Working with Qualitative Data: Cleaning and Coding

Celia Emmelhainz @anthrolib
Anna Sackmann @sackmanna
UC Berkeley

Introductions

Celia Emmelhainz
Qualitative Research Librarian

Anna Sackmann
Data Services Librarian

What you can expect to learn today:

How to clean messy spreadsheet data

How to anonymize textual data

Understand different approaches to coding qualitative data

How to import and code qualitative data in Taguette

Today's topics:

The qualitative research lifecycle

Cleaning qualitative survey data in spreadsheets

Anonymizing textual data

Importing data into Taguette

Tagging / coding texts

Thinking conceptually / clustering by theme

Exporting data for use in presentations and publications

Packaging qualitative data for future re/use

The Qualitative Research Lifecycle

5. Share
Check with respondents
Present and publish
Deposit your data

1. Plan
- Ask ethnographic questions
- Determine best methods
- Select a research site

Qualitative Research Lifecycle

4. Code
Analyze data
Cluster themes
Identify key trends

2. Connect
Collect data from people:
- Interviews
- Focus groups
- Observations
- Surveys

3. Clean
Organize data, transcribe audio, anonymize, prep for coding

Tidy Data / Cleaning Data

Spreadsheets!

We're working in Excel today, but there are plenty of other options!

- LibreOffice
- Google Sheets
- Apple Numbers

Spreadsheets!

We're working in Excel today, but there are plenty of other options!

- LibreOffice
- Google Sheets
- Apple Numbers

What matters: always save or export your spreadsheet as a .csv file.

Spreadsheet Rules

A few rules of thumb:

1. Only use the spreadsheet for data entry
2. Never create figures in the spreadsheet
3. Organization within the spreadsheet:
   a. Utilize tabs
   b. Keep track of your clean-up and anything else you do to the raw data

Spreadsheet organization

1. Enter variables (the thing being measured, question being asked) in columns
2. Each observation gets its own row (e.g. survey response)
3. Never put more than one piece of information in a cell

Recipient Email	Do you have permanent status?	How is the pandemic affecting your work now?	Years as a librarian
librarian@berkeley.edu	yes	I have more work	17
librarian@ucsf.edu	no	I have taken on additional responsibility.	2

Spreadsheet organization continued

4. Never create more than one table in a sheet

5. Don't necessarily create an additional table in another tab

6. Avoid merging cells

Instead:

- Could you add another column to the existing sheet?

Naming and formatting

Even when working with qualitative data, we're likely to have some quantitative data.

For example:

- Dates:
  - July 14, 2014 → 41822
  - Jul-10 → 1-Jul-10
- Follow ISO 8601! YYYYMMDD
  - August 1, 2021 → 20210801
  - Extended ISO 8601: YYYYMMDDhhmmss

Exporting Data from Spreadsheets

Never save spreadsheet data as a proprietary spreadsheet file.

Why?

- The software may cease being used.
- Versions of proprietary software may treat the data differently.
- Other software may not be able to open the file.

Anonymizing Data

Q: Why would we anonymize data?

What's good about these anonymizations? What isn't working?

Example A: "So my first workplace was X which was about X minutes from my home in X. My best colleagues from day one were X, X and X and in fact, I am still very good friends with X to this day. X lives in the same parish still with her husband X and their X X."

Example B: "So my first workplace was [name] which was about 20 minutes from my home in Norwich. My best colleagues from day one were Andy, Julie and Louise and in fact, I am still very good friends with Julie to this day. She lives in the same parish still with her husband Owen and their son Ryan."

Tips on removing sensitive information:

Switch **direct identifiers** (name, date, place) or generalize them:

> Anna → [**Beth**] *or* [**Friend 2**]

Remote or generalize **indirect identifiers,** which can ID when combined:

> born in 1996 → born in [**the 1990s**]

> a white data librarian → white [**academic**] librarian

> in Berkeley → in [**the Bay Area**], [**California**], or [**the Western US**]

Also consider redacting sensitive comments or data points:

> "I hate my boss!" → [**comment redacted**]

> "Treasure is at lat **-122.258560**, long **37.872589**" → [**location redacted**]

Track changes in an encrypted anonymization log!

Interview and page number	Original	Changed to
Int1		
p1	Age 27	Age range 20-30
p1	Spain	European country
p3	Manchester	Northern metropolitan city / provincial city
p2	20th June	June
p2	Amy (real name)	Moira (pseudonym)
Int2		
p1	Francis	my friend
p8	Station Road primary school	a primary school

Try it out!

- choose a sample interview
- read and make notes separately
- discuss in your group

How did it go?

Break (10 min)

Moving into Taguette

Why are we using Taguette today?

- Free and open source!

- Flexible tagging options (created by you)

- Lower barrier to entry → a good option for novice learners

- Data and tags stay your own → easy to import and export!

Getting Started with Taguette

Use their server: app.taguette.org

    Register at: app.taguette.org/register

Install a local copy on Windows or Mac: www.taguette.org/install.html

    (Current installer at gitlab.com/remram44/taguette/-/tags/v1.0.0)

    Windows is 108 MB -- do before workshop??

Login to Taguette at app.taguette.org

TAGUETTE Home   Taguette Guide ☑ Log in
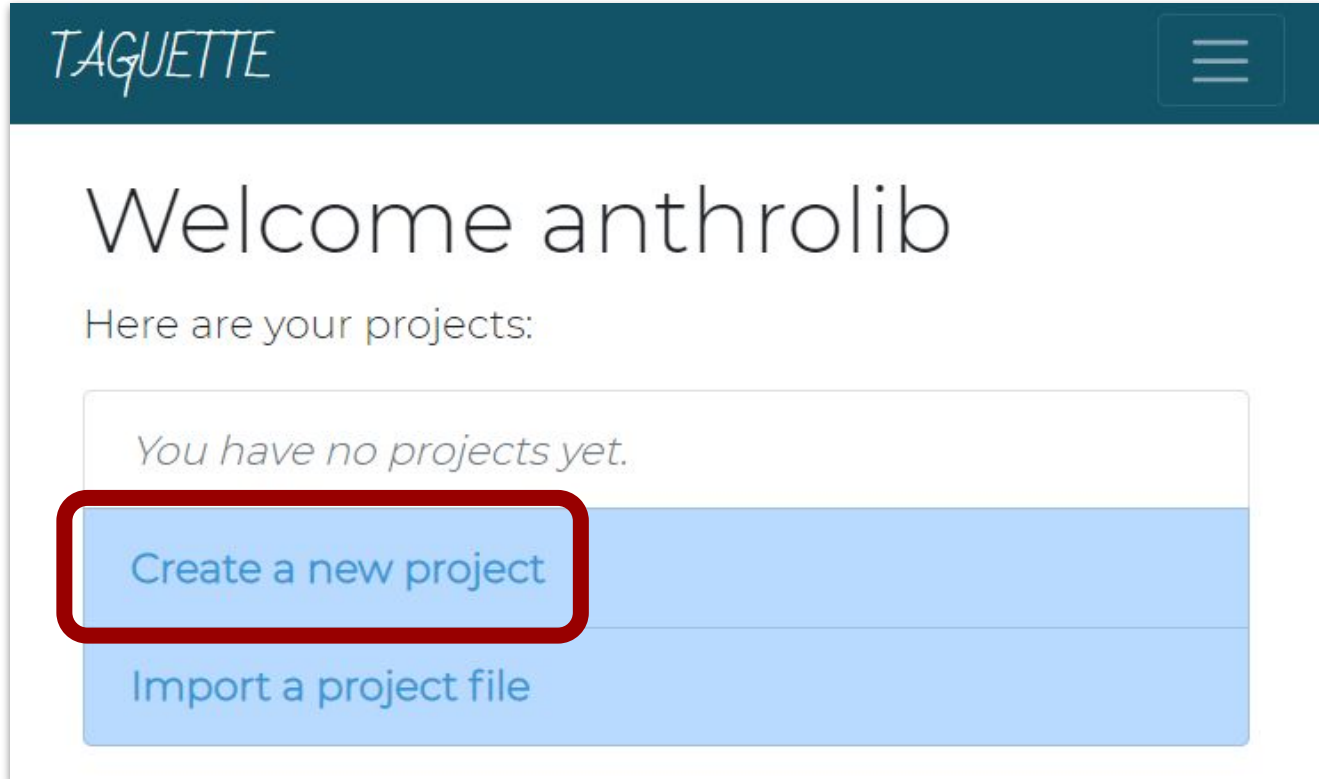
Welcome

TAGUETTE is an open-source web-based document tagging tool for qualitative data analysis.

Using this tool, you can upload a collection of documents, create a hierarchy of tags, and annotate portions of documents with tags and notes that you can recall and organize later.

Register now for free and get started!

Create a new project:

Add a memorable title and description:

Add collaborators (optional)

"Add a document" and select file:

Project info   Documents

Highlights

Add a document

There are no
documents in this
project yet.

☐ Backlight

Load a document on the left

Add a document   ✕

Enter a document name   Choose File   No file chosen

● Left to right   ○ Right to left

Description
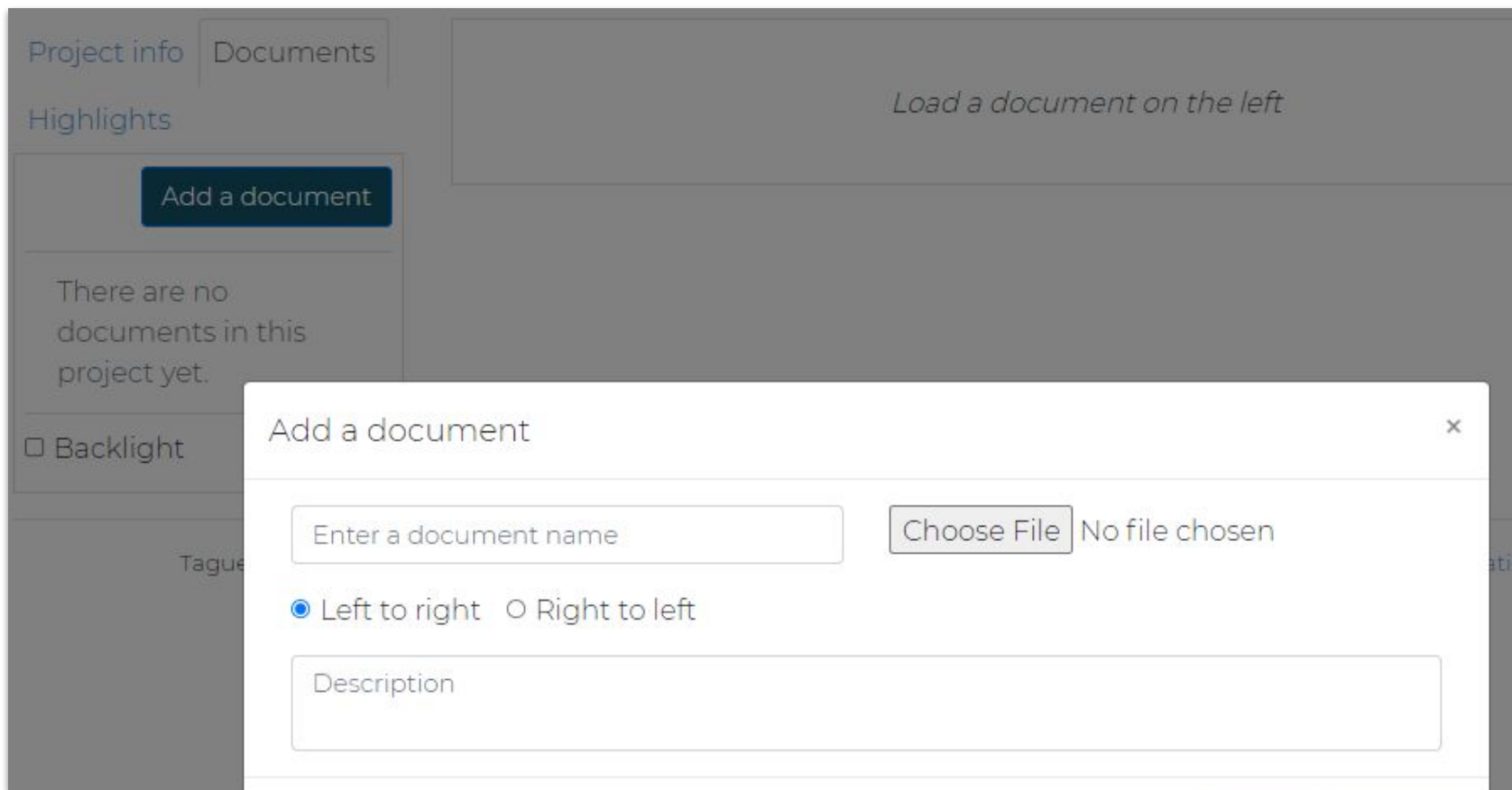
Add a title and contextual information to each file:

Add a document   ×

Int2 Meruert | Choose File | Librarian Int...leaned.docx

● Left to right   ○ Right to left

Interview with Bulbul, a Central Asian librarian.

Import   Cancel

What types of data could you gather and code?

What Taguette handles:

Taguette uses Calibre to convert your documents to HTML for tagging. It can work with:

- Microsoft Word files (.docx)
- LibreOffice files (.odt)
- PDFs (.pdf)
- PLain text (.txt)
- Rich text (.rtf)
- HTML (.html)

It doesn't currently support audio, video, image, or spreadsheet files. You may find it easier to code those in a commercial qualitative software!

10 minute break!

What do I do with this data?

1. Read

Gather your anonymized research materials and read them over.

You can highlight quotes and add comments directly to them in some software (or using 'comment' feature in Google Docs or Word, much as you would with a friend's essay).

You can also take notes ('memos') on each document and your thoughts in another notebook:

What does this remind you of?

What's interesting?

What do you want to know more about?

Try it!
Read one interview and take notes.

2. Create codes and apply them to the text.

Merge similar tags (optional)

Highlight text related to your tags

Apply a tag to your highlighted section:

Add a new tag to a highlighted section:

View your highlighted sections

Project info   Documents

Highlights

See all highlights

Create a tag

career-exploration 1 Edit

childhood-dreams 3 Edit

family-advice 0 Edit

family-concerns 1 Edit

family-consideratio 0 Edit
ns

interesting 0 Edit

Export this view ▼

Q: What profession were you most interested in as a child?

A: From my childhood, I really liked to visit the school library, and I got this dream to be a librarian. When I finished my school, I decided to apply and study for library science. I'm not close to math and physics, but more to humanities. That's why I chose library science.

Int1 Bulbul childhood-dreams

Q: What profession were you most interested in as a child?

A: I wanted to be a teacher, I was mostly inspired by my teachers in the village school. I grew up in a small village in [Kazakhstan], and all the teachers were very big people, very smart, and so I wanted to be like them. From grade 6, I started to dream about becoming an English teacher.

Int2 Meruert childhood-dreams

Q: What profession did you want as a child?

A: I wanted to be an English language teacher, but then I realized teaching isn't a very grateful [appreciated] job in [Tajikistan]. Then I wanted to be a lawyer or prosecutor, but then I understood that I'm too fair-minded for that work. [implies that she wouldn't take bribes].

Int3 Botagoz childhood-dreams

View your highlighted sections

Project info   Documents

Export this view ▾

Highlights

See all highlights

Create a tag

career-exploration ① Edit

childhood-dreams ③ Edit

family-advice ⓪ Edit

family-concerns ① Edit

family-consideratio
ns ⓪ Edit

interesting ⓪ Edit

Q: What profession were you most interested in as a child?

A: From my childhood, I really liked to visit the school library, and I got this dream to be a librarian. When I finished my school, I decided to apply and study for library science. I'm not close to math and physics, but more to humanities. That's why I chose library science.

Int1 Bulbul childhood-dreams

Q: What profession were you most interested in as a child?

A: I wanted to be a teacher, I was mostly inspired by my teachers in the village school. I grew up in a small village in [Kazakhstan], and all the teachers were very big people, very smart, and so I wanted to be like them. From grade 6, I started to dream about becoming an English teacher.

Int2 Meruert childhood-dreams

Q: What profession did you want as a child?

A: I wanted to be an English language teacher, but then I realized teaching isn't a very grateful [appreciated] job in [Tajikistan]. Then I wanted to be a lawyer or prosecutor, but then I understood that I'm too fair-minded for that work. [implies that she wouldn't take bribes].

Int3 Botagoz childhood-dreams

View your highlighted sections

Taguette highlights: childhood-dreams

Q: What profession were you most interested in as a child?

A: From my childhood, I really liked to visit the school library, and I got this dream to be a librarian. When I finished my school, I decided to apply and study for library science. I'm not close to math and physics, but more to humanities. That's why I chose library science.

Document: Int1 Bulbul **Tags:** childhood-dreams

Q: What profession were you most interested in as a child?

A: I wanted to be a teacher, I was mostly inspired by my teachers in the village school. I grew up in a small village in [Kazakhstan], and all the teachers were very big people, very smart, and so I wanted to be like them. From grade 6, I started to dream about becoming an English teacher.

Document: Int2 Meruert **Tags:** childhood-dreams

Try it!

Try coding text in Taguette.

Thinking conceptually about coding

3. Look deeper

Once you've tagged topics, trends, or emotions, explore further:

- Use "merge" to cluster together sparse or related themes

- If you've got too many codes, set some aside. Focus on your core question or most well-supported finding!

- Focus on one or two larger codes. Look back at the quotes you highlighted for sub-themes you hadn't thought of before.

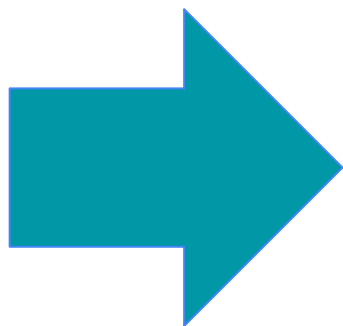| COLUMN 1 *Raw Data*	COLUMN 2 *Preliminary Codes*	COLUMN 3 *Final Code*
[1] The closer I get to retirement age, the faster I want it to happen. I'm not even 55 yet and I would give anything to retire now. But there's a mortgage to pay off and still a lot more to sock away in savings before I can even think of it. I keep playing the lottery, though, in hopes of winning those millions. No luck yet.	*"retirement age"* *financial obligations* *dreams of early retirement*	[1] RETIREMENT ANXIETY

Saldana, Johnny (2009). The Coding Manual for Qualitative Researchers. London: SAGE. p. 17.

Change the angle you're looking at!

Clustering codes using memos:

14 November 2007
EMERGENT PATTERNS, CATEGORIES, THEMES, AND CONCEPTS: MIDDLE SCHOOL HELL

MIDDLE SCHOOL HELL seems to be an umbrella code for TWEEN ANGST and THE LOST BOY. Barry says in another interview that those years were a "DEAD PERIOD" for him – an evocative In Vivo Code. But I think I focus on the particular by using MIDDLE SCHOOL HELL to represent many students within that age range.

However, don't discount "DEAD PERIOD" as a major theme or concept at this time. As I continue analysis, that code may be more conceptual than I think at this point in the study. Something in me doesn't want to let it go. A "DEAD PERIOD" can occur during any portion of a person's life, while TWEEN ANGST is limited to a specific age range.

from Saldana, J. (2009). The Coding Manual for Qualitative Researchers, p. 36.

Clustering quotes into overarching themes

Yik Yak   37 UPVOTES

When the guy next to you is whispering his essay aloud in the library, and it sounds like Parseltongue 🐍🐍🐍

www.inthelibrarywiththeleadpipe.org/2017/yik-yak

100 yik yaks from across the country → 4 main themes

1. Peer dynamics in the library

2. Expectation for quiet

3. Frustrations of studying

4. Seeking peer help with study aids

→ leads to recommendations for librarians; see bit.ly/libyikyak

Try it!

Go back and read your text from a different angle. Add new codes.

Exporting and packaging data

Export: highlighted interviews

Documents

Add a document

Edit

ert   Edit

ȷoz   Edit

Interview #1, with [Bulbul]

Real name: [redacted]

Pseudonym: Bulbul

Date: 01/10/2019

Location: Google Hangouts

Context: [Bulbul] is in her mid-twenties and graduated with a BA in librarianship. She spent time abroad in an exchange program and is now looking for further education abroad. She'd like to do her MLS in another country, and is also under strong family pressure to get married.

Q: What profession were you most interested in as a child?

Export this view ▾

HTML

DOCX

PDF

Exported view: tagged interview

Context: [Bulbul] is in her mid-twenties and graduated with a BA in librarianship. She spent time abroad in an exchange program and is now looking for further education abroad. ==She'd like to do her MLS in another country, and is also under strong family pressure to get married.== *[family-concerns]*

==Q: What profession were you most interested in as a child?==

==A: From my childhood, I really liked to visit the school library, and I got this dream to be a librarian. When I finished my school, I decided to apply and study for library science. I'm not close to math and physics, but more to humanities. That's why I chose library science.== *[childhood-dreams]*

Export: your codebook of tags

Exported view: codebook

	A	B	C
1	tag	description	number of highlights
2	interesting	Further review required [this is an automated Taguette code]	0
3	family-advice	Advice that the respondent was given about their career.	0
4	childhood-dreams	Respondent mentions a childhood dream to become a library	3
5	family-considerations	Sections that address the impact of family members on librarian career choices.	0
6	family-concerns	Ways that concern for family members (or from family members? This is unclear) impact librarian career choices.	1
7	career-exploration	Librarian describes the process of exploring a new career.	1

Export: all highlights, with their tags

Exported view: all tags

|	A	B	C	D
	id	document	tag	content
	1161030	Int1 Bulbul	family-concerns	She'd like to do her MLS in another cou
	1161034	Int1 Bulbul	childhood-dreams	Q: What profession were you most inte
	1161043	Int1 Bulbul	career-exploration	Prompt: How do you know that for sur
	1161059	Int2 Meruert	childhood-dreams	Q: What profession were you most inte
	1161060	Int3 Botagoz	childhood-dreams	Q: What profession did you want as a c

Why share data for re-use?
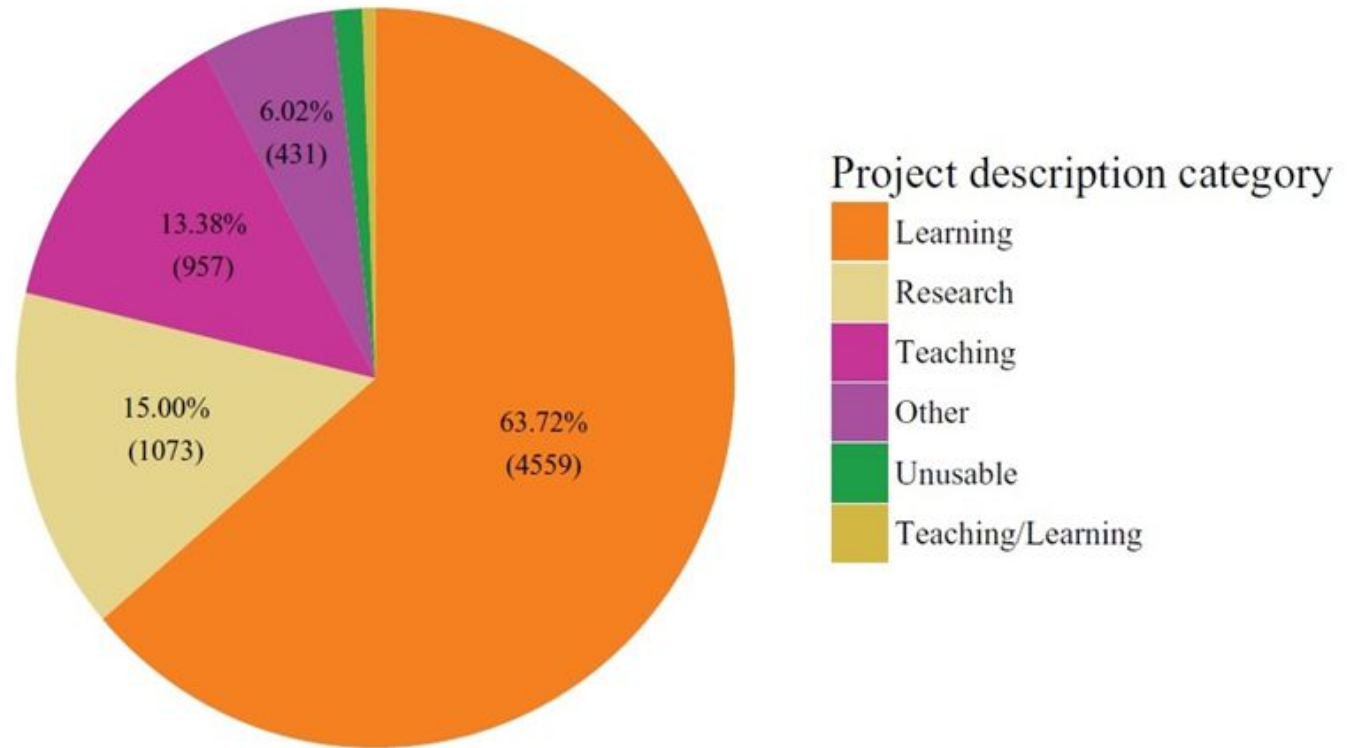
Figure 2. Reuse purposes of qualitative data downloaded from UK Data Service during the period 2002-2016.
Image from Bishop and Kuula-Luumi (2017) Revisiting Qualitative Data Reuse: A Decade On, p. 5

Why use a qualitative data repository?

Expertise in handling textual and image data that may identify specific people

Someone else hosts your data and grants others permission to use it

Allows people to search for and cite your data

Provides a permanent ID and location for your data

Options

ICPSR

Based out of the University of Michigan (USA)

Accepts text, image, audio, and video data; contact before depositing

They recommend anonymization + secondary IRB approval/data use agreements

You can write the costs of data curation into a grant (contact for estimates); campus membership may also offset some costs

Options

Based out of the University of Syracuse (USA)

Focused on issues around storing, securing, and coding qualitative data

Approach before archiving; your campus may have a membership

Packaging qualitative data for reuse

What do you need in order for others to understand your qualitative data?

- Research methods and practices (including the informed consent process) that are fully documented
- Data collection instruments such as interview questionnaires
- Steps taken to remove direct identifiers in the data (e.g., name, address, etc.)
- [Qualitative Data Codebook](#)

Further resources

[Getting started with Taguette](#)

[Taguette Frequently Asked Questions](#)

[Qualitative Data Repository](#)

[Tidy Data for Librarians](#)

[Tutorials on Ethnographic Data Management](#)

Questions?

Anna: asackmann@berkeley.edu

Celia: emmelhainz@berkeley.edu