Data Management for Qualitative Research

The Ohio State University, July 23, 2018

Sebastian Karcher (Qualitative Data Repository)

Agenda

- Research and the data lifecycle
- The value of planning and DMPs
- Intervention points in data lifecycle
- Transparency and data documentation
- Keeping your data safe and secure
- *Exercise:* Evaluating DMPs

Definitions

Research data management is caring for, facilitating access to, preserving and adding value to research data throughout its lifecycle.

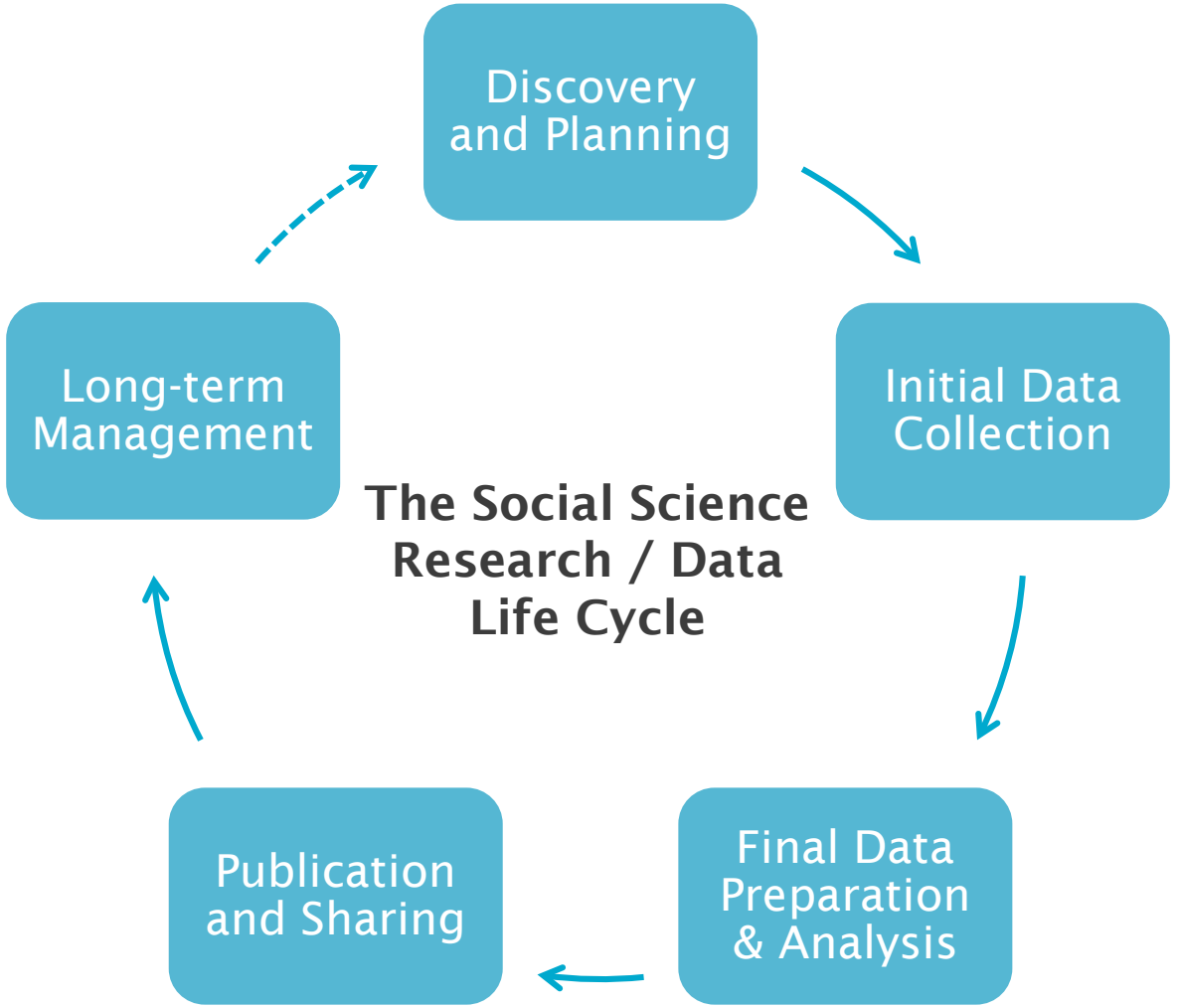
Source: University of Edinburgh Information Services

A **data management plan** (DMP) helps researchers consider *during the research design and planning stage*, how the data will be managed *during the research process itself* and potentially shared *afterwards* with the wider research community.

Why manage research data well ?

- Your data creation is likely to be expensive
- Your data underpin your published findings
- Good quality data = good quality research
- Protect your data from loss, destruction
- Compliance with ethical codes, data protection laws, journal requirements, funder policies
- **To benefit your future self**

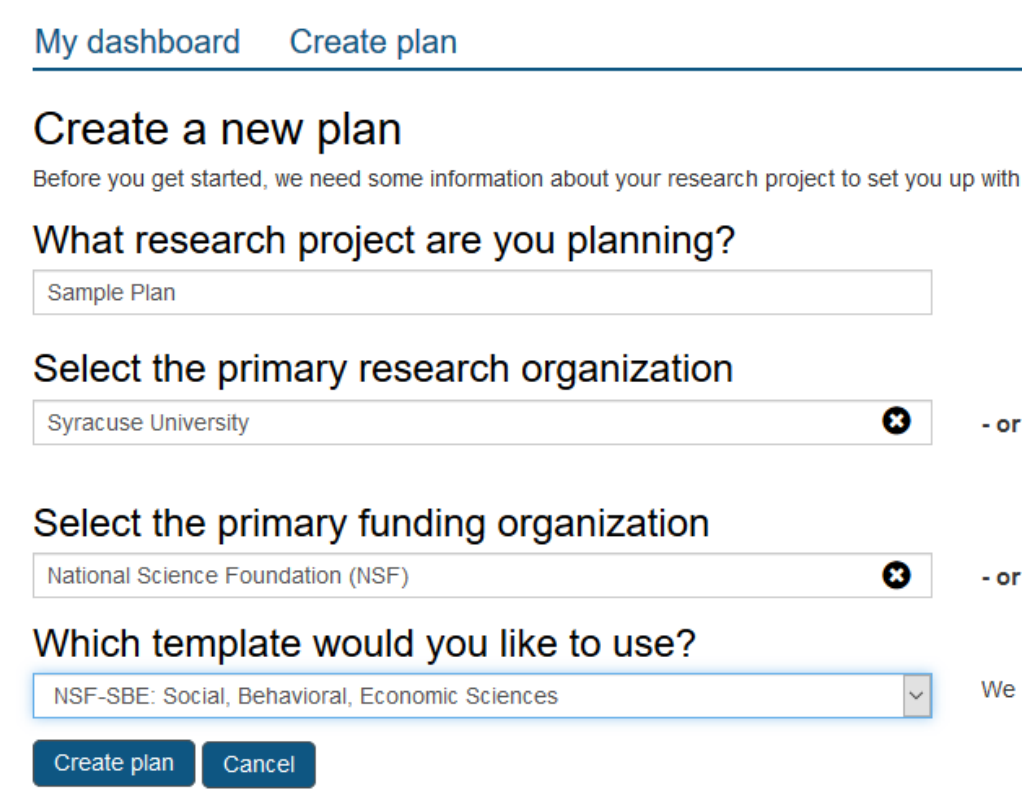
Research / Data lifecycle

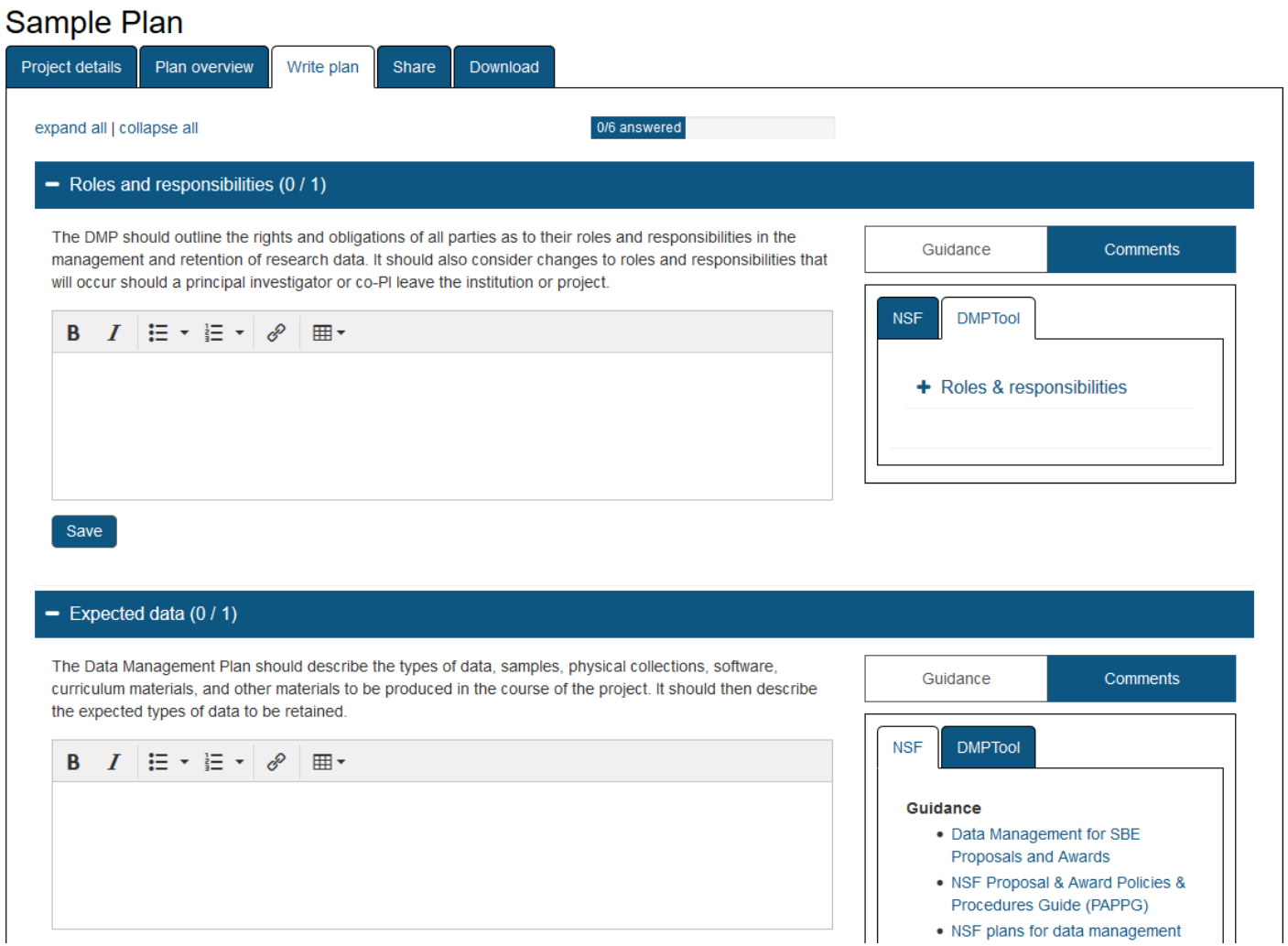


Based on Green and Gutmann, 2007

There's an App for That

https://dmptool.org





Topics of a DMP

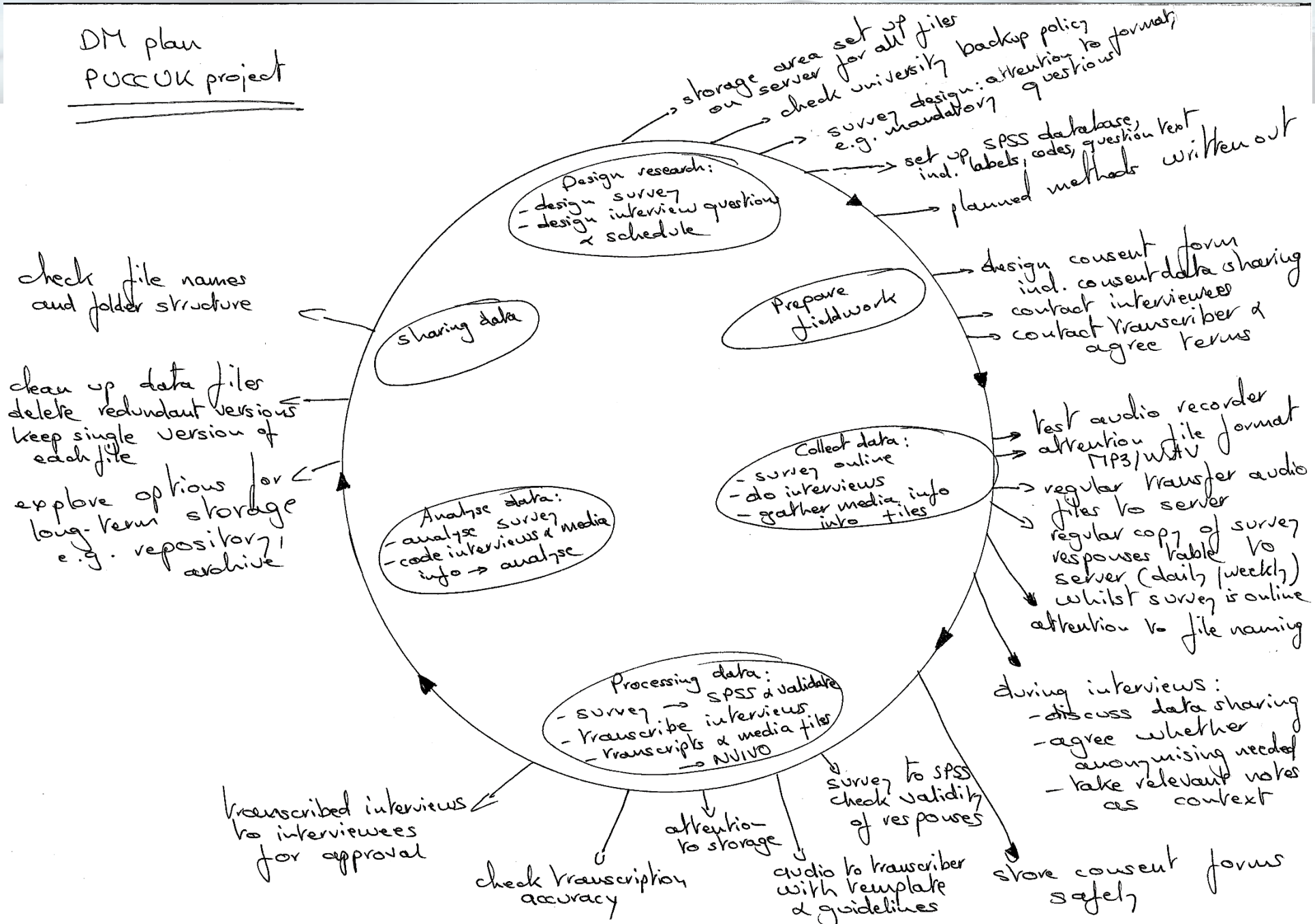
- Kinds of data that are being created
- Any applicable data sharing policies
- File formats
- Data descriptions, standards & metadata
- Data storage
- Access and use, incl. appropriate restrictions
- Intellectual property ownership / copyright
- Human participant constraints
- Roles and responsibilities in a team
- Budget for data activities

DMP Checklist will be handed out: available at

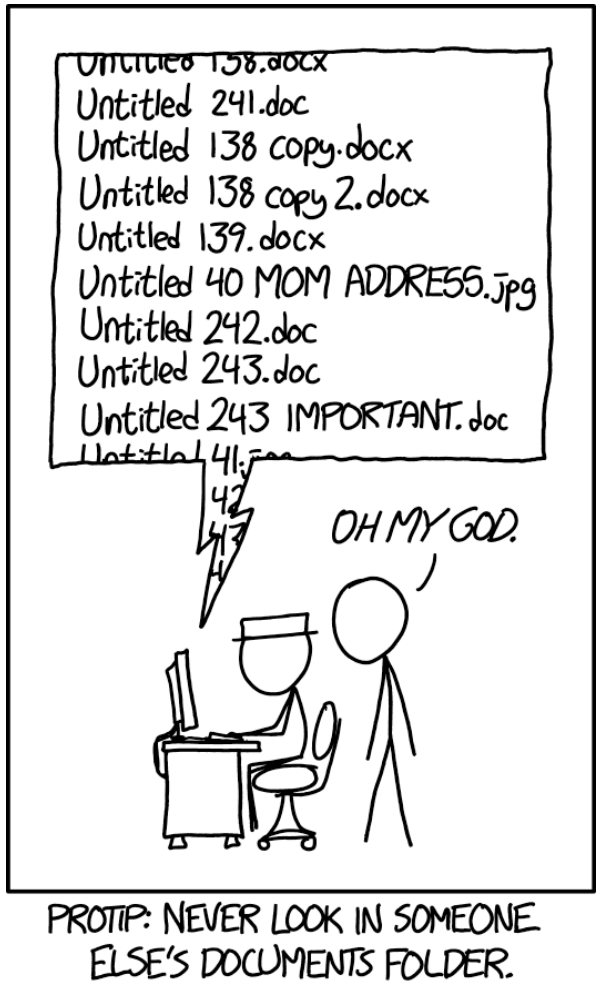
Key planning issues

- Know your legal, ethical and other obligations towards research participants, colleagues, research funders and institutions
- Know your institution's policies and services: storage and backup strategy, research integrity framework, IPR policy, local or other recommended data repository
- Assign roles and responsibilities to relevant parties
- Implement and review management of data during project meetings and review
- Include cost of data management into research applications / research proposals / DMPs

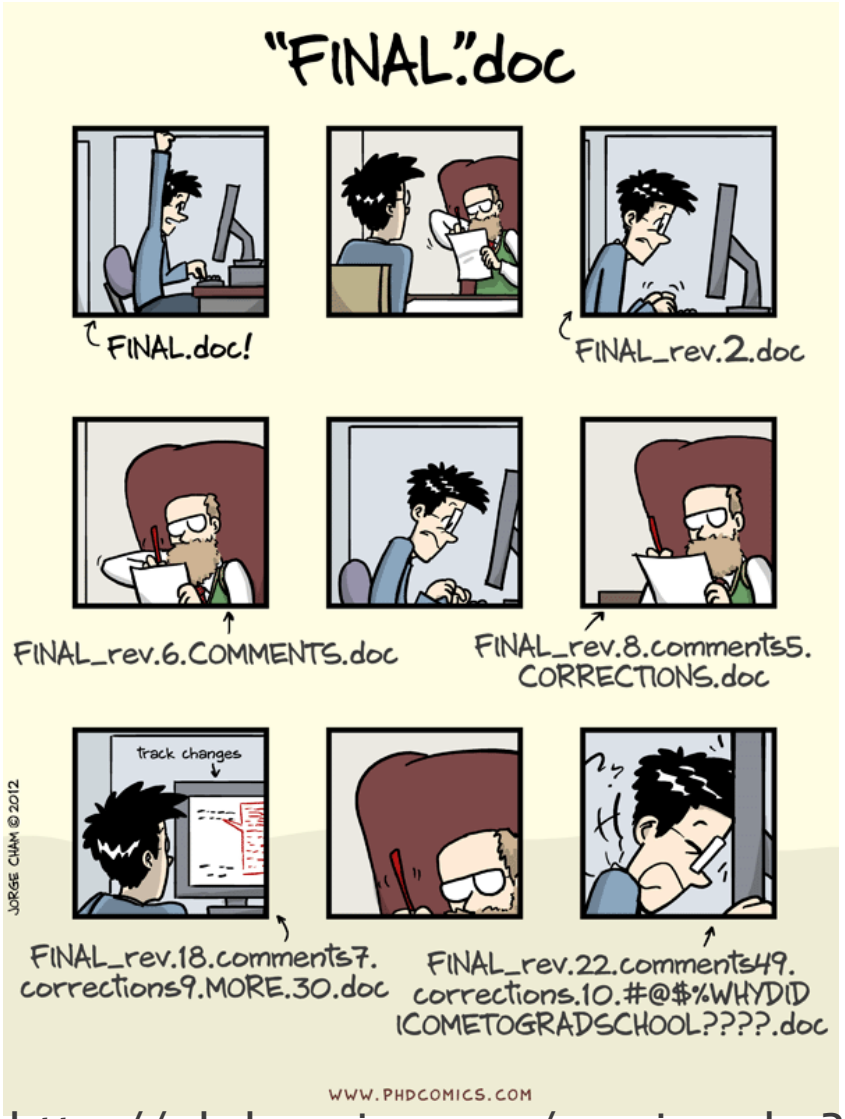
DM plan
PUCCUK project



Start with the basics…



https://xkcd.com/1459/



http://phdcomics.com/comics.php?f=1531

Why document your data and processes?

- Enables you to understand data when you return to them
- To make data and research understandable to others, i.e. reusable and verifiable
- Helps avoid incorrect use/misinterpretation
- Data documentation is critical for sharing the data via a repository in order to:
  - Supplement a data collection with documents such as user guide(s) and data listing
  - Ensure accurate processing and archiving
  - Create a catalog record for a published data collection

Guiding question: If using your data for the first time, what would a new user need to know to make sense of it?

What should be captured?

- Contextual information about project and data
  - Background, project history, aims, objectives, hypotheses
  - Formal publications based on data collection
  - Final reports, working papers, lab books
  - User guides / ReadMe – type files orienting secondary users to the data
  - Anything else you as the creator of these data think would be useful for future comprehension…

What should be captured?

- Data collection methodology and processes
  - Data collection process and sampling choices
  - Instruments used: questionnaires, show-cards, interview schedules, topic guide
  - Temporal/geographic coverage
  - Data validation - cleaning, error-checking
  - Compilation of derived variables (QUAL EX: codes you develop as part of content analysis)
  - Citations for any secondary data sources used

What should be captured?

- Information on data files structure
  - Inventory of files
  - Relationships between those files
  - Units, records, cases…
  - Multiple versions

- Variable-level documentation
  - Labels, codes, classifications
  - Missing values
  - Derivations and aggregations

- Data confidentiality, access and use conditions
  - De-identification carried out (de-identification protocol)
  - Participant consent and copyright conditions/forms/procedures
  - Access or use conditions of data

Consider documentation early on

- Good documentation and metadata depends on what you can provide
- What you can provide depends on what you can remember
- Start gathering meaningful information *from as early on in the research process as possible*

Project-level documentation Example: data list

- Data listing provides an at-a-glance summary of data files

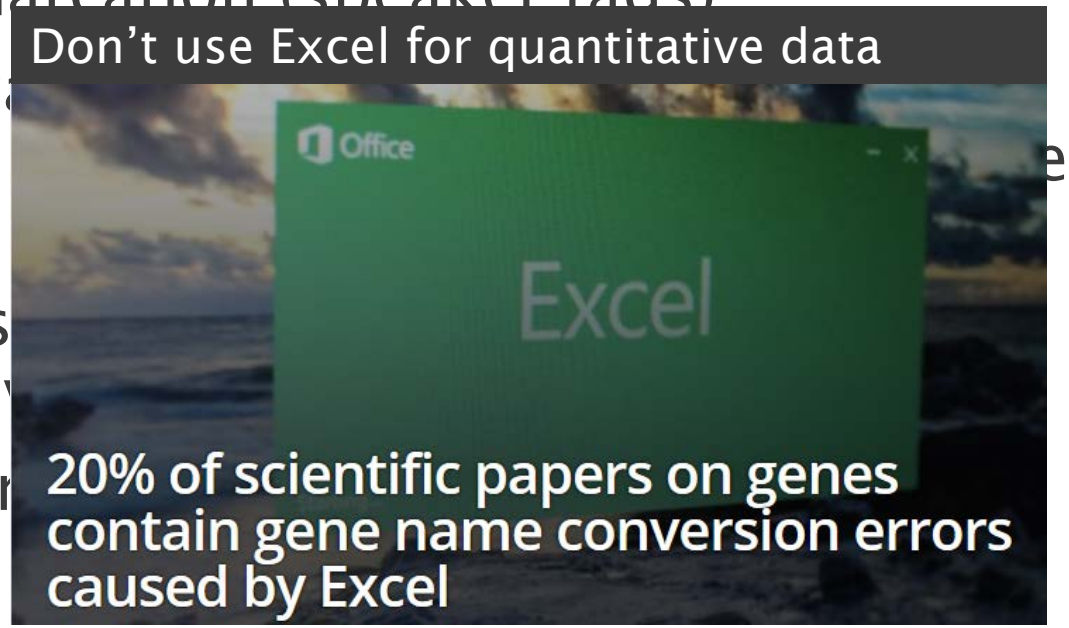
QDR Project: Cassese				
Visuals and Text, Trump				
Headline	**Publication/Website name**	**Author**	**Date**	**Original URL**
Horror-Clown! How papers around the world reacted to Do	The Independent	Lily Pickard	November 9, 2016	http://www.ind
How did this monster get created? The decades of GOP lie	Salon	Heather Cox Ri	July 19, 2015	http://www.sal
I collect monsters. Visual Monsterisation Strategies.	icollectmonsters.tumblr.com	icollectmonster	November 13, 2016	http://icollectm
Marvel Comics' Latest Villain is Monstrous Donald Trump	comicbook.com	Lucas Siegel	July 1, 2016	http://comicbo
Naked Donald Trump statues pop up in several cities	The Hill	Nikita Vladimiro	August 18, 2016	http://thehill.cc
The Frankenstein Monster Speech: Trump Was a Rampagi	Alternet	Chauncey DeV	October 14, 2016	http://www.alte
Trump's Monstrous Call	The Huffington Post	Jedediah Purdy	December 7, 2015	http://www.huf
Trumpmonsters	Storify	Jeffrey Jerome	July 2016	https://storify.c

Clinton Headlines | Clinton Images | Trump Headlines | Trump Images | **Trump Visuals and Text**

File-level documentation suggestions

Embed documentation in your data files

- Interview transcript speech demarcation (speaker tags)
- Document header with brief deta
  interviewer name (unless trying
  details, context
- Stata/R/SPSS: variable attributes
  (label, code, data type, missing
- Excel: document properties, wor



https://doi.org/10.1186/s13059-016-1044-7

In practice: Documentation in transcript

Raúl L. Madrid Interview with Felipe Quispe,
Leader of Movimiento Indigena Pachacuti (MIP)
La Paz, Bolivia, July 29, 2004

Quienes crearon el MIP?

Un poco de introducción. Ese movimiento indígena viene desde los anos 70s. Luego hemos estado en los Ayllus Rojos. Se ha transformado en el Ejercito Tupak Katari. Por dos años actuamos en acciones revolucionarias. Nos capturaron. Nos encarcelaron. Estuve 5 años en la cárcel. En 1997 salí. Todavía tengo proceso.

Source: https://doi.org/10.5064/F6MS3QNV

Metadata – data about data

- Highly structured documentation

- Data collection metadata examples:
  - Components of a bibliographic reference
  - Core information that a search engine indexes to make the data findable

- International standards/schemes
  - Data Documentation Initiative (DDI)
  - Dublin Core

Excerpt from QDR catalog record metadata

Metrics 15 Downloads

Contact

Data for: When do the dispossessed protest? Informal leadership and mobilization in Syrian refugee camps

Clarke, Killian B. 2018. "Data for: When do the dispossessed protest? Informal leadership and mobilization in Syrian refugee camps". Qualitative Data Repository. https://doi.org/10.5064/F6CN723S. QDR Main Collection. V1. UNF:6:/UwXUp2V/F3WLS344J/Sxg==

Cite Data Project

Learn about Data Citation Standards.

Description

Project Summary

Refugees are often considered to be among the world's most powerless groups; they face significant structural barriers to political mobilization, including often extreme poverty and exposure to repression. Yet despite these odds refugee groups do occasionally mobilize to demand better services and greater rights. This paper examines varying levels of mobilization among Syrian refugees living in camps and informal settlements in Turkey, Lebanon, and Jordan in order to explain how marginalized and dispossessed groups manage to develop autonomous political strength. I explain the surprisingly high levels of mobilization in Jordan's Za'atari Camp, compared to the relative quiescence of refugees in Turkish camps and Lebanese informal settlements, as the product of a set of strong informal leadership networks. These networks emerged due to two unique facets of the refugee management regime in Jordan: 1) the concentration of refugees in the camp, and 2) a fragmented governance system. In Turkey and Lebanon, where these two conditions were absent, refugees did not develop the strong leadership networks necessary to support mobilization. I develop this argument through structured comparison of three cases and within-case process tracing, using primary source documents from humanitarian agencies, contentious event data, and 87 original interviews conducted in the summer of 2015.

Data Abstract

Files Description

The data are of two types: humanitarian documents and web-scraped material on protest events. The humanitarian documents are organized by topic. First are two documents with statistical information on the camp. Following these are files with the meeting minutes for the weekly Camp Coordinating Meeting in Zaatari (50 total). Then come meeting minutes for the community mobilization initiative. Next come two security reports and the camp governance plan. Finally, there are four maps. A full list of these files, including their original URL, generated by QDR curators, is also included as documentation. Most of these documents are in PDF format.

Second, the catalog of protest events in Lebanon are included as a spreadsheet. Each event listed includes the details of the event as well as links to web content (usually news articles or social media posts) corroborating the occurrence of the event. These links were also archived using perma.cc by QDR, and the perma.cc links are included as well. This list is included in its original Excel (.xlsx) format as well as tab-separated values (.tsv)

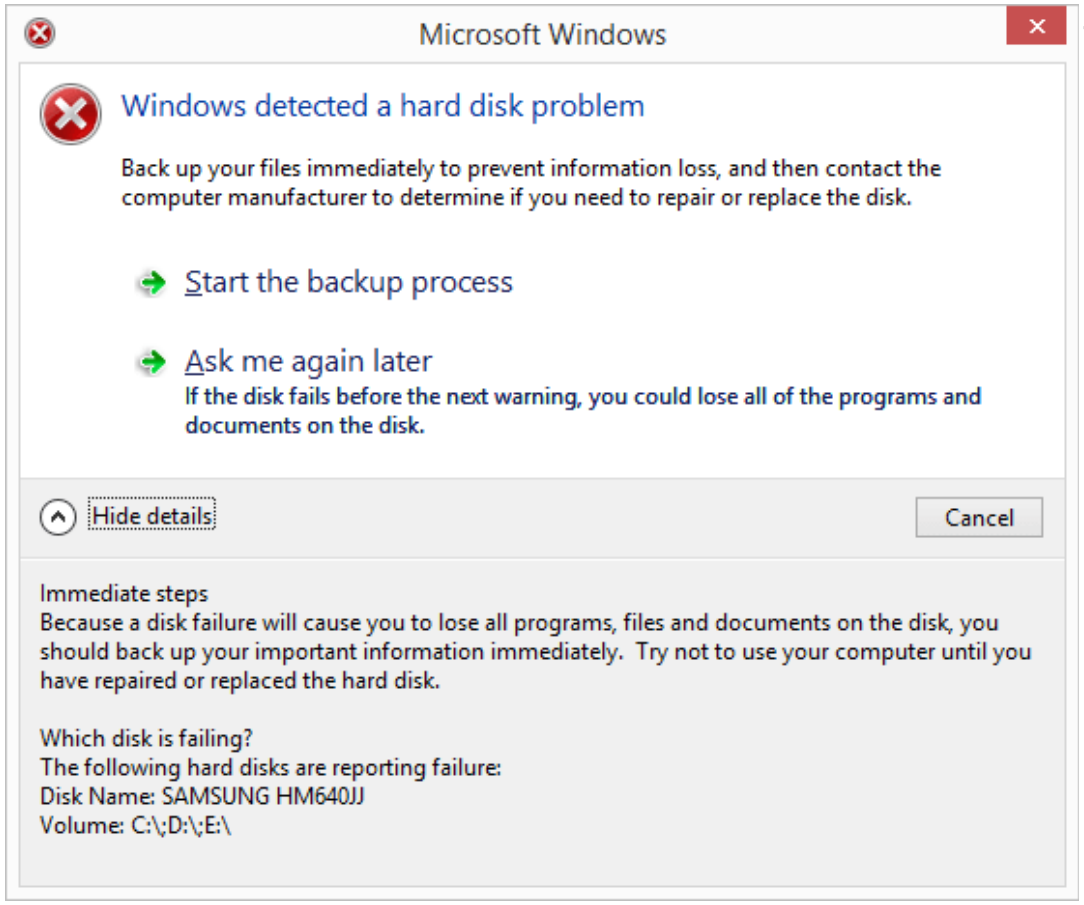
Social Sciences

refugees (ICPSR Subject Thesaurus) https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/index
world problems (ICPSR Subject Thesaurus) https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/index
social protest (ICPSR Subject Thesaurus) https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/index
social networks (ICPSR Subject Thesaurus) https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/index
United Nations (ICPSR Subject Thesaurus) https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/index
humanitarian aid (ICPSR Subject Thesaurus) https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/index
international assistance (ICPSR Subject Thesaurus) https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/index

Time Period Covered	Start: 2012-06-01 ; End: 2016-01-01
Date of Data Collection	Start: 2015-06-01 ; End: 2015-09-01
Type of Data Project	Qualitative data project; Supplementary data project
Geographic Coverage	Syrian Arab Republic Jordan Turkey Lebanon Middle East
Language	Arabic; English
Distributor	Qualitative Data Repository (Syracuse University) (QDR) https://qdr.syr.edu/
Distribution Date	2018-05-22
Depositor	Clarke, Killian

Stuff happens: Dissertation nightmares

Source: lilysussman.wordpress.com



Source: http://graphics-unleashed.com/2015/04/smart-warning-predicts-imminent-hard-disk-failure/

Your can lose your data in various ways (one at least gives you a good story…)

Backing-up data

- It's not a case of *if* you will lose data, but *when* you will lose data!
- Digital media are particularly fallible
- Keep additional backup copies
  - Rule #1: 3 versions in 2 locations
  - Rule #2: Regular, automatic, incremental
  - Check that backups work; copy data files to new media every 2-5 years
- Protect against: software failure, hardware failure, malicious attacks, natural disasters, YOURSELF!

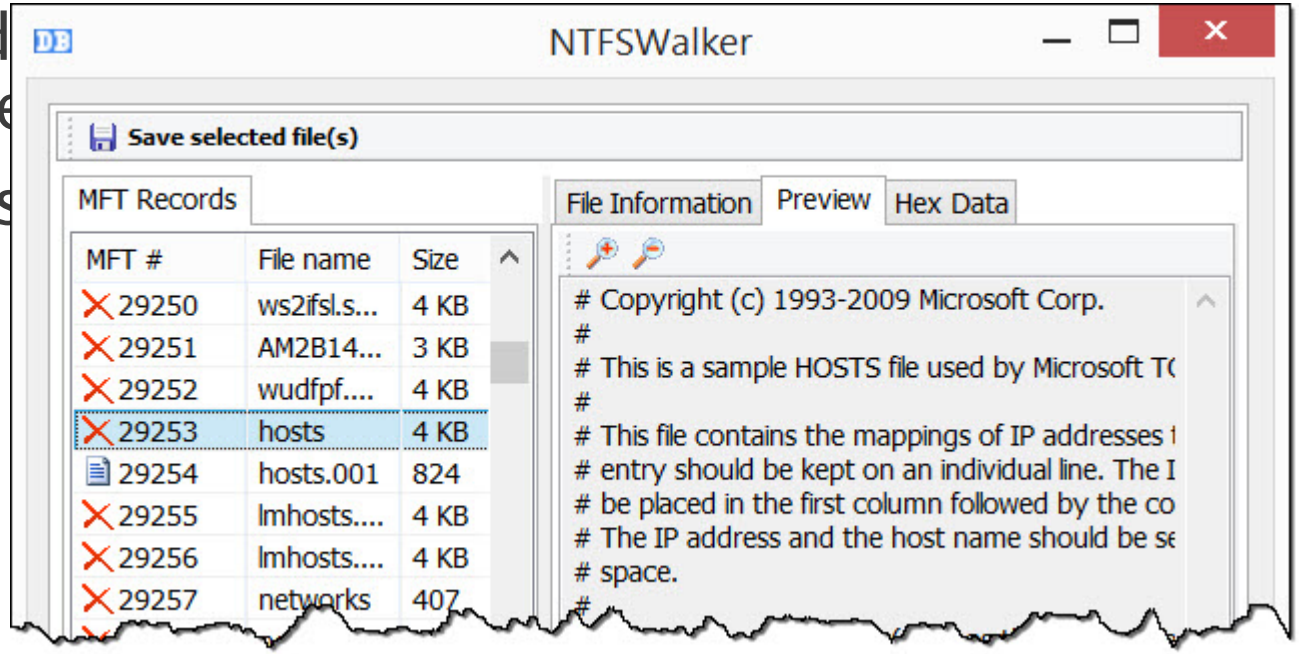
Cloud storage services

- Online or ‘cloud’ services increasingly popular
  - DropBox, Box.com, OneDrive, Google Drive etc.
  - Very convenient
  - Background syncing
  - Mobile apps available
- Use, but use with care:
- Consider if appropriate, as services can be hosted outside your country (personal data laws)
- Encrypt anything sensitive (e.g.: VeraCrypt) or
- Look for services with end-to-end encryption, aka, “zero knowledge”
  - Often paid; did you budget for that?
- Your university’s IT may have rules & services for this

Data destruction

Beware of mandates to destroy the data but, if required, keep the following in mind:

- When you delete a file from a hard drive, it's still retrievable – even after emptying the recycle bin
- Files need to be overwritten (id
  random data to ensure they are
- Free file and folder-shredding s



How to share data ethically/ legally

- Obtain informed consent, including ***explicitly*** for data sharing and long-term preservation / curation
- Protect identities, e.g., not collecting personal details when unnecessary during data collection; de-identification after the fact
- Regulate access where needed (all or part of data), e.g., by group, use, time period
- Securely store personal or sensitive data (separately)

Planning is key! (again)

- Collecting identifying information
  - Avoid collecting unless necessary
  - Where confidential: Keep directly identifying info separate and secure

- Informed consent – an active process
  - Be careful with restrictions in consent script
  - Oral vs. written consent
  - Cultural context
  - Ask for permission for data sharing explicitly (& include in IRB application)

In practice: wording in consent form / information sheet

We expect to use your contributed information in various outputs, including a report and content for a website. Extracts of interviews and some photographs may both be used. We will get your permission before using a quote from you or a photograph of you.

After the project has ended, we intend to archive the interviews at . . . Then the interview data can be disseminated for reuse by other researchers, for research and learning purposes.

The interviews will be archived at . . . and disseminated so other researchers can reuse this information for research and learning purposes:

- ❑ I agree for the audio recording of my interview to be archived and disseminated for reuse
- ❑ I agree for the transcript of my interview to be archived and disseminated for reuse
- ❑ I agree for any photographs of me taken during interview to be archived and disseminated for reuse

De-identifying qualitative data

- Removing or replacing information in text can distort data, make them unusable, unreliable or misleading: A balance to preserve context
- Remove direct identifiers, or replace with pseudonyms – often not essential research info
- Avoid blanking out; use pseudonyms or replacements (Identify replacements)
- Plan or apply editing at time of transcription
- Consistency within research team /project
- Keep de-identification log of all replacements or removals made; keep separate from anonymized data files

De-Identification Requires Context Expertise

Entrevistador: ¿Y en qué barrio pensás, digamos, cuando, vos en qué barrio..?

Entrevistada: Bueno, yo soy BARRIO 1 y BARRIO 2. BARRIO 2 y BARRIO 1.

Entrevistador: Y, digamos, si tuvieras que, de esos referentes que conocés de diferentes partidos, dar un número, ¿te animás a dar un número, cuántos son?

Entrevistada: Y, son unos cuantos, son muchos eh, yo entiendo que son más de cincuenta

Entrevistador: Bueno, cincuenta me parece...

Entrevistada: Te digo, por ejemplo, en MUNICIPIO 1 hay [detalla el número] bibliotecas populares, reconocidas por la CONABIP, que es una institución nacional que las agrupa y

De-identification log

EXAMPLE: Anonymization Log of Interview Transcripts		
Interview/Page	Original	Changed To
Int1		
p1	Spain	European country
p2	E-Print Ltd	printing company
p2	20th of June 1989	Summer of 1989
Int2		
p1	Amy	Ms. Z
p2	Francis	Ms. Z's neighbor

Data Sharing and Copyright

COPYRIGHT – an intellectual property right assigned automatically to the creators of "original works of authorship" (*title 17, U.S. Code*), which prevents unauthorized copying and publishing of an original product

- Who owns copyright?
- Copyright and research materials
- Interviews and copyright
- Clearing copyright – before reproduction, sharing, SDA
- Data repositories hold no copyright

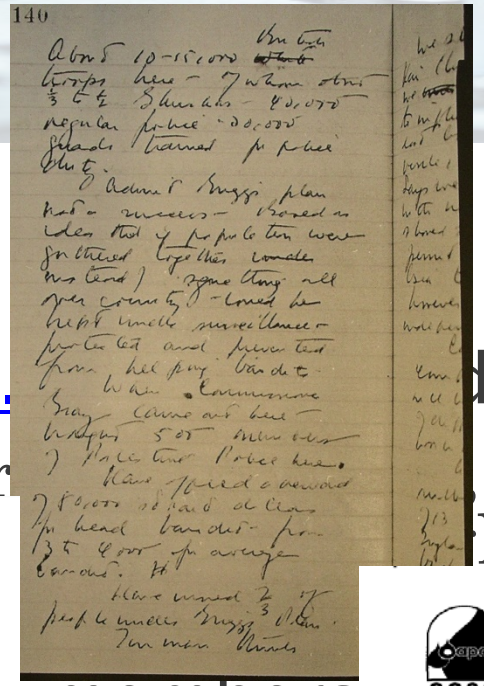
Fair Use and Public Domain

- Fair use exemption = key part of copyright law that permits the unlicensed use of copyrighted material under some circumstances (study, teaching, quotations, criticisms, review)
- Fair use claims in research/scholarship because of the way in which data are used
  - Transformative
  - Not involving a large amount
  - Not likely to affect the potential market value of items
  - Is used for academic / non-commercial purposes
- Public Domain
  - US: Government documents
  - Copyright has run out

Sharing Data in a Repository

- Offer stability over time
- Assign unique DOIs to data and ensure long-term preservation of digital assets.
- Institutions may require this
- Can manage data in a way that maintains their understandability and usability for the scientific community → CURATION!
- Makes data more visible/easier for other scholars to discover, access, use, cite
- Interoperability across disciplines
- Allows scholar to set limits on who can access the data and how they can be used

What Is QDR?



- Online since 2014: [qdr.syr.](#) ded
  QDR curates, stores, preser an
  h qualitative and
  members
  ce: Interna
  published

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.

Naucalpan de Juárez 2006-2009

GERENCIA DE AGUA POTABLE

HORARIOS DE SERVICIO DE AGUA POTABLE POR COLONIA EN NAUCALPAN

No.	POBLACION	TIEMPO DE SERVICIO	HORARIO DE SERVICIO (HORAS)
	PUEBLOS		
1	LOS REMEDIOS	Diario	8 hrs.
2	SAN ANTONIO ZOMEYUCAN	c/3er día	24 hrs.
3	SAN BARTOLO NAUCALPAN (NAUCALPAN CENTRO)	Diario	24 hrs.
4	SAN ESTEBAN HUITZILACASCO	Diario	24 hrs.
5	SAN FRANCISCO CHIMALPA	No existe infraestructura hidráulica operada por este Organismo	
6	SAN FRANCISCO CUAUTLALPAN	Diario	24 hrs.
7	SAN JOSÉ RÍO HONDO	Diario	24 hrs.
8	SAN JUAN TOTOLTEPEC	Diario	24 hrs.
9	SAN LORENZO TOTOLINGA	c/3er día	24 hrs.
10	SAN LUIS TLATILCO	Diario	14:00 a 6:00 hrs
11	SAN MIGUEL TECAMACHALCO	Diario	24 hrs.
12	SAN RAFAEL CHAMAPA	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	

Access Controls – QDR's Practices

- Data deposited are *not* in the public domain or open access
  - Only accessible after registration – data users agree to legally binding General Terms & Conditions of Use
  - Use of data is only allowed for specific (research and teaching) purposes
  - → *e.g., not make efforts to identify any individuals*

- Several degrees of stricter access regulation for sensitive data (case-by-case) available
  - *conditional online access; depositor-approved users; offline access; time-based embargo*

Group Exercises

Group A You brought a DMP	Group B You brought notes to create a DMP	Group C You did not bring own project
• Exchange your DMP with neighbor. Assess their DMP using the rubric handout (*20mins*) • Discuss DMPs with neighbor (*10 mins each)*	• Work on your DMP (*40 mins*) • Use Research Lifecycle handout to plan steps • Use DMPtool to help structure your writing: https://dmptool.org	• Assess the DMP handed out using the rubric handout. (*20mins*) • Compare notes with your neighbors (*10mins*) • In small group (4-6), discuss benefits & challenges of using rubric (*10mins*)