of projects, we know that the "default" academic research practices tend to reinforce existing power structures, and there are many examples of data collection and use (historical, but also unfortunately contemporary) that actually cause harm. This is especially true when data involve or can affect vulnerable or marginalized people--but the data don't have to be *about* people to require careful consideration (E.g., open city information about pest control or road repair reports/requests could affect housing prices).

We also know that the research ecosystem is set up to incentivize new projects and products, rather than careful relationship-building, maintenance, and thoughtful processes. These are areas where institutional change needs to take place to better support researchers. In the meantime, however, we do our best.

And of course, the landscape keeps getting more complex. Some of these links, for example, point to ongoing discussions about how to ask for consent for future (possibly undefined) uses of data; whether genomic data are actually de-identifiable; and with the advent of big data, the danger of different datasets being combined to identify individuals even when the original datasets did not seem to pose this danger. This is not to say we should avoid sharing data, however--indeed, research fatigue is part of the challenge facing many communities who are asked to participate in research again and again, sometimes answering the same questions needlessly if those data are collected and never shared for reuse.

Additionally, making research data not just available but usable and accessible, is a topic that still needs research and standardization (though this is another reason data management best practices such as good organization and machine readability, are important in their own right).

This can seem overwhelming, but it's likely that only a subset of these considerations will apply to your own work, and in the last year or two, there has been an explosion in available resources on these topics. We have linked a few on this slide, and another project I'm working on is compiling a directory that we are organizing by project stage. There are also groups such as the Detroit Digital Justice Coalition (DDJC) and Detroit Community Technology Project (DCTP) that have been exploring how to advance equitable practices for collecting, disseminating and using open data for quite a while. I recommend starting with a tool such as the Data Ethics Canvas, which offers general questions to ask yourself about data at each stage of your project.

Data management planning

Kristin Briney
@KristinBriney

Data management is hard and everyone is bad at it. This includes data managers more often that we care to admit. Data management is basically fighting against chaos.

(2 min.)
Rachel

As the Tweet says, *[read slide].* But the truth is, we're all always engaging in data management, in our personal as well as professional lives--it's just a matter of whether we have been able to plan our actions, or are going it ad hoc. And the more clearly we understand our own end goals, the better job we can do at being transparent, knowing what we can promise, and keeping our word. Now I'm going to talk very briefly about the data planning process, before we have a short activity to give you a chance to reflect on your own work or experiences in light of the different parts of data management.

Planning for data: Why?

You want to...

- Create and document organization that allows you and others to more easily understand your work
- Articulate and follow through on your commitments

You have to...

- Meet legal and other regulatory requirements (for protected data types)
- Meet funding agency requirements

(4 min.)
Rachel

There are a variety of reasons planning for your data is important; we've already covered some of them, but these are a few of the most common. In addition to the ethical aspects, good planning makes running a project and communicating your results easier and less stressful.

Many funding agencies have required data management plans for a while now; and many are moving toward data sharing requirements as well, which requires especially good planning. However, there is still sometimes a ways to go in aligning requirements such as data sharing with appropriate infrastructure and support for researchers.

I mentioned it just now, but most commonly, data planning takes the form of a document called a data management plan. ***Who here has created a data management plan before?***

A data management plan usually lays out, in just a few pages, the framework for how your data will be created, organized, stored,

preserved, and shared. The goal of this process is to demonstrate that you have thought about how to do this work effectively rather than planning out every detail. You can’t get too deep in just a few pages, but your planning will be much more successful if you do it with an eye toward identifying decision points and thinking substantively about what you need to do--making things easier for your future self--rather than just checking off boxes on a requirement.

Planning for data: How?

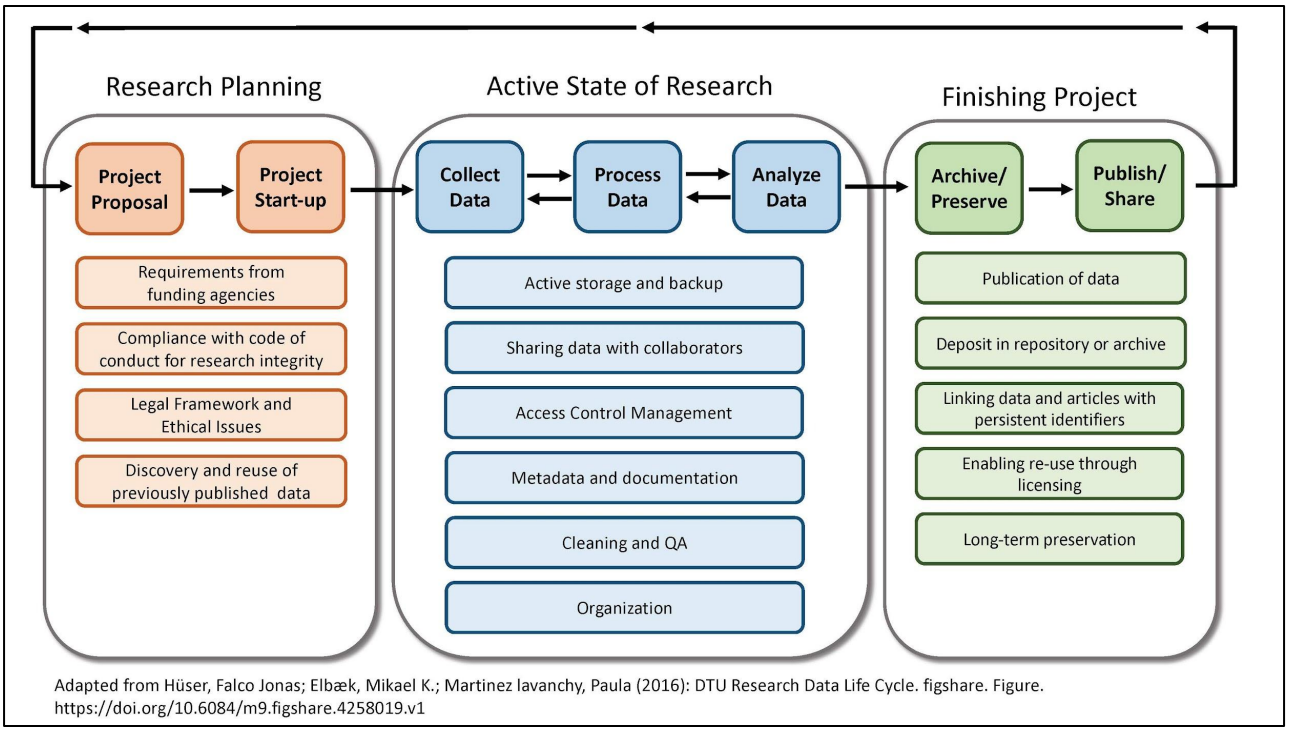
- **What types** of data will you collect or create? (Text, image, video? File type(s)?) Do you need a workflow?
- **How much** data will you collecting or create? (# of files? MB/GB/TB?)
- How will you ensure **reliability and validity** of the data?
- How will you **document** your process and decisions? What will you share?
- **Who needs access** to the data, and what training will they need to use and manage the data appropriately? **Who is responsible**?
- **How long** do you need access to the data?

(2 min.)
Rachel

Finally, we get to the "how," which Matt will be going over in more detail in the second part of the presentation. To be clear: data planning and management requires budgeting time and labor, and you should allocate some resources to this process in order to be successful. (Luckily, there is finally starting to be more recognition of the labor required to manage data well. Some grants allow budgeting for data consultations, preservation, positions, etc. as part of their proposal). Ethical data and research practices require slowing down, asking questions, and making space, which we know can be difficult and disincentivized within the research ecosystem.

There are obviously many ways to approach this task, but broad questions like these are often helpful to help where there are gaps you may need to consider further. *[Read slide?]* (This is true to some extent at any stage for a project--if you don't know where to start, where/who could you go to for help? If you're already in the middle of your project, where have you been running into problems?)

Next we'll be looking at a few tools that help with this process.



Adapted from Hüser, Falco Jonas; Elbæk, Mikael K.; Martinez lavanchy, Paula (2016): DTU Research Data Life Cycle. figshare. Figure. https://doi.org/10.6084/m9.figshare.4258019.v1

(3 min.)
Rachel

One tool for data management planning is what we (in the library world) often call the "data lifecycle;" this is usually shorthand to refer to a diagram of the activities and stages involving data throughout a research project. This is one way to structure your thinking--of course these differ from project to project, but it can be helpful to see a visual overview of the types of things you may need to consider, and explicitly ask yourself which activities apply to your work, and what resources you already have at your disposal to meet these needs (including tools, people, information sources, and your own experience). So for instance, *[Read slide]* A lot of these are what Matt is going to be addressing in the second half of the workshop.

Data lifecycle diagram - image description

Phase 1: Research Planning

Project proposal > Project startup

- Requirements from funding agencies

- Compliance with code of conduct for research integrity
- Legal Framework and Ethical Issues
- Discovery and reuse of previously published data

Phase 2: Active State of Research
Collect Data > Process Data > Analyze Data

- Active storage and backup
- Sharing data with collaborators
- Access Control Management
- Metadata and documentation
- Cleaning and QA
- Organization

Phase 3: Finishing Project
Archive/Preserve > Publish/Share

- Publication of Data
- Deposit in repository or archive
- Linking data and articles with persistent identifiers
- Enabling re-use through licensing
- Long-term preservation

(May lead back to Phase 1)

(Diagram adapted from Hüser, Falco Jonas; Elbæk, Mikael K.; Martinez lavanchy, Paula (2016): DTU Research Data Life Cycle. figshare. Figure. https://doi.org/10.6084/m9.figshare.4258019.v1)

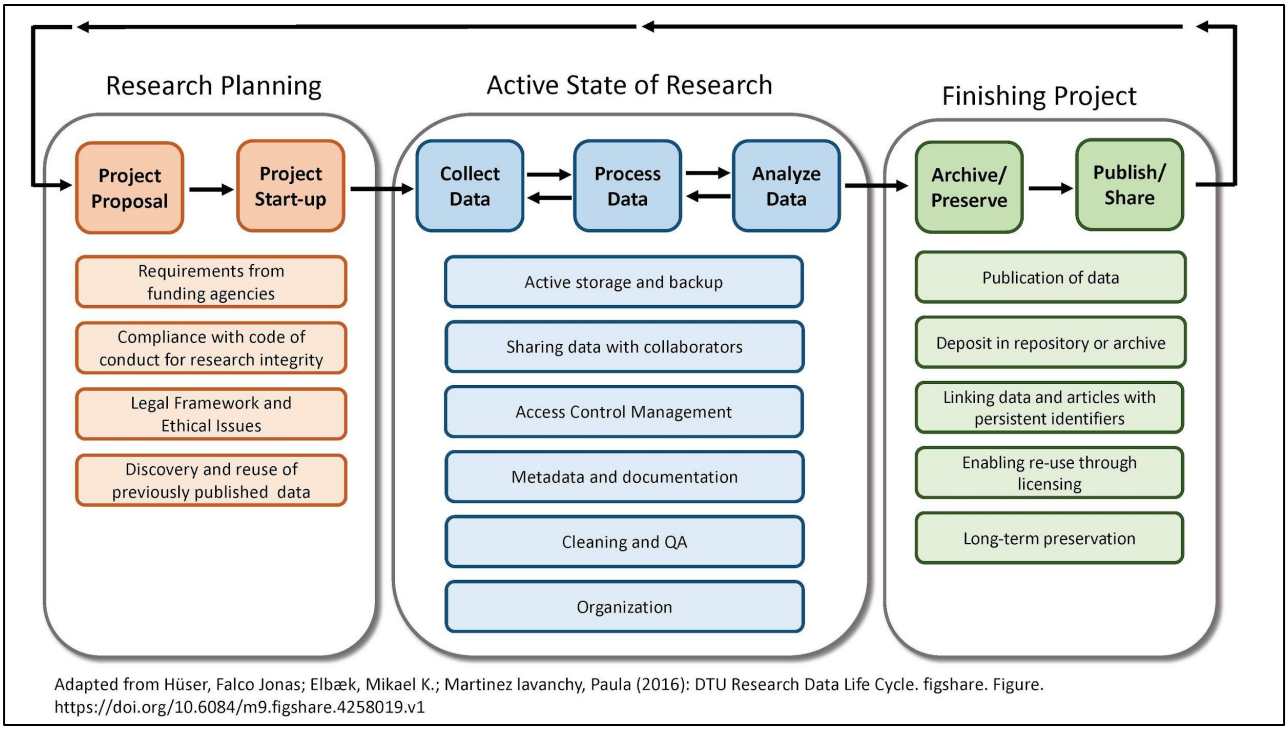
Activity/Break

1. Data Planning Checklist
2. DMPTool

(12 min.)
Rachel

Here are two more tools that can be useful for data management planning. The Data Planning Checklist is a series of questions that cover the entire project life, and the DMPTool guides creation of data management plans at the beginning of a project for different funding agencies (including the NEH). We will drop the links to these in the chat, and I will also put the data lifecycle diagram back up on the screen.

We're going to give you about 10-12 minutes to think about your own projects and reflect on them using one of these tools. Please also feel free to ask any questions or put them in the chat. If you need to take a break, this would also be a good time to do that. We're hoping that everyone is able to come out of this with one area identified where they might need to do more problem solving, or could improve their own practices. (Matt will also be talking about some specific ways to implement best practices in the second half of the workshop, which will hopefully provide some concrete ideas to build on).



Adapted from Hüser, Falco Jonas; Elbæk, Mikael K.; Martinez lavanchy, Paula (2016): DTU Research Data Life Cycle. figshare. Figure. https://doi.org/10.6084/m9.figshare.4258019.v1

(1-3 min.)
Rachel

Links for chat:

1. Planning checklist: https://docs.google.com/document/d/16O_X9yr6175zw9Fot_dBpobbVjka70jvuTOCUyrTyh0/edit?usp=sharing
2. DMPTool: https://dmptool.org/

We're going to move on to the second section of the workshop now, but before we do, is anyone willing to share anything that came up for them, either in terms of questions or realizations around your own work and process, or challenges in applying these tools?

Data Organization and Metadata

What is metadata?

- Metadata describes the content, quality, condition, and other characteristics of data.
- Metadata is standardized, structured information about an object that facilitates functions associated with that object.
  - Discovery, management, rights and access control, reuse, etc.
- Provides the context around the data

Matt

So, we've covered a lot in terms of how to plan for managing your data. Now let's jump forward to when you're just ready to jump into your research and start collecting data. Maybe you've finally gotten that big grant, or your dissertation topic just got approved, or maybe that assignment is just due in a couple of weeks.

So we're going to be talking today about data organization, data description, and documentation. We're going to start at the most granular level, looking at description and organization for individual data points or research objects you may be collecting, and working out to broader levels of description, things like file organization and project documentation.

One thing to consider in effective data management is describing individual data points, or applying metadata to your research data.

[Read info on slide]

Data can mean very different things depending on to whom you're talking; what discipline you're in; or what type of project you're working on.

Context is required to correctly interpret data; to know what it can and can't (or shouldn't) be used to do.

Next slide

Metadata

Metadata Type	Example Properties	Primary Uses
Descriptive metadata	Title Author Subject Genre Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management
Structural metadata	Sequence Place in hierarchy	Navigation
Markup languages	Paragraph Heading List Name Date	Navigation Interoperability

Riley, Jenn. Understanding Metadata: What is Metadata, and What is it For: A Primer. NISO, 2017

Matt

Metadata can be a complicated topic, so I'll be giving you something of a crash course, so don't be alarmed if you still have a lot of questions afterwards. That's why I'm here as a metadata specialist, to help you through the specifics.

Metadata can be applied to research objects in many different ways, including creating data dictionaries, applying appropriate column headers to spreadsheets, or using markup languages.

This chart provides a good breakdown of the major types of metadata that are applicable to many data types. Along with each metadata type, the chart provides some examples of metadata elements for each type, and primary uses.

[briefly go over slide]

Next slide

Finding a metadata standard

Research Data Alliance Metadata Directory

FAIRsharing

Open Metadata Registry

Linked Open Vocabularies

BioPortal

TK (Traditional Knowledge) Labels

Matt

If you are interested in applying metadata to the data that you collect, there are a lot of established metadata standards available, usually designed by people in various research disciplines to conform to the data types or research objects they use most often, or the research practices and methodologies they use. Not every discipline has established metadata standards, but many do.

Metadata standards are tools that are useful in some circumstances but may not always apply. Metadata standards are often built for very specific data types or disciplines. Think about how your audience will expect to interact with your data, and what viewpoint it may be representing (disciplinary, etc.)

[briefly talk about resources on the slide]

When should you use a metadata standard?

If data is “self-describing”, a metadata standard may not be necessary, but you may still consider adding additional metadata to provide more context, depending on the audience

Interoperability needs - who might you need to share this data with, and what software will you be using for your research? Does it require a certain metadata standard?

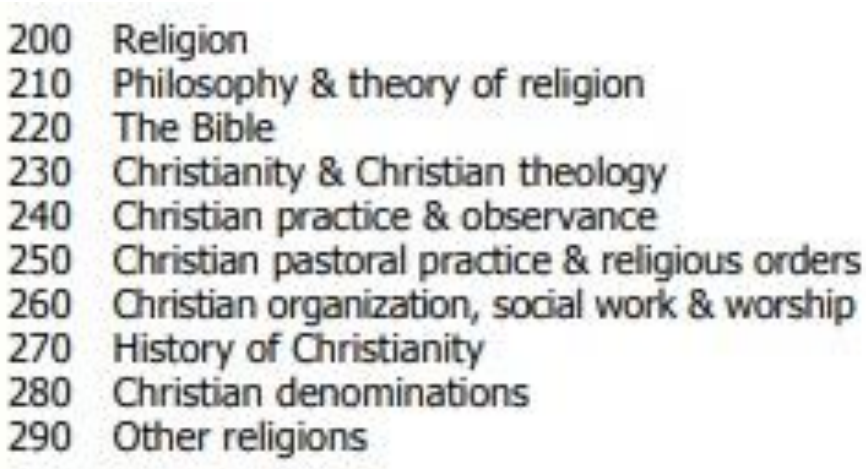
Considering audience - generalist vs. specialized

Even if a standard isn’t needed for your project, can be useful to build on or

think about your research space

Next slide

Critical Assessment of Metadata



Matt

Applying appropriate metadata requires critical assessment, so you'll want to consider biases under which metadata standards/controlled vocabularies may have been created, and also be aware/critical of historically problematic categorizations or language, documenting departures from those where you make them.

There are a lot of examples of this in existing systems. Not taking a critical approach can lead to biases and other consequences in your data and findings. Dewey Decimal Classification provides a perfect example of bias built into an organization and classification system. Religion is subdivided into 10 top-level categories, meant to broadly encompass all known religions. You can see that the first nine sub-categories are all aspects of Christianity. The tenth subdivision is "Other religions", which covers all other religions on the planet (representing the majority of humans on the planet).

So from this, you can get a sense of what the creators of this description and classification thought was important and what their worldview was. But it isn't really an accurate portrayal of religious practice world-wide. So it is important to assess the description and organization you apply to your data critically, and to document the

decisions you made regarding how to describe your data (especially if they deviate from traditional practice).

Next slide

Critical Assessment of Metadata

- Be aware of what metadata is collected, as some software packages don't always make that apparent.
- Make sure you get consent before collecting metadata.
  - https://www.wired.com/story/a-weird-mit-dorm-dies-and-a-crisis-blooms-at-colleges/
- Know the potential pitfalls of using certain types of software
  - https://eusprig.org/research-info/horror-stories/

Matt

Also, be aware of what metadata you are collecting or producing, especially if you are using commercial software for data collection, and make sure you get consent from any human research subjects before collecting metadata.

Here is a link to a recent Wired article on the unintended harm of hidden metadata. It tells the story of a survey that MIT gave its students (which Michigan designed, incidentally) on student mental health and wellbeing. MIT embedded location metadata in the survey, which they didn't tell students about, and it gave them the ability to know what dorm each respondent lived in. They then used that information to actually close down a dorm because they thought that students there had worse outcomes, but it was a flawed analysis and generated quite a bit of backlash. You can read the full story if you are interested.

And be aware of potential pitfalls of using certain types of software to collect, create, or manipulate metadata. Here is an entire site run by the European Spreadsheet Risk Assessment Interest Group that collects real-world stories of massive errors found in spreadsheet data and their consequences. One of my favorites from recent years is a study finding that there are errors in about 25% of all published genetics research papers because Microsoft Excel automatically changes certain data without the user's consent or knowledge.

Next slide

Documentation

- Who is responsible for what, and for how long?
  - Plan for knowledge transfer
- Decision points (methods, transparency)
  - Data formats collected
  - Vocabularies, classification and coding schemes
  - Workflows
  - Methods of analysis
- How many people need to understand it? For how long?

Matt

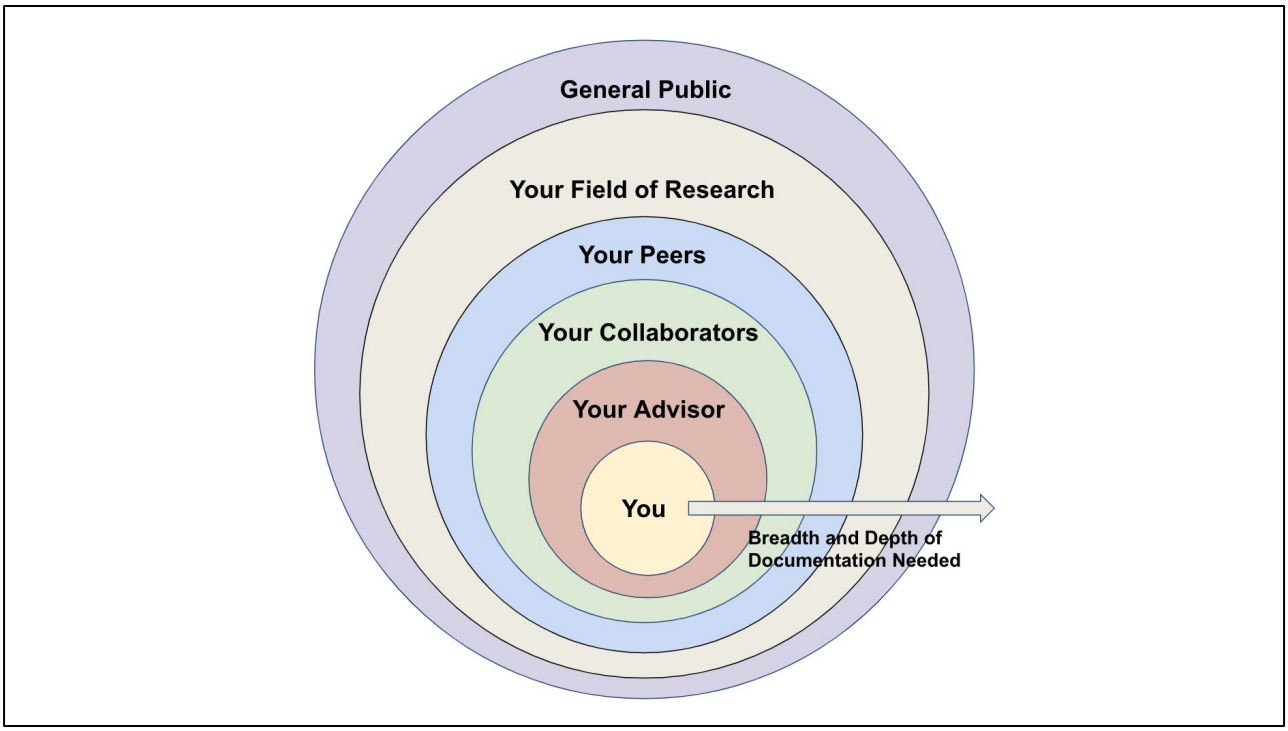
Along with metadata to describe the individual data points or research objects you are using, you'll likely want to create documentation for your data as a whole.

Documentation serves multiple purposes, both for your project team, as well as for others who may want to reuse your data later.

Documentation can allow you to:
Delineate responsibilities for your data and plan for knowledge transfer
Record decisions you and your team have made regarding the data, including:
What data formats are being collected
What vocabularies, classifications, and coding schemes are being implemented
Workflows
Explain the analytical approach for the project and justify its selection
Make a plan for how to format the data for analysis

When creating your documentation, consider the audience. Who will be using this documentation? How long do you expect the data to be useful?

Next slide



Matt

This graphic provides a representation of how to create documentation based on your intended audience.

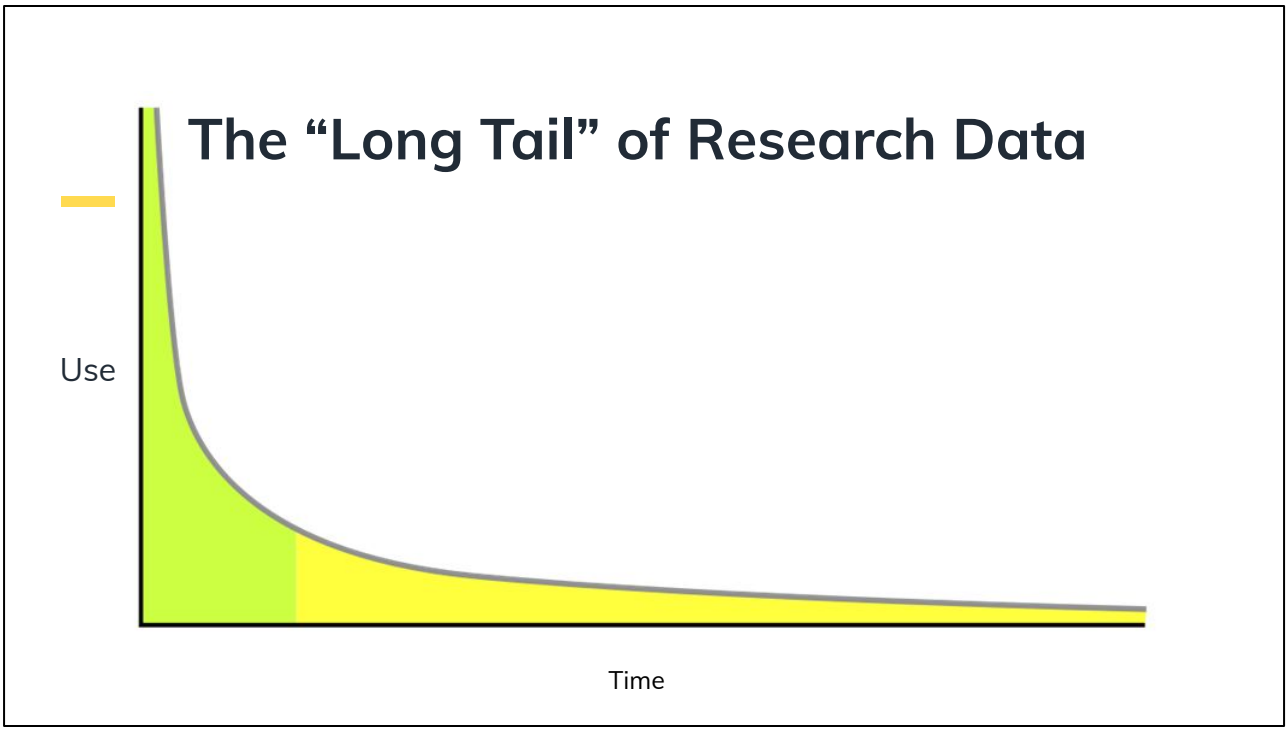
Requirements will likely change for each level of documentation

Different audiences require different levels of specificity, terminology, etc.

If you are creating the documentation just for your own use, it can likely be more narrow in scope. As you include more people, the breadth and depth (level of detail and explanations of the data) will likely need to increase significantly.

It is a good idea to consider your audience from the outset of your research, so you can plan appropriately for the time it will take to create sufficient documentation.

Next slide



Matt

One of the most compelling reasons to take the time to create good documentation for your research data is that effective data management and documentation can increase the “long tail” of research data.

The “long tail” refers to the length of time your research data is useful, and to how many people. Often research data is used and reused most immediately after it is published. Use tends to drop significantly after that (of course, each discipline is different), but there is often a long period of time (known as the “long tail”), where the research data is still being used and reused by a smaller subset of people.

If you create sufficient documentation for your data so that people can understand the data, methodologies, and tools used in the analysis, that data will likely be useful for a long time, thus increasing the length of the “long tail”. Additionally, this also applies to how long your data can be useful to you. Having sufficient documentation means that you won’t have to rely on your own memory of your data and analysis if and when you revisit the data years after the initial research ends.

Next slide

What files do you have?

- Research data
  - Spreadsheets
  - Images
  - Videos (edited, unedited)
  - Text (transcripts)
- Project documentation
  - Grant applications
  - IRB documentation
  - Process/policy/procedure documentation (how-to)
  - Financial/budget
- Technical documentation
  - Change logs
  - Architecture documentation
  - ID/Password management

Matt

Let's talk a bit about creating an organizational system or structure for your research data.

File organization is often very contextual. This is a high-level example of file organization that generally works across a number of different types of research.

It may be helpful to organize all of your research data in one area (including sections for raw data, processed data, etc.), project documentation in another area, and technical documentation in a third, all subdivided appropriately.

Take the time to work out a system for organizing your files that makes sense to you and your project team and stakeholders, and that will help make your workflows more efficient.

Next slide

Organization

Documenting your organizational system:

- Where data is stored
- How it is organized

Keep multiple copies in multiple locations - have a backup plan! (3-2-1)

Plan for eventualities like staff turnover, etc.

Matt

Once you have an organizational system worked out, you'll want to be sure to document that system, detailing how the data is organized and where it is stored.

Think about how your organization might reflect your workflow, and work in duplication or backups--so for a research project involving interview data, folders for raw interview transcripts, cleaned transcripts, and versions of cleaned transcripts suitable for preservation or sharing.

It's also important to have backups and redundancies built into your data organization. A good general rule is "3-2-1": Keep at least three (3) copies of your data, and store two (2) backup copies on different storage media, with one (1) of them located offsite

Documenting your organizational system can help you plan for things like staff turnover, ensuring that knowledge of the project doesn't get lost as people leave the research team.

Next slide

File naming

File naming is generally highly contextual to individual research projects.

The most important thing is to be consistent and descriptive in naming and organizing your files.

Include within your documentation information on the naming convention you used.

Smithsonian Data Management Best Practices - Naming and Organizing Files

Matt

Now, we couldn't have a workshop on data management without having at least one slide on file naming conventions, since it's a very common question that researchers have: "how do I name my research files?"

[Read info on slide]

Generally helpful to think about what elements will allow you to distinguish effectively between files (including across folders), without being visually overwhelming to scan.

The Smithsonian has a very good, three-page guide to file naming best practices.

Next slide

Tools

- How to choose/considerations (tools don't solve process problems)
- Accessibility/limitations
- Working with sensitive data - UM Sensitive Data Guide to IT Services
- SPG 601.07 - Responsible Use of Information Resources
- File formats, openness, interoperability
- Tools should make your life easier/save you (and others!) time
    - Consider setting a time limit for deciding

Matt

Rather than list actual tools (which can vary widely by discipline, data type, and other factors), we wanted to provide you with some considerations for how to choose the appropriate tools for your research.

Accessibility/limitations:
Cost - what are the benefits and downsides of purchasing software? Especially if the software uses a proprietary data format?
Ease of use - how long will it take for you and your team to learn and implement the tool? Is that investment of time and effort worth the benefit?
Access - is the software licensed? For how long? Is it actively supported, or at risk of falling into obsolescence?

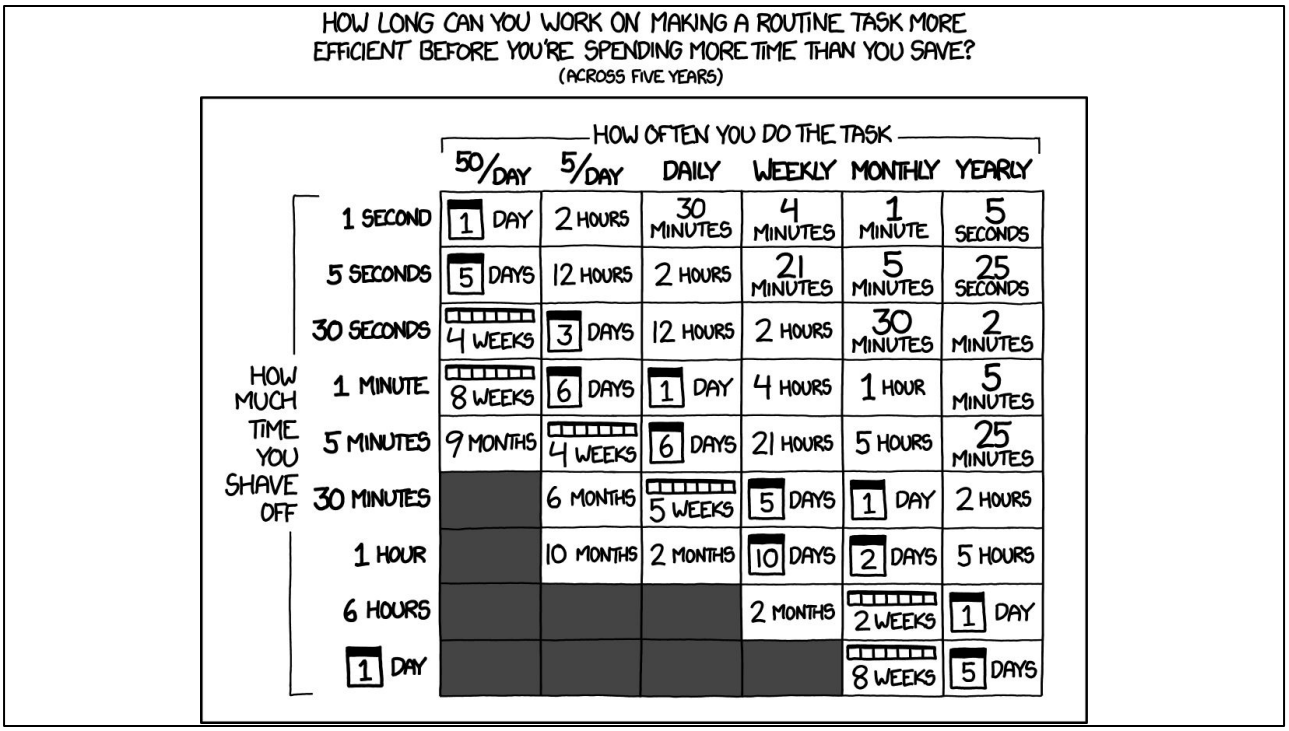
Consider which file formats the tool you want to use supports, if it is interoperable with other tools that you are considering or need to use.

Working with sensitive data:
IT Services provides a great guide to choosing the right tools for working with various kinds of sensitive data.
[Link for the chat: https://safecomputing.umich.edu/dataguide/?q=home]

The university also has a Standard Practice Guide that covers the responsible use of information resources

Don’t over-do it! Tools are supposed to make your research easier, not more complicated or difficult. Ask yourself what benefits the tool brings to your research, and consider setting a time limit for investigating, deciding on, and implementing tools.

Next slide



Matt

https://xkcd.com/1205/

This is an xkcd comic that illustrates that point. It's a graph that tells you how long you can spend on trying to make a routine task more efficient before you are spending more time than you would save. It's a great reminder not to get too bogged down in trying to make things more efficient, or else you run the risk of actually slowing down your research.

Next slide

Sharing and Preserving

- Who will/should have access?
- What to share?
- Context needed to interpret
- Lifespan
- Transparency, reproducibility, limitations/uncertainty
- Credit labor appropriately

Matt

Generally, the last step in managing your data is how you share and preserve the data long-term. There are a number of questions to consider about how best to do this, including:

Who will or should have access to the data?
Are there regulations or ethical considerations that dictate how the data can be shared?
Context is also required to correctly interpret data; to know what it can and can't (or shouldn't) be used to do, so how do you make that clear to people who view or reuse the data? What documentation and metadata do you need to include with the data you share?
What is the expected lifespan of the data? How long will it be useful and relevant? This is often very different based on discipline.
How do you ensure transparency and reproducibility of the data?
How do you communicate limitations in the data or uncertainty of the findings?

How can you appropriately credit the labor that went into creating, processing, and analyzing the data?

Next slide

Sharing and Preserving

Repositories for long term preservation and access:

Domain-specific repositories (ICPSR, tDAR)

Institution-specific repositories (Deep Blue Data)

General data repositories (Dryad, Open Science Framework, Figshare)

Matt

In terms of sharing and preserving research data, there are many repositories available.

[Read info on slide]

Each repository will have its own policies for access and retention of data, so do your research and make sure you choose what is best for you and your data. In particular, if you need long-term retention and access for you or others, consider things like what will happen to your data if the organization loses funding, and whether it will provide any services for migrating or keeping your data accessible as formats and softwares change.

Some repositories may cost money; it can be useful to plan for this ahead of time and work it into your funding proposal if applicable.

Next slide

Further resources

U-M Library Data Services (finding and mining data, data planning, metadata consultation, sharing and preserving data, visualization)

- UK Data Service
- ICPSR Guidelines for Effective Data Management Plans
- Digital Curation Centre Data Management Plan Checklist
- PLOS Ten Simple Rules for Creating a Good Data Management Plan
- DMPTool
- SPARC - Data Sharing Requirements by Federal Agency

Matt

So that's all of the content we have prepared for today. We know it's a lot, and we definitely expect that you'll still have lots of questions, so we wanted to wrap up with a few slides on where you can find more information and help with managing your data.

We have put together a list of online resources that can help you with data management.

We also have a number of resources and services here in the library that can help you with data management planning.

Don't be overwhelmed--this is just a starting point, and what we hope you take away from this session is one practice or area where you feel like you can take steps to improve your data management (and make your work easier!).

Next slide

We can help!

Digital Scholarship

Digital Scholarship Office Hours

- First Tuesdays, 3:30 – 4:30 PM
- Third Tuesdays, 3:30 – 4:30 PM
- https://umich.zoom.us/j/258694314

Matt

We also have office hours.

Next slide

Readings

- Briney, Kristin. (2015). Data Management for Researchers : Organize, maintain and share your data for research success. Research skills. Exeter: Pelagic Publishing.
- Ross MW, Iguchi MY, Panicker S. Ethical aspects of data sharing and research participant protections. Am Psychol. 2018 Feb-Mar; 73(2):138-145. https://pubmed.ncbi.nlm.nih.gov/29481107. PMID: 29481107.
- See Slide 12 for additional ethics readings

Matt

Here are a few suggested readings if you would like to look further into the practical aspects of data management, or some of the ethical questions that spring up around different types of data.

Next Slide

How was this?

We'd love to hear what you think about what you liked in this workshop and what could have been improved.

Feedback Form: https://umich.qualtrics.com/jfe/form/SV_6EbiwabxlYM1z2m

With that, I want to thank everyone for coming, and we'll open up the rest of the time we have for questions. Please note that there will also be a very short survey for this session, so please fill it out if you can--this will really help us better understand what worked for this session, and what we could improve for next time. We'll put the link in the chat.

So let us know if you have questions! Feel free to put them in the chat, or raise your hand.