Session 17: Reproducibility and Transparency

October 26, 2020

Onikepe Owolabi

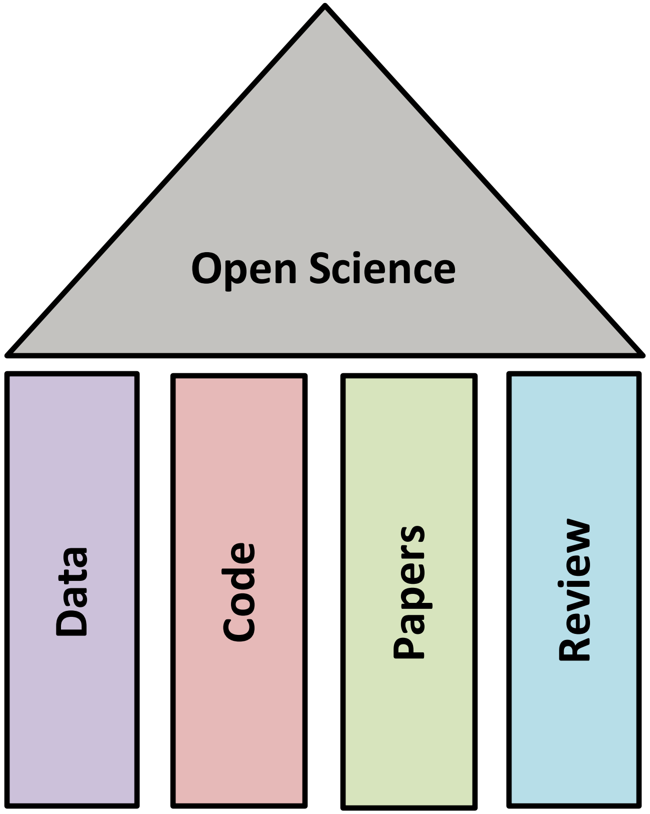
Open Science

- Loosely: science that is transparent in its methods, and accessible to everyone.

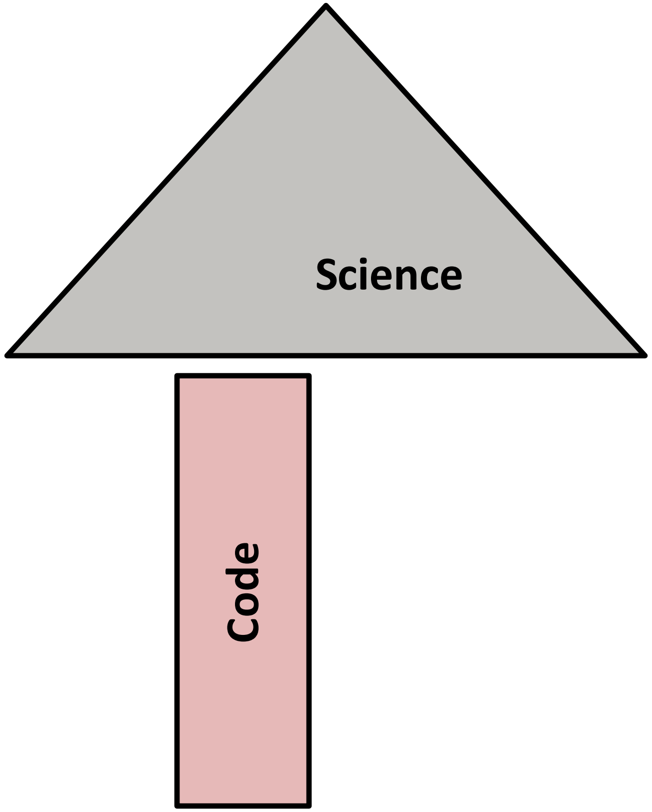
- Practically: means a lot of different things to different people.

- Ties into both the **equitable distribution** of scientific knowledge, as well as the **reliability** of results.

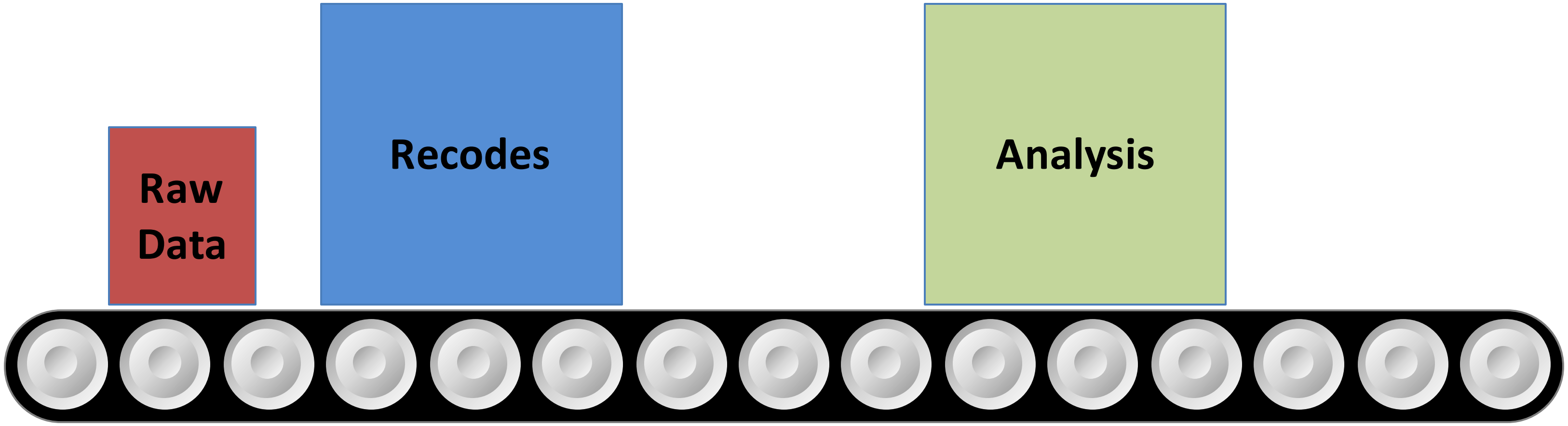
Open Science



Science



What is Reproducibility?



Raw Data

Recodes

Analysis

Replicability vs. Reproducibility:

- **REPLICABLE**: *When you (or someone else) performs the original analysis on* ***new data*** *and get the same (or similar) results.*

- **REPRODUCIBLE:** *When you (or someone else) performs the original analysis on the* ***same data*** *and can recreate all of your published results.*

Why wouldn’t work be reproducible?

- A lot of tiny decisions
- Analysis happened in a lot of different places
- Not enough documentation
- Murphy’s Law

Stata? Excel? A piece of scrap paper? Your head?

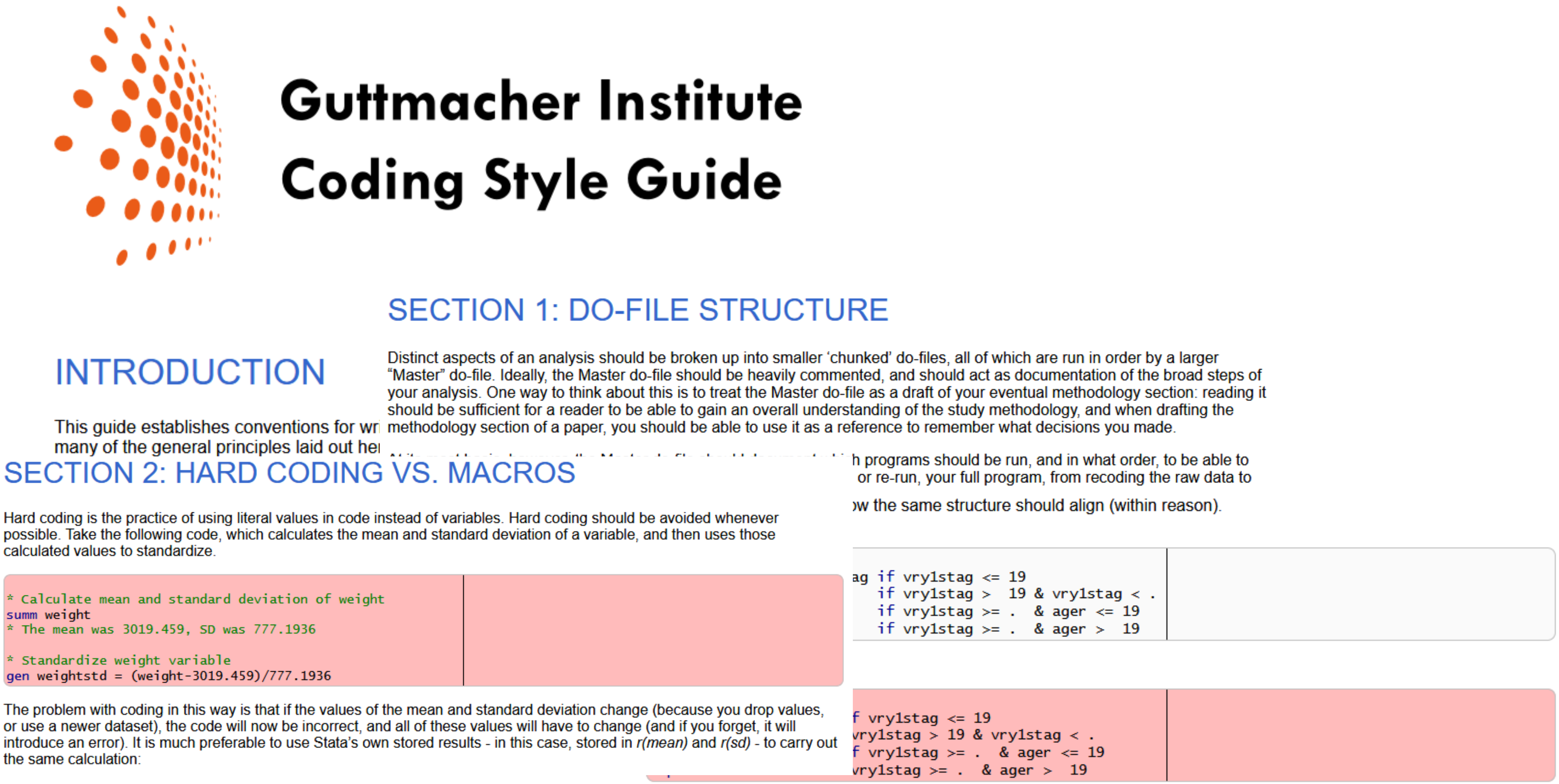
Why is reproducibility important for abortion research?

- Transfer of knowledge
- Economies of scale
- Easier to check results

Protects our work

- As abortion researchers, it’s incredibly likely that our work will be challenged
- Having well-checked and well-documented code makes responding to criticisms easier (and less stressful)

How do you actually do it?



Guttmacher Institute Coding Style Guide

INTRODUCTION

This guide establishes conventions for wri
many of the general principles laid out her

SECTION 1: DO-FILE STRUCTURE

Distinct aspects of an analysis should be broken up into smaller ‘chunked’ do-files, all of which are run in order by a larger “Master” do-file. Ideally, the Master do-file should be heavily commented, and should act as documentation of the broad steps of your analysis. One way to think about this is to treat the Master do-file as a draft of your eventual methodology section: reading it should be sufficient for a reader to be able to gain an overall understanding of the study methodology, and when drafting the methodology section of a paper, you should be able to use it as a reference to remember what decisions you made.

h programs should be run, and in what order, to be able to
or re-run, your full program, from recoding the raw data to

ow the same structure should align (within reason).

```
ag if vry1stag <= 19
   if vry1stag >  19 & vry1stag < .
   if vry1stag >= .  & ager <= 19
   if vry1stag >= .  & ager >  19
```

```
f vry1stag <= 19
vry1stag > 19 & vry1stag < .
f vry1stag >= .  & ager <= 19
vry1stag >= .  & ager >  19
```

SECTION 2: HARD CODING VS. MACROS

Hard coding is the practice of using literal values in code instead of variables. Hard coding should be avoided whenever possible. Take the following code, which calculates the mean and standard deviation of a variable, and then uses those calculated values to standardize.

```
* Calculate mean and standard deviation of weight
summ weight
* The mean was 3019.459, SD was 777.1936

* Standardize weight variable
gen weightstd = (weight-3019.459)/777.1936
```

The problem with coding in this way is that if the values of the mean and standard deviation change (because you drop values, or use a newer dataset), the code will now be incorrect, and all of these values will have to change (and if you forget, it will introduce an error). It is much preferable to use Stata’s own stored results - in this case, stored in *r(mean)* and *r(sd)* - to carry out the same calculation:

Principles of Reproducibility

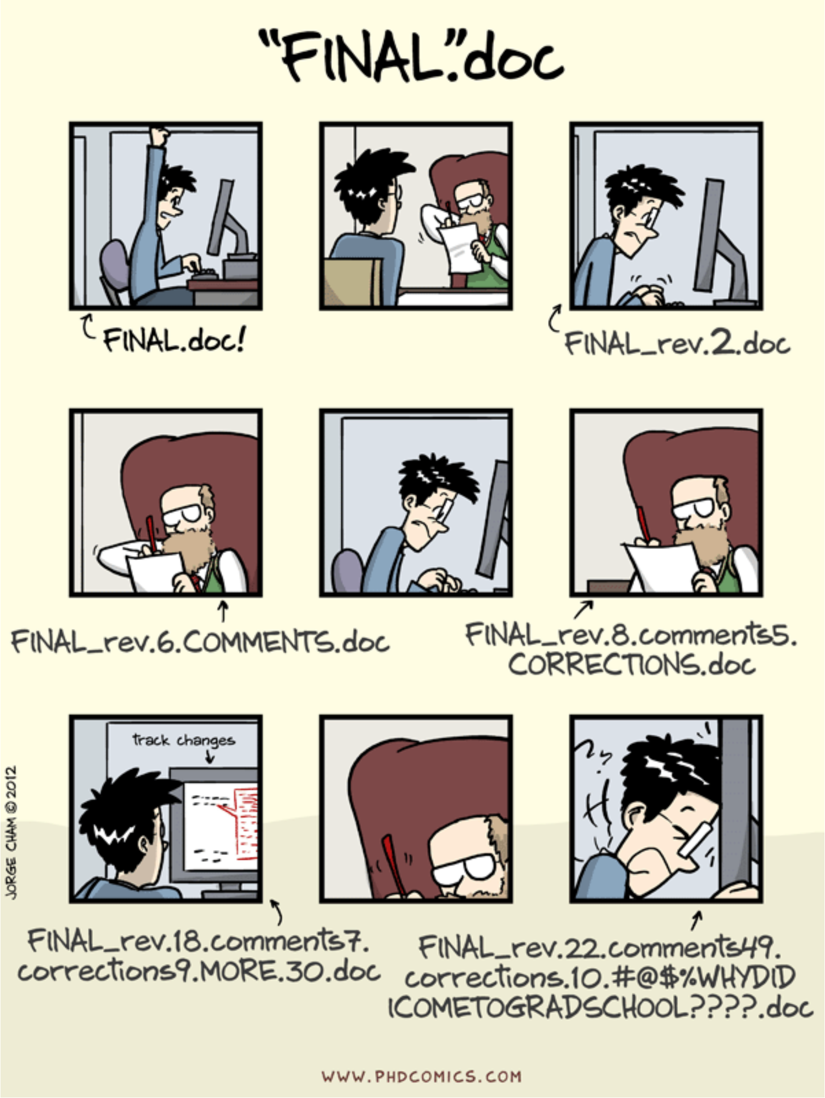
- All steps in an analysis project, including data cleaning and management, should be reproducible.

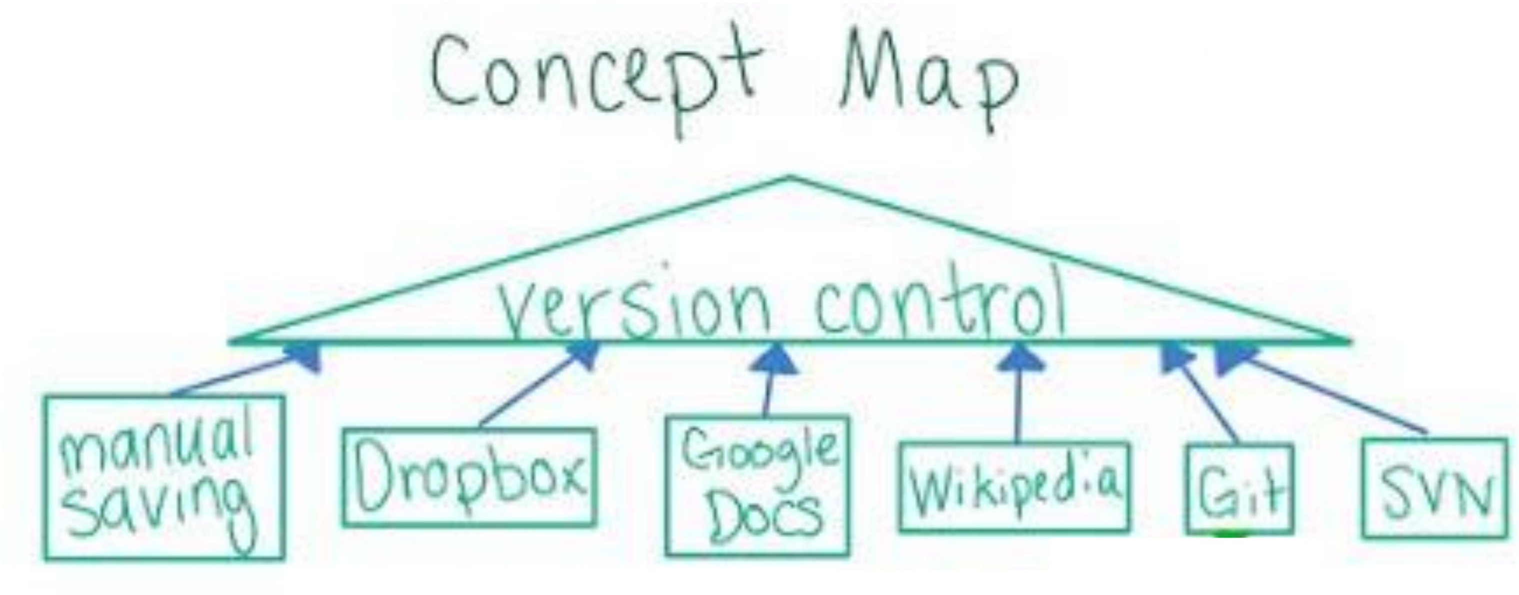
- Every program is commented well enough that a research assistant with basic programming experience could follow each step.

- Every number and figure in text and tables can be easily traced back to the code that produced it.

Version control

What is version control?




Concept Map
version control
manual saving
Dropbox
Google Docs
Wikipedia
Git
SVN

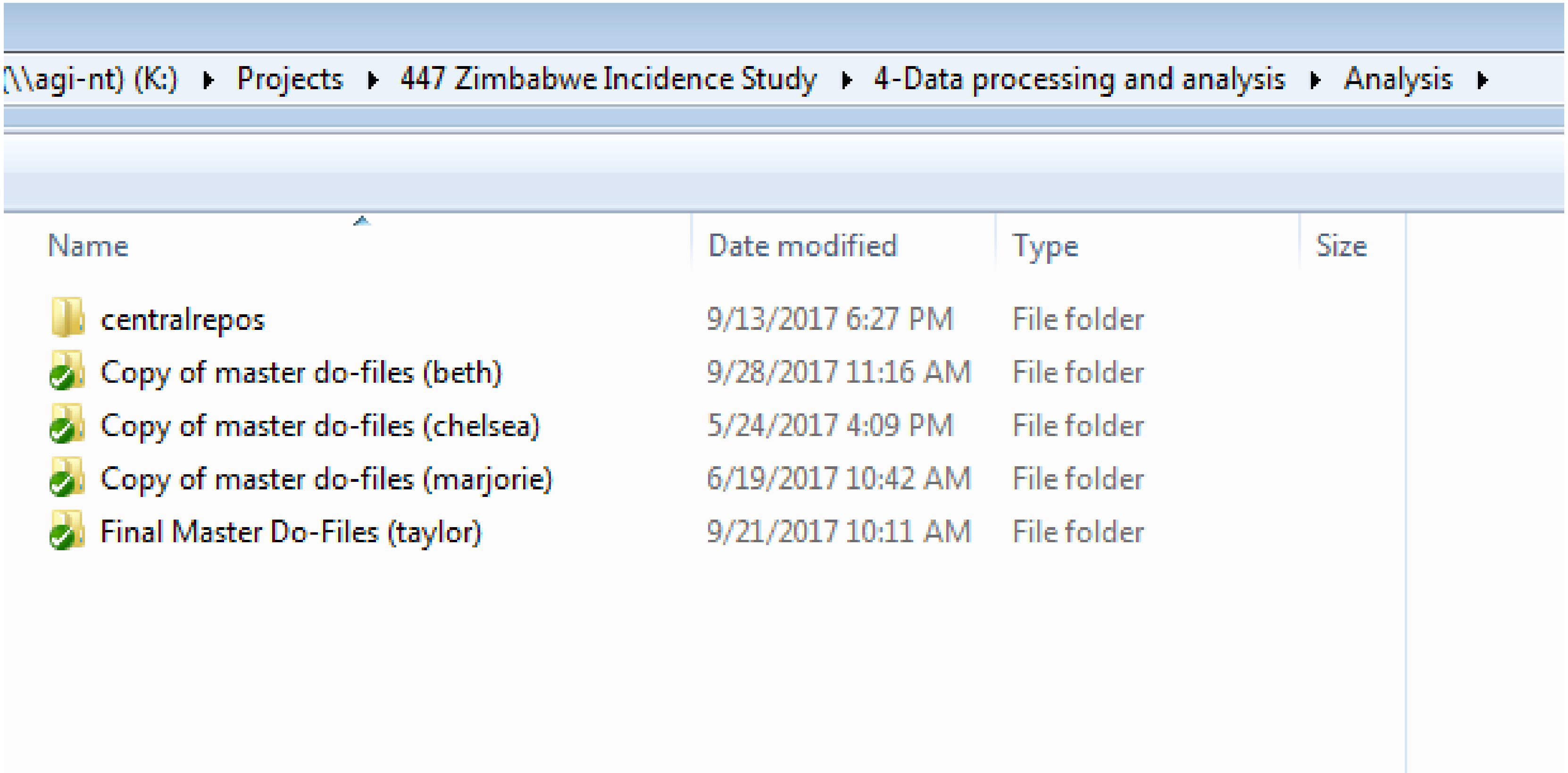
Version control is a way....

- to keep track of your files
- to track incremental changes in work over time
- for team members to all work in the same files simultaneously
- for team members to collaboratively work on code
- to review code written by others
- to back up your files
- to go back and find old versions of your files

Folder Structure

- **One folder will serve as your repository, a central place where changes to your files are tracked and stored**
  - – The repository does not contain your actual files
- **A second set of folders will be your working folders, where your actual files will sit**
  - – After pulling from the repository, your files in a folder will be up-to-date

Example of folder structure

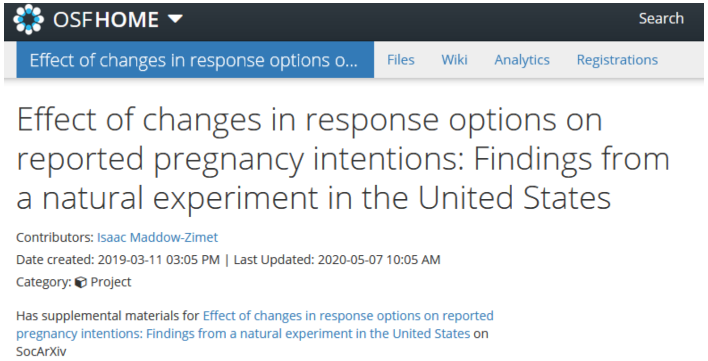


\\agi-nt) (K:) ▸ Projects ▸ 447 Zimbabwe Incidence Study ▸ 4-Data processing and analysis ▸ Analysis ▸

Name	Date modified	Type	Size
centralrepos	9/13/2017 6:27 PM	File folder	
Copy of master do-files (beth)	9/28/2017 11:16 AM	File folder	
Copy of master do-files (chelsea)	5/24/2017 4:09 PM	File folder	
Copy of master do-files (marjorie)	6/19/2017 10:42 AM	File folder	
Final Master Do-Files (taylor)	9/21/2017 10:11 AM	File folder	

How are we doing it? (cont.)

- Code review
- Public code



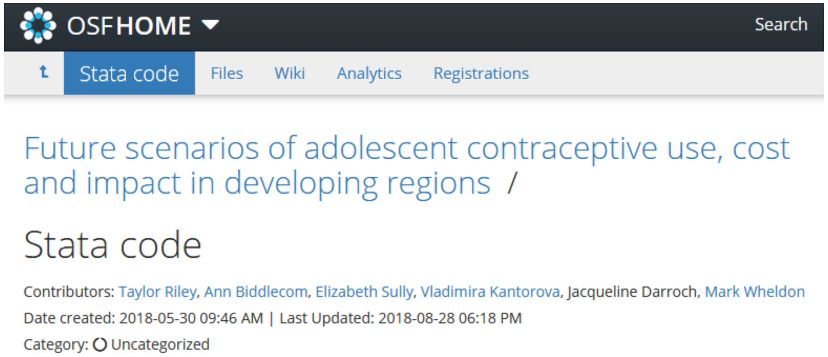
Effect of changes in response options on reported pregnancy intentions: Findings from a natural experiment in the United States

Contributors: Isaac Maddow-Zimet

Date created: 2019-03-11 03:05 PM | Last Updated: 2020-05-07 10:05 AM

Category: Project

Has supplemental materials for Effect of changes in response options on reported pregnancy intentions: Findings from a natural experiment in the United States on SocArXiv



Future scenarios of adolescent contraceptive use, cost and impact in developing regions /

Stata code

Contributors: Taylor Riley, Ann Biddlecom, Elizabeth Sully, Vladimira Kantorova, Jacqueline Darroch, Mark Wheldon

Date created: 2018-05-30 09:46 AM | Last Updated: 2018-08-28 06:18 PM

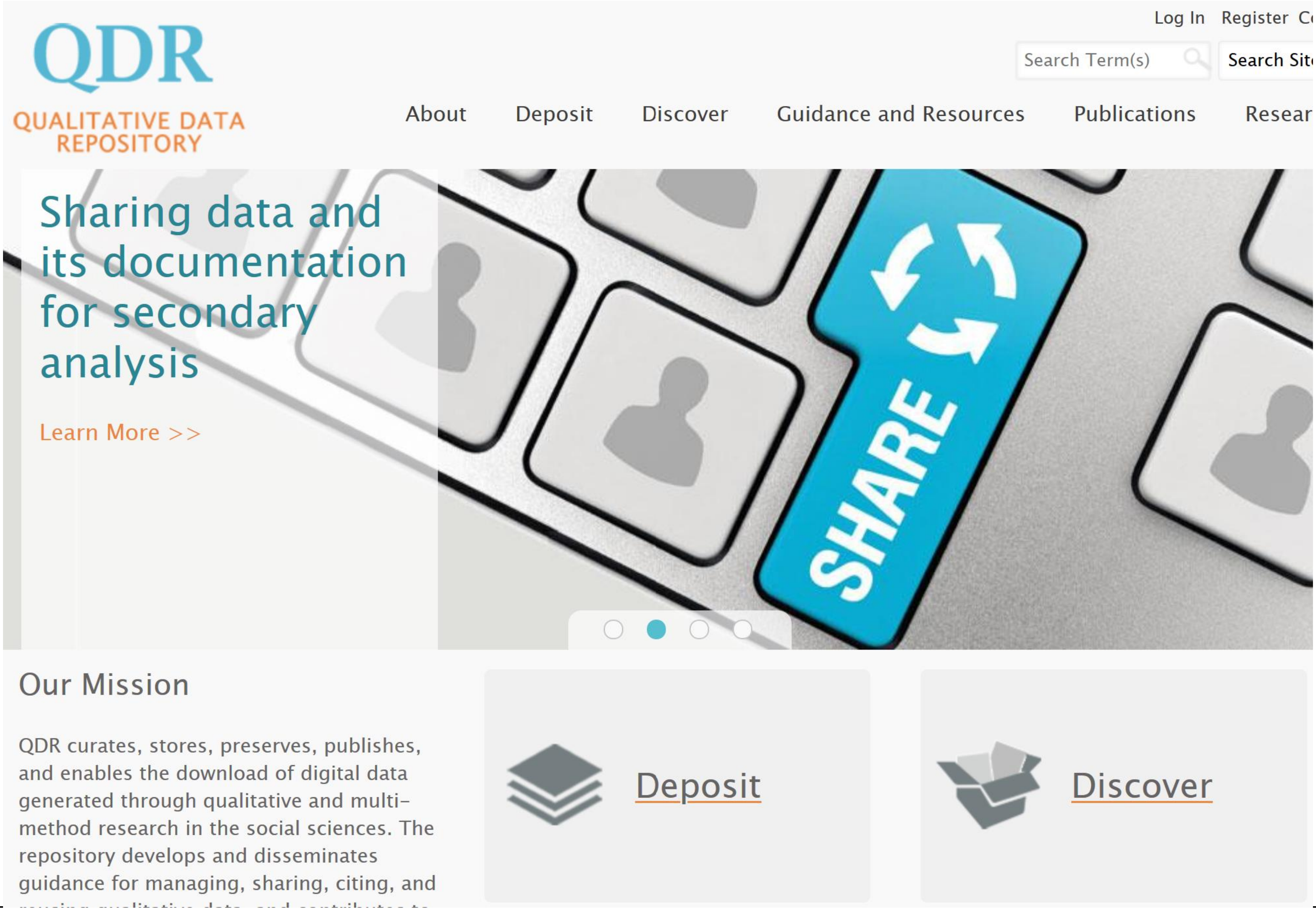
Category: Uncategorized

- Trainings, trainings, trainings… training, trainings, trainings

Resources

- Style guide (publicly available! It’s at https://guttinst.github.io/)
- Guide to exporting tables programmatically in Stata (will be public soon!)
- Public code for our projects on OSF
- What else would be helpful?

Transparent Qualitative Research with Sensitive Data



Log In Register C

Search Term(s) Search Sit

QDR

QUALITATIVE DATA REPOSITORY

About Deposit Discover Guidance and Resources Publications Resear

Sharing data and its documentation for secondary analysis

Learn More >>

○ ● ○ ○

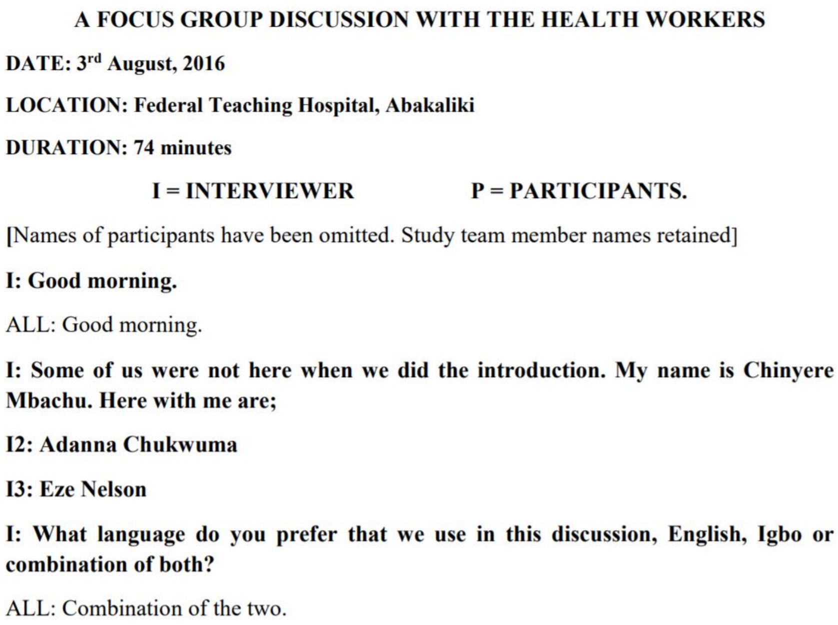
Our Mission

QDR curates, stores, preserves, publishes, and enables the download of digital data generated through qualitative and multi-method research in the social sciences. The repository develops and disseminates guidance for managing, sharing, citing, and reusing qualitative data, and contributes to

Deposit

Discover

Qualitative data



A FOCUS GROUP DISCUSSION WITH THE HEALTH WORKERS

DATE: 3rd August, 2016

LOCATION: Federal Teaching Hospital, Abakaliki

DURATION: 74 minutes

I = INTERVIEWER P = PARTICIPANTS.

[Names of participants have been omitted. Study team member names retained]

I: Good morning.

ALL: Good morning.

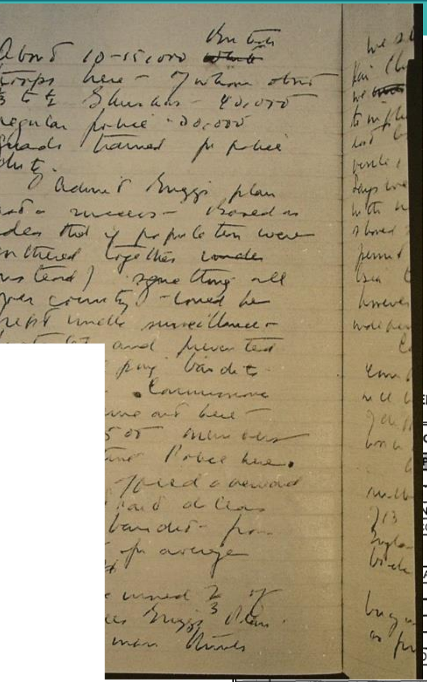
I: Some of us were not here when we did the introduction. My name is Chinyere Mbachu. Here with me are;

I2: Adanna Chukwuma

I3: Eze Nelson

I: What language do you prefer that we use in this discussion, English, Igbo or combination of both?

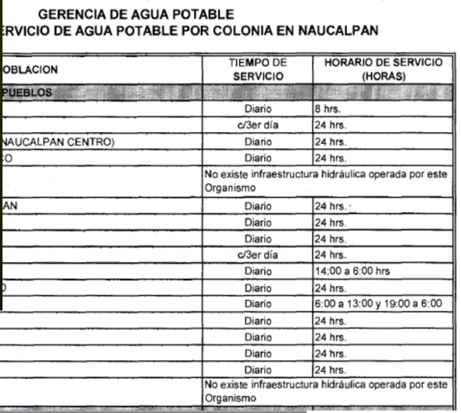
ALL: Combination of the two.



140

About 10-15,000 ~~[illegible]~~ British troops here — of whom about $\frac{1}{3}$ to $\frac{1}{2}$ [illegible] — 40,000 regular police · 30,000 guards trained for police duty.

Admiral Briggs plan had a success — based on idea that if population were gathered together (would be instead) something all over country — could be kept under surveillance — [illegible] and prevented [illegible] paying bandits [illegible] Communist [illegible] have not been [illegible] 50 or [illegible] members [illegible] Police here. [illegible] offered a reward [illegible] paid at [illegible] bandit — from [illegible] for average [illegible] [illegible] 2/3 of [illegible] Briggs plan. [illegible] man [illegible]



GERENCIA DE AGUA POTABLE

ERVICIO DE AGUA POTABLE POR COLONIA EN NAUCALPAN

	OBLACION	TIEMPO DE SERVICIO	HORARIO DE SERVICIO (HORAS)
	PUEBLOS		
		Diario	8 hrs.
		c/3er día	24 hrs.
	NAUCALPAN CENTRO)	Diario	24 hrs.
	O	Diario	24 hrs.
		No existe infraestructura hidráulica operada por este Organismo	
	AN	Diario	24 hrs.
		Diario	24 hrs.
		Diario	24 hrs.
		c/3er día	24 hrs.
		Diario	14:00 a 6:00 hrs
	O	Diario	24 hrs.
		Diario	6:00 a 13:00 y 19:00 a 6:00
13	SANTA CRUZ ACATLAN	Diario	24 hrs.
14	SANTA CRUZ DEL MONTE	Diario	24 hrs.
15	SANTA MARÍA NATIVITAS	Diario	24 hrs.
16	SANTIAGO OCCIPACO	Diario	24 hrs.
17	SANTIAGO TEPATLAXCO	No existe infraestructura hidráulica operada por este Organismo	

What is qualitative research and why would you do it?

- Theory-building: do your qualitative work *before* your survey, then test the theories/pathways you saw in the qualitative data

- Interpreting or illuminating quantitative findings

- Showing the mechanisms of associations that the data show us exist, but we don't understand

- … and more

Can qualitative research be be reproducible?

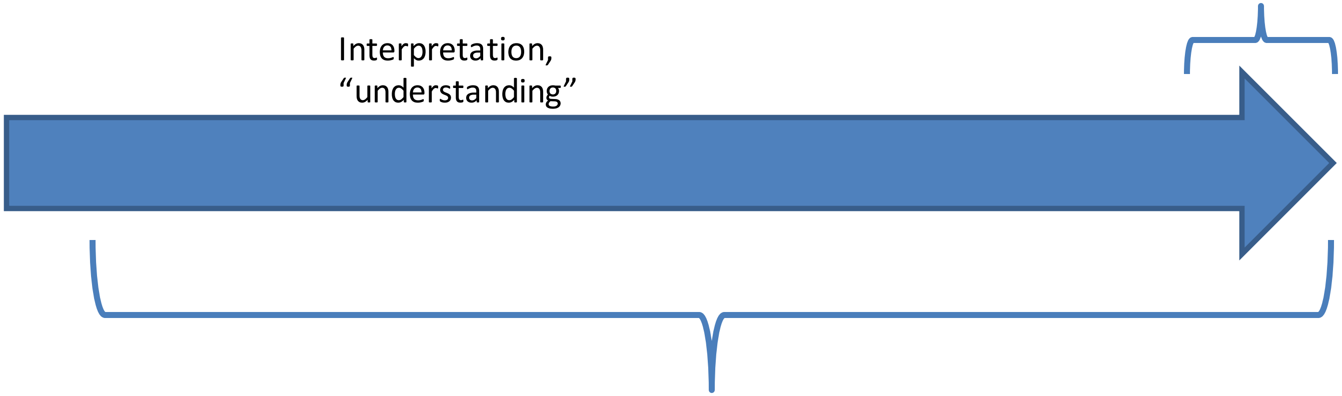
- Challenges
  - Appropriate?
  - Ethical?
  - Feasible?

Transparency vs. Reproducibility

- **Aspects of transparency**
  - Data access
  - Analysis
  - Production

= reproducible for quant

Transparency vs. Reproducibility: Epistemology



Reproducibility

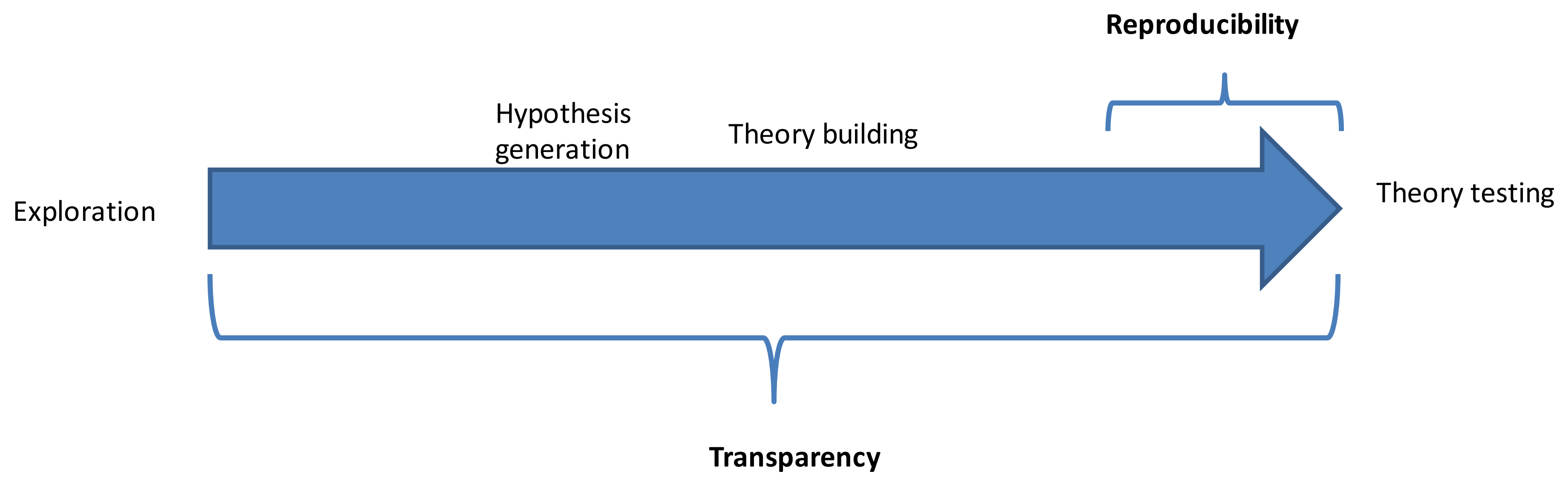
Interpretation, “understanding”

Solipsism: Only author can understand their own data

Positivism: Search for universal laws

Transparency

Transparency vs. Reproducibility: Research cycle



Data Access: Ethics

Consent

De-identification

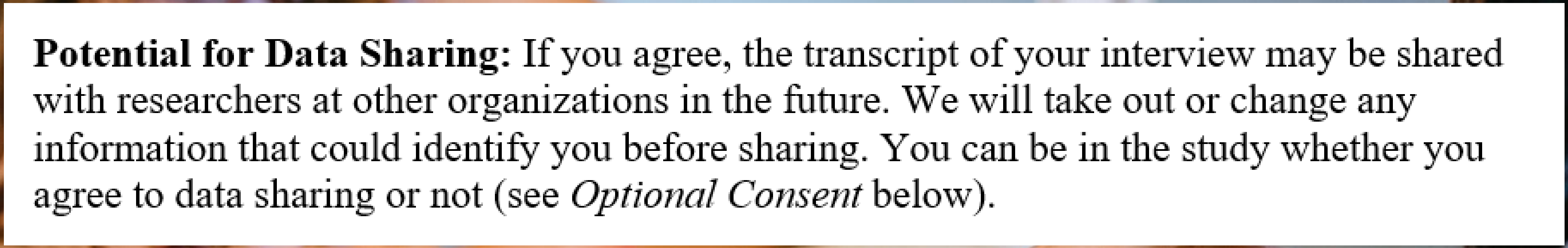
Removal of identifying details

Might not be enough!
Withhold or restrict access to transcripts

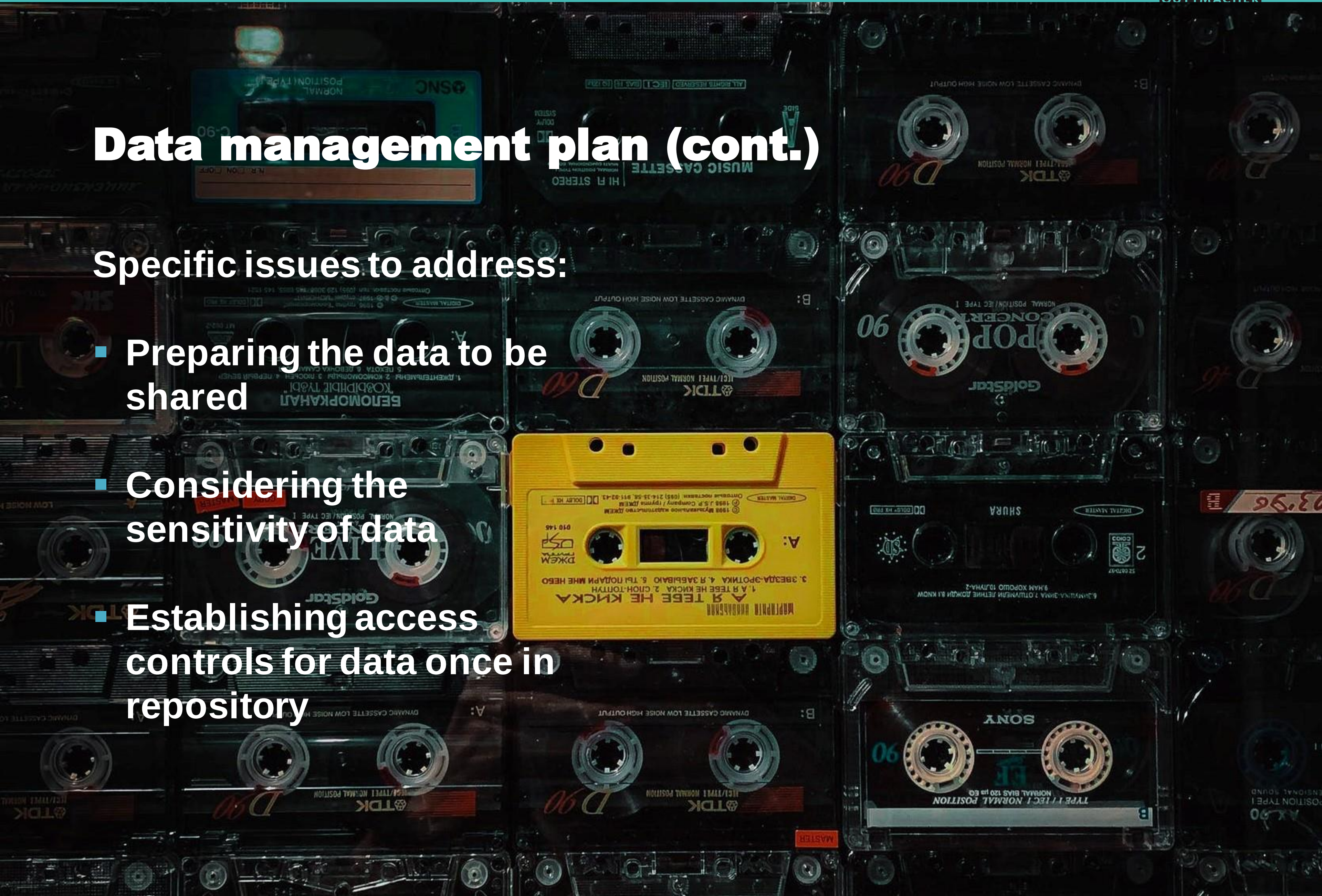
Secure storage

Appropriate access conditions

Data sharing in IRB documentation / informed consent form



Potential for Data Sharing: If you agree, the transcript of your interview may be shared with researchers at other organizations in the future. We will take out or change any information that could identify you before sharing. You can be in the study whether you agree to data sharing or not (see *Optional Consent* below).



Data management plan (cont.)

Specific issues to address:

- Preparing the data to be shared
- Considering the sensitivity of data
- Establishing access controls for data once in repository

Why make a transparency plan?

- **To help you follow the principles of transparency throughout the lifecycle of your project**
  - To facilitate potential data sharing
  - To help you generate transparency-related materials as you go
- **To document your activities for future work**
- **To help you distinguish between internal documentation and what's needed for inclusion in a public repository**

Excerpts from a checklist for qualitative transparency

Stage/category	Document/item	Internal only?	For sharing?	Notes
All/meta	Methodological underpinning and study justification		X	Only necessary as a separate piece if it is not included in the IRB application
Data collection Analysis	Data management plan		X	
Data	Recordings of interviews	X		Deleted per IRB requirements
Analysis	Coding scheme with node descriptions		X	

OSF Preregistration

- Publicly post research plan and pre-analysis plan, ideally before you start collecting data

- No requirements on what/how much information to include when you preregister.
  - OSF preregistration for is just a guide for what could be included

- Inclusion decisions depend on:
  - Goals of preregistering
  - Project type/scope

OSF Preregistration

- Monday’s training will cover:
  - Overview of Open Science Framework
  - Benefits of preregistering
  - Tutorials on how to use the site, how to preregister
- Homework:
  - Look over the preregistration form and Ghana example
    - Think about how this might be used to preregister your study
  - Bring questions to training