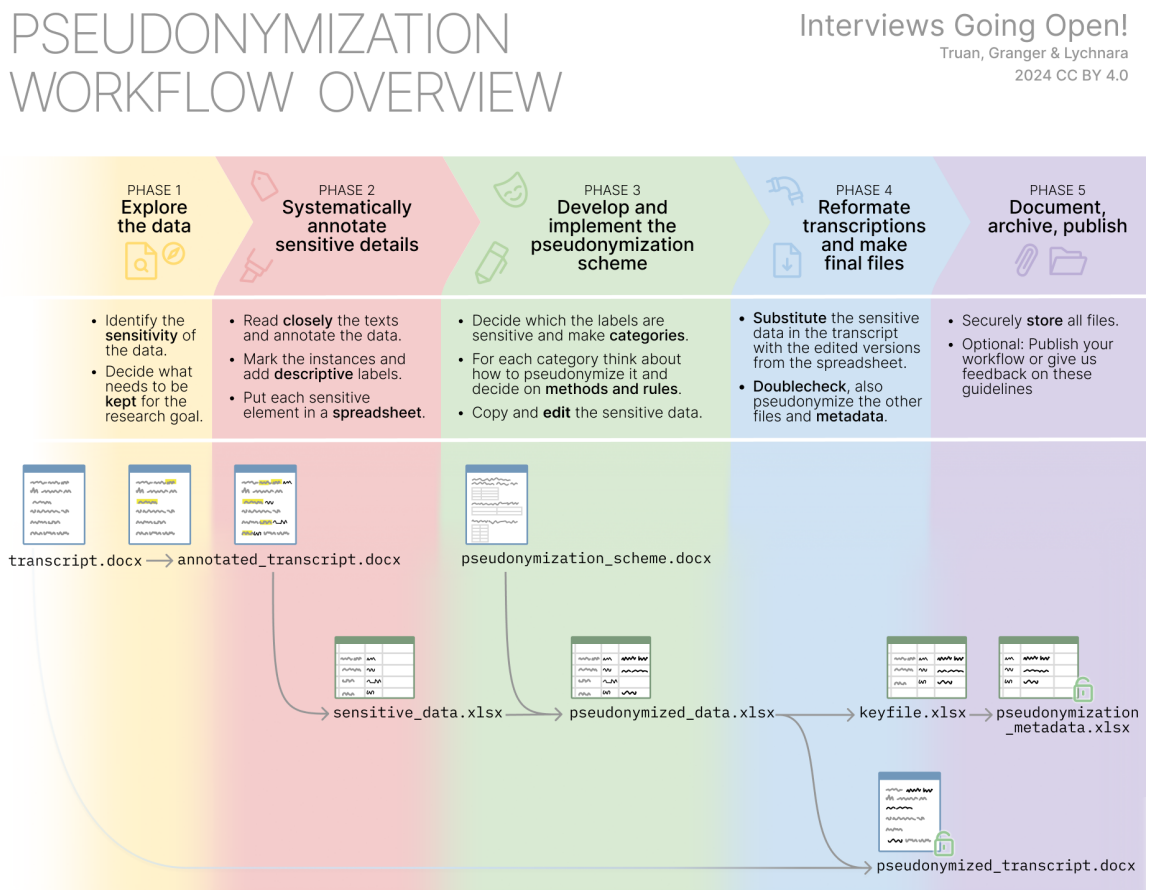
Interviews Going Open! How to Pseudonymize Sensitive Interview Data: A Detailed Step-by-Step Guide (With Time Stamps)

Naomi Truan, Sophie Granger, Jo Lychnara

How to Pseudonymize Sensitive Interview Data: A Detailed Step-by-Step Guide (With Time Stamps)

Naomi Truan, Sophie Granger, Jo Lychnara



Overview of the pseudonymization workflow 'Interviews Going Open!'
Figure by Sophie Granger

In this resource, you can follow a step-by-step description of a research data workflow involving the pseudonymization of interviews conducted in a small community in which individuals may be easily identifiable.

Read further if you are interested in open data, FAIR principles, process documentation, or are about to submit a grant and need to identify how many hours pseudonymizing potentially sensitive data may take.

A separate document (Truan, Granger & Lychnara 2024) details the pseudonymization scheme used to pseudonymize the interview data in a tight-knit community in a way that ensures privacy while maintaining data utility for research or analysis.

Cite as: Truan, Granger & Lychnara 2024a, Interviews Going Open! How to Pseudonymize Sensitive Interview Data: A Detailed Step-by-Step Guide (With Time Stamps). ⟨halshs-04743263⟩

Last update: 17.12.2024

Project Overview

Pseudonymizing Interviews in the Humanities & Social Sciences

Title: Interviews Going Open! How to Pseudonymize Sensitive Interview Data: A Detailed Step-by-Step Guide (With Time Stamps)

Keywords: qualitative research, interviews, sensitive information, pseudonymization, anonymization, guidelines, funding, time management, grants

Key values: open science, open data, FAIR principles

Project (origin, grants): This document has been produced as part of the Leiden University Centre for Digital Humanities Small Grant 2024 for the project "Interviews Going Open! Developing Interdisciplinary Guidelines on How to Publish Qualitative Interview Data as Open Data".

Authors: Project leader: Dr Naomi Truan; Research Assistants: Sophie Granger and Jo Lychnara. See below for Author contribution CRediT statement.

Abstract: The project *Interviews Going Open!* addresses the challenge of sharing sensitive interview data within the humanities and social sciences. Interviews are valuable sources of information, but concerns about privacy and ethical considerations often hinder their open accessibility. Another common challenge is how to make "thick description" (a property of qualitative fieldwork) suitable for metadata (which is often criticized as oversimplification in qualitative research). This handout documents the process, including the number of hours needed for each task and our questions as they arise. In doing so, we identify challenges faced by qualitative researchers and offer lowkey, realistic, practical, and tailored solutions based on a corpus of 25 interviews in a small and possibly easily recognizable community.

Author contribution CRediT statement: (see Allen, O'Connell & Kiermer 2019)
* denotes equal contribution
Conceptualization: Naomi Truan
Project Administration and Supervision: Naomi Truan
Writing - Original Draft Preparation: Naomi Truan, Sophie Granger, Jo Lychnara
Writing - Review & Editing: Naomi Truan, Sophie Granger, Jo Lychnara
Pseudonymizing: *Sophie Granger, *Jo Lychnara
Data management: *Naomi Truan, *Jo Lychnara
Data formatting: *Sophie Granger, *Jo Lychnara
Coding: Sophie Granger *Figures*: Sophie Granger
Acknowledgments: Andrew Hoffman & Femmy Admiraal advised us during the process.

Table of Contents

Pseudonymizing Interviews in the Humanities & Social Sciences

Research Data Management Workflow

Pseudonymizing Interviews in the Humanities & Social Sciences

Project's aims

> "Any sets of research ethics guidelines and dicta will be ineffective if researchers do not have embedded into their practice strong values establishing ethical behavior built on the principle of care" (Boellstorff et al. 2012: 129)

The project's aims are twofold. Firstly, it seeks to **pseudonymize existing interviews** to protect the interviewees' identities and personal data while preserving the metadata crucial for qualitative research.

The main question we address is: How can we navigate a **small-scale project on a tight budget** in just four months while ensuring **ethical practices**, even without the resources to double-check every decision with the interviewees whose data is being pseudonymized?

Yet the objective goes beyond this specific set of interviews. By using this corpus as a sample, the project aims to **showcase a methodology for pseudonymizing personal data and create a practical guide**, including how much time each step takes. This is particularly important because, despite the growing body of research on pseudonymization and de-identifying sensitive data (Campbell et al. 2023), there is a lack of documentation detailing the technical necessities and potential issues that may arise. The current lack of practical information makes it challenging for newcomers or people handling new datasets to **estimate how much time and budget** (for example for research assistants) they should allocate to handle a corpus of e.g. semi-structured interviews.

This Research Data Management Workflow shows that it is possible to fully pseudonymize rich and sensitive interview data, including process documentation, in around **130 hours**.

How much time should I plan to pseudonymize an already transcribed corpus of 25 interviews?

130 hours, divided into 2 student assistants, is a good start
(budget of 3,500€ in the Netherlands)

The interview data

The data used to experiment with this workflow originates from a research project on the language ideologies of French speakers who also use German in Berlin (Truan 2024; Truan under review). The corpus consists of 25 semi-structured interviews in French (mean length: 50') which were audio- or video-recorded. The corpus amounts to around 30 hours of recorded interviews, transcribed in 160,333 words or 477 pages (A4, Word).

The interview data in a nutshell:

- 25 semi-structured interviews in French
- mean length: 50'
- around 30 hours of recorded interviews
- 160,333 words or 477 pages

The data collection and transcription were handled before this project, while Naomi Truan was working at the University of Leipzig (grant: Leipzig Flexible Fund; student assistants: Jun An Chen, Matilde Mondo). The interviews were transcribed thoroughly (word level), but may still present minor inaccuracies (e.g. inaudible words).

Disclaimer: The interview data is provided in Word format rather than an interoperable format such as XML or TEI due to time constraints and a limited budget. While XML or TEI would offer enhanced flexibility and long-term preservation benefits, the resources required to implement these formats were not available during the project. Prioritizing the timely completion of the pseudonymization, we opted for a widely accessible and manageable format that meets the immediate needs of the project. However, meeting interoperability standards remains the goal for future data management.

The second phase, which is the object of this workflow, was made possible thanks to the LUCDH Small Grant 2024, with the pseudonymization workflow as a target.

For whom? How to use this document

This document is intended for researchers interested in making sensitive interview data FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) and possibly open, i.e. "data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike"[1].

[1] https://opendatahandbook.org/guide/en/what-is-open-data/, Open Data Handbook, accessed on 14.08.2024.

This Research Data Management Workflow should not be viewed as a best practice guide with mandatory solutions, but rather as a **practical example of implementing FAIR and Open Science principles** when collecting sensitive data.

To give more tangible information on how much time should be planned for data pseudonymization specifically, we include the number of hours allocated per task, hoping that this information will be helpful for scholars **applying for grants** and looking for **examples of project management and budget**.

This workflow paper was published along with a practical document that details the pseudonymization scheme for this project and tips for using files and data efficiently (Truan, Granger & Lychnara 2024). .

Pseudonymization – What do we mean?

Disclaimer: *The notes below may benefit from review by a GDPR/privacy expert. If that's you and you have anything to add or modify, please get in touch!*

Deidentification

- Deidentification is an **overarching term** used to describe processes that remove personal data, including both anonymization and pseudonymization.

Pseudonymization

- Pseudonymization is the process of **removing personal identifiers from data and replacing those identifiers with placeholder values**.
- The process reduces the identifiability of personal data without fully eliminating the link to the original identity.
- This aligns with the **GDPR's risk-based approach** rather than a binary standard, with pseudonymization being part of a broader set of "technical and organizational measures" outlined in Article 32[2].
- The pseudonymized data cannot be matched to an identifiable person, as the **personal information is stored separately in a key file**.
- In our project, since the key files are not uploaded to the data repository, the uploaded data remains pseudonymized under GDPR definitions, as the separation of the key files ensures that the data cannot be directly linked to individual identities.

Anonymization vs Pseudonymization

While pseudonymization and anonymization may have been conflated in the literature (Saunders, Kitzinger & Kitzinger 2015), anonymization and pseudonymization are two

[2] https://gdpr-info.eu/art-32-gdpr/, accessed on 17.12.2024.

distinct methods for handling personal data, each with different privacy implications (see DeLacey 2024):

- *Anonymization* involves removing all direct and indirect identifiers, such that the data can no longer be used to identify individuals, and no identification key exists. This process is challenging, especially with qualitative data like interviews, and once anonymized, the data no longer qualifies as personal data under the GDPR.
- In contrast, *pseudonymization* is less secure because, while direct identifiers (like names or social security numbers) are removed, the data can still potentially be linked back to individuals using indirect identifiers or additional data points. Pseudonymization operates on a spectrum, ranging from least to most identifiable, and the specific method used depends on the research project and data type. Unlike anonymization, pseudonymized data still qualifies as personal data under the GDPR.

Personal data

- According to the European Commission, personal data is "any information that relates to an identified or identifiable living individual"[3].
- This broad understanding of 'personal data' means that in handling interviews, we need to go beyond obvious personal data such as name and home address to encompass **other elements which, considered together, may lead to the identification of a specific individual** (think: work, family configuration, languages used, etc.).

Sensitive data

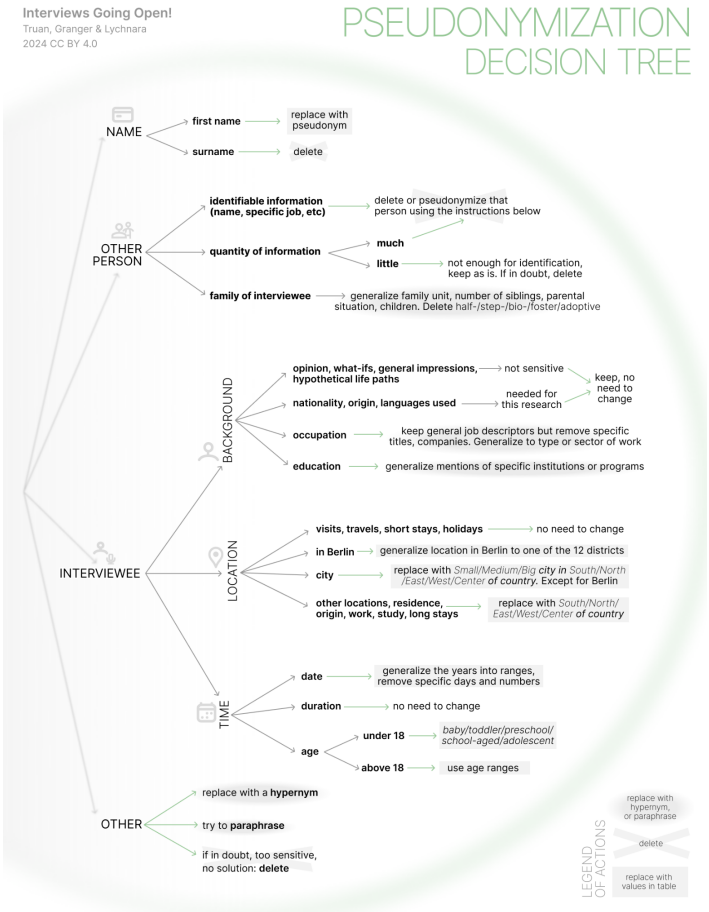
- Sensitive data, also known as special category data under the GDPR, includes personal data that reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for uniquely identifying a person, data concerning health, or data concerning a person's sex life or sexual orientation.
- This type of data is considered more **sensitive because its misuse could significantly impact individuals' privacy and freedoms**. Due to its sensitive nature, it is subject to stricter processing conditions and protections under data protection laws.

[3] https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en, European Commission, accessed on 14.08.2024.

Decision tree

Pseudonymizing Interviews in the Humanities & Social Sciences

- The decision tree reflects the pseudonymization scheme of our project. It gives an overview of the types of edits.
- The goal of a decision tree and scheme tables are to determine **the most suitable method** for masking identifiers while preserving data utility.
- The decision tree **systematically guides the pseudonymization process** based on several factors:
    - type of data;
    - level of identifiability;
    - purpose of the research;
    - required balance between data protection and usability.
- A decision tree can be designed to help to **decide which identifiers to remove or mask**. It could also address how to handle indirect identifiers to ensure compliance with privacy regulations, like the GDPR.



Decision tree: Overview of the pseudonymization scheme 'Interviews Going Open!'
Figure by Sophie Granger

Pseudonymization Scheme

Pseudonymizing Interviews in the Humanities & Social Sciences

- The tabular format for the pseudonymization scheme gives more room for details than the decision tree. Besides the instructions, it explains the **rationale behind the pseudonymization** and what is needed for the research purposes, allowing the team to make informed edits when dealing with sensitive information.
- The objective is to **protect the data while maintaining its usefulness for analysis** for our specific research purposes.

Category	Explanation	Pseudonymization Instructions	Examples
Names	Refers to names mentioned in the transcript whether they are of the interviewees or other individuals	Keep an alternative name and stay consistent throughout the document. Recurrent names should be replaced with pseudonyms.	Transcribed: John and Sarah met in Berlin. John moved there in the Fall, and Sarah followed a year later.
			Pseudonymized: Denis and Emma met in Berlin. Denis moved there in the Fall, and Emma followed a year later.
Place of origin	Helps define a person's origin and identity prior to migration	Use generic regions (North/South/East/West/Central of Country X) instead of specific places of origin unless otherwise stated.	Transcribed: I grew up in Mainz.
			Pseudonymized: I grew up in a small town in West Germany.
Nationality	Ties to the cultural aspects of identity	Keep all instances where nationality is mentioned, as this information is relevant to analyze the data and key to the project on migration and multilingualism.	No change incorporated: She is German.
Languages used	Reflects cultural and linguistic background	Keep all instances of languages used since it is a key aspect of the (in our case, sociolinguistic) analysis.	No change incorporated: She speaks German, English, and French fluently.
Dates	Significant for understanding the timeline of life events	Keep but generalize the years into ranges using the "Early decade" and "Late decade" instruction for sensitive dates.	Transcribed: In 85' I went to Berlin.
			Pseudonymized: In the early 80s I went to Berlin.
Education and studies	Key for professional and intellectual background	No need to pseudonymize unless specific institutions are mentioned.	Transcribed: She studied linguistics at Leiden University.
			Pseudonymized: She studied linguistics at a university in the South of the Netherlands.
Occupation	Defines the person's current professional life	Keep job titles but remove specific companies or institutions if mentioned.	Transcribed: He worked as a software engineer at Google.
			Pseudonymized: He worked as a software engineer at a tech company.

Category	Explanation	Pseudonymization Instructions	Examples
Age	Provides context for life experience	Generalize age into categories according to language acquisition phases: 0-1 → baby, 1-2 → toddler, 3-6 → preschool 7-11 → school aged children 12-17 → adolescence For people 18+, use predefined age ranges such as: 18-24, 25-34, 35-44, etc.	Transcribed: She was 35 years old. Pseudonymized: She was 35-44 years old.
Place of residence/ other geographic locations	Indicates current or past location of the interviewees	Replace specific city or neighborhood names with population-based descriptions or geographic generalizations (North/South/East/West/Central of Country X).	Transcribed: She lives in Paris. Pseudonymized: She lives in a large city in the West of France.
Other biographical information	Details about a person's life that help describe their background, experiences, and identity	Pseudonymize sensitive information by generalizing personal life details.	Transcribed: She moved after her divorce. Pseudonymized: She moved after a change in her personal family situation.
Family	Can influence personal motivations and experiences	Generalize family relationships by using terms like 'siblings' or 'parent' instead of specifying 'half-sibling', 'step-sibling', 'bio-parent', etc.	Transcribed: My brother and sister. Pseudonymized: My siblings.
Direct mention of another person	Refers to mentions of individuals in the narrative who were not interviewed.	If the individual is not part of the interview, anonymize and/or delete mentions unless relevant.	Transcribed: My neighbor, John, never moved out. Pseudonymized: My neighbor, Jack, never moved out. OR: My neighbor never moved out.

[for more details also see Truan, Granger & Lychnara 2024, Interviews Going Open! Pseudonymization Strategies: Protecting Data While Preserving Utility]

Workflow

Pseudonymizing Interviews in the Humanities & Social Sciences

Manual pseudonymization — Why?

- Manual pseudonymization is crucial for **data integrity and confidentiality**, protecting the interviewees' identities.
- The process supports a **consistent, systematic approach**, especially with multiple researchers involved.
- Automatic pseudonymization may be faster but often requires additional checking, which can be time-consuming and less thorough.
- The main question we address is: How can we navigate a **small-scale project on a tight budget** in just four months while ensuring **ethical practices**, even without the resources to double-check every decision with the interviewees whose data is being pseudonymized?

Task phases and time allocation

Phase	Step	Time per step (h)	Time per phase (h)
Phase 1: Explore the data `Main working file: transcript_annotated.docx`	*Step 1*: Set up a working diary	0,1	
	Step 2: Familiarize yourself with the data	18	
	Step 3: Reflect	1	**19**
Phase 2: Systematically annotate sensitive details `Main working file: sensitive_data.xlsx`	*Step 1*: Categorize	21	
	Step 2: Decide on the strategy	7	
	Step 3: Reflect	1	**30**
Phase 3: Develop and implement the pseudonymization scheme `Main working file: pseudonymized_data.xlsx`	*Step 1*: Develop a consistent methodology	7	
	Step 2: Finalize the pseudonymization	23	**30**
Phase 4: Reformate transcriptions and make final files `Main working file: transcript_pseudonymized.docx`	*Step 1*: Implement and automate	8	
	Step 2: Check and finalize transcripts	2	
	Step 3: Pseudonymize metadata	6	**16**
Phase 5: Document, archive, publish `Main working file: whole folder`	*Step 1*: Document	1	
	Step 2: Ask for informed consent to make the (pseudonymized) data open	2	
	Step 3: Archive	1	
	Step 4: Publish	1	**5**
Administrative tasks and meetings	General administrative tasks and meetings	30	**30**
Total project hours			**130**

Going further? How you could expand this resource

Following an idea by Andrew Hoffman, Service Scientist for Research Data Management at the University of Leiden, future research could explore the development of a **programmable tool** such as a **spreadsheet** that allows researchers to input project parameters (e.g., number of transcript pages, expected outputs, team size) to automatically calculate essential metrics like FTEs (hour allocation in the Netherlands), budgets, and other resource requirements for project proposals. This tool could aim to address the lack of accessible research data management (RDM) resources available during the pre-award phase. Research could focus on how to design such a tool to accommodate the heterogeneity of projects while making it useful for researchers, data stewards, and project managers in planning and executing projects more effectively

Phase 1: Explore the data

***Step 1: Set up a working diary* (needed for future phases)**

`File: any text document, a spreadsheet may be handy to record hours`

- Create a detailed log to record daily activities, decisions, and progress related to the pseudonymization process.
- Include notes on challenges encountered, solutions implemented, and any procedure updates or changes.

Step 2: Familiarize yourself with the data

```
Files: - copy of the full transcribed interviews
       - transcript metadata
```

- In this phase, we worked on a copy of the full corpus of interviews in a single document—a Word document containing the transcripts of 25 interviews and the annotation scheme = 160,333 words or 477 pages.
- We quickly read through the transcripts and the metadata documents with information about the speakers (age, gender, occupation, languages used, etc.).
- On a small sample, we decided to **highlight sections containing identifiable information**.
- This version focused on **direct personal data** like names, ages, and locations.
- Each interview (around 15 pages / 4,700 words) took about 15-40 minutes to highlight direct identifiers.

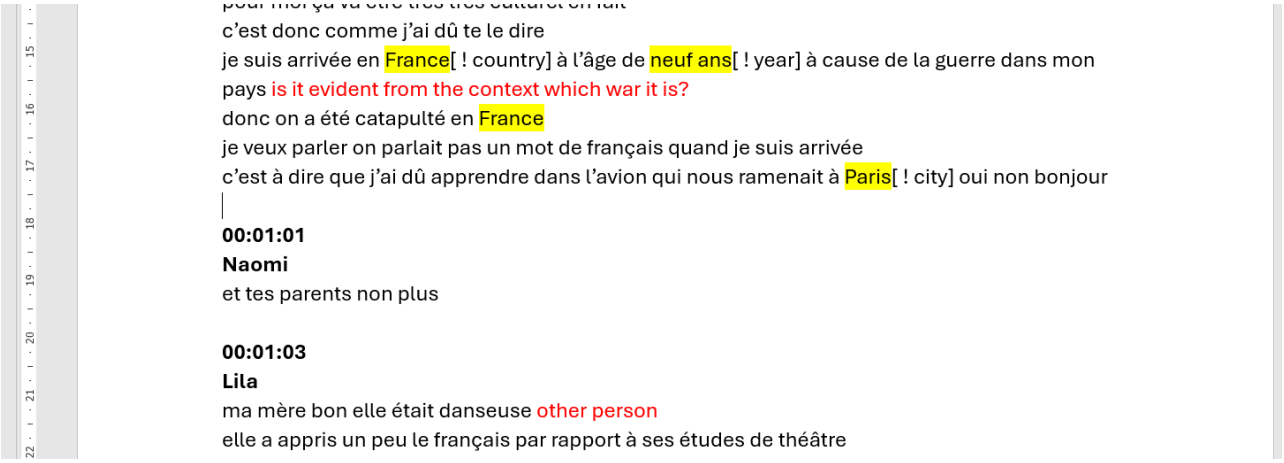


Figure 1: screenshot of the full transcript with highlighted sensitive information (here replaced with bogus data to protect the interviewee)

Time estimate:
Plan 1-3 minutes per 1 page of a fully transcribed interview to spot direct personal data.

Step 3: Reflect

Files: working diary (see step 1)

- The interviews contained many life stories, and the individuals are part of a small, tight-knit community, making them easily recognizable.
- Given the varying interpretations of what constitutes personal information. We focused on details that, when **combined with other parts of the interview**, could reveal the participants' identities.
- We decided that going forward we will also **pseudonymize additional details**, including **indirect information** and **biographical data**.
- We also discussed how to handle details about **individuals mentioned in the interviews, aside from the interviewee**.

Time estimate:
Phase 1 tasks and meetings: 29 hours

Phase 2: Systematically annotate sensitive details

Step 1: Categorize

`File: - annotated transcript`
`- sensitive data spreadsheet`

- After discussing in Phase 1 which types of information constitute sensitive information, we added category labels such as `[ ! location]`.

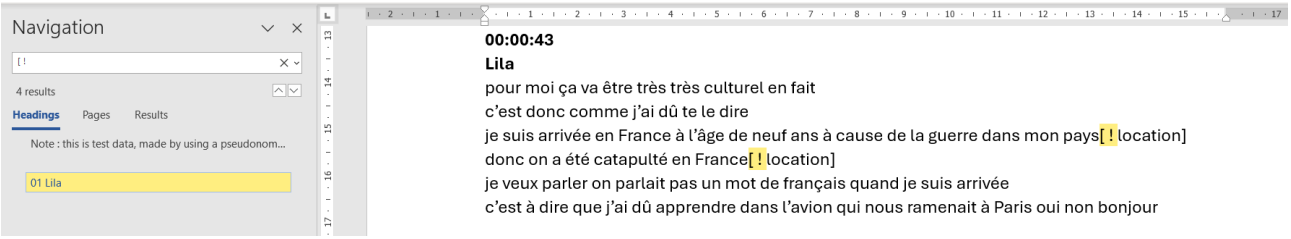


Figure 2: screenshot of the transcript with annotated sensitive information and type labels (here replaced with bogus data to protect the interviewee)

- After marking up the potentially sensitive data, we switched from working directly in the transcription to using a separate **spreadsheet.**
- Each row in the spreadsheet represents a sensitive data point with potentially identifiable information.
- We recorded the interview number, timestamp, and speaker's name to easily locate the utterance in the full transcript document.
- Categorizing the data made it easier to **locate and review ambiguous information without needing to revisit the full interview context**.
- We flagged words and expressions that often co-occur with sensitive information for potential future use.

	A	C	D	E	F	G	H	I
1	**Interview**	**Time stamp**	**Sensitive Info, original text**	**Type of sensitive Info**				
2	1	0:00:43	je suis arrivée en France à l'âge de neuf ans à cause de la guerre dans mon pay	année d'arrivée				
3	1	0:00:43	donc on a été catapulté en France	biographie				
4	1	0:00:43	ma mère bon elle était danceuse	famille				
5	1	0:00:43	elle a appris un peu le français par rapport à ses études de théâtre	famille				
6	1	0:00:43	donc on est arrivé en France quand moi j'avais neuf ans	année d'arrivée				
7	1	0:01:50	ce passage de ma langue maternelle qui était le serbo-croate	langue maternelle				
8	1	0:01:50	dehors on jouait dehors on parlait serbo-croate	langue maternelle				
9	1	0:01:50	heu le fait d'arriver à neuf ans à apprendre une autre langue	année d'arrivée				
10	1	0:01:50	mon frère a 5 ans de plus	famille				
11	1	0:03:34	c'est à dire que moi j'ai eu neuf ans pendant la guerre	biographie				
12	1	0:03:34	j'avais à peine huit ans quand j'ai arrêté d'apprendre ma langue	biographie				
13	1	0:03:34	donc en fait ils ont passé peut-être les deux premières années où j'étais à l'éco	biographie				
14	1	0:05:01	en France toutes mes études je les ai faites en France	études				
15	1	0:05:01	c'est à dire que je lisais pas j'allais pas à des cours spécial pour réviser le serbo-	langue maternelle				
16	1	0:05:01	et arrivée en Bosnie ce que je faisais quand j'allais en vacances	biographie				
17	1	0:05:01	donc j'le baragouine oui ça fait 20 ans	ans				
18	1	0:08:45	alors justement 1994 revenons 18 ans en arrière	année d'arrivée				
19	1	0:08:49	donc ça fait 18 ans que t'es ici non y a eu une interruption	année d'arrivée				
20	1	0:08:52	la première fois que j'ai découvert Berlin 1954	année d'arrivée				
21	1	0:08:52	je viens d'être acceptée dans la huitième meilleure prépa de France	études				
22	1	0:09:08	de m'envoyer en plus de se débarrasser de moi au Goethe Institut ici en cours	études				

Figure 3: screenshot of the spreadsheet with potential sensitive sentences (here modified to protect the interviewee)

Time estimate:
This meticulous process took 30 to 90 minutes per interview.

Step 2: Decide on the strategy to be implemented

Files: - updated working diary
- sensitive data spreadsheet

- The **three methods for handling various types of sensitive information** involved:
    - *replacing* (e.g. names and ages of interviewees)
    - *paraphrasing* (e.g. names of cities)
    - *deleting* (e.g. sensitive details about family members or personal information like medical history)

[also see Truan, Granger & Lychnara 2024, Interviews Going Open! Pseudonymization Strategies: Protecting Data While Preserving Utility]

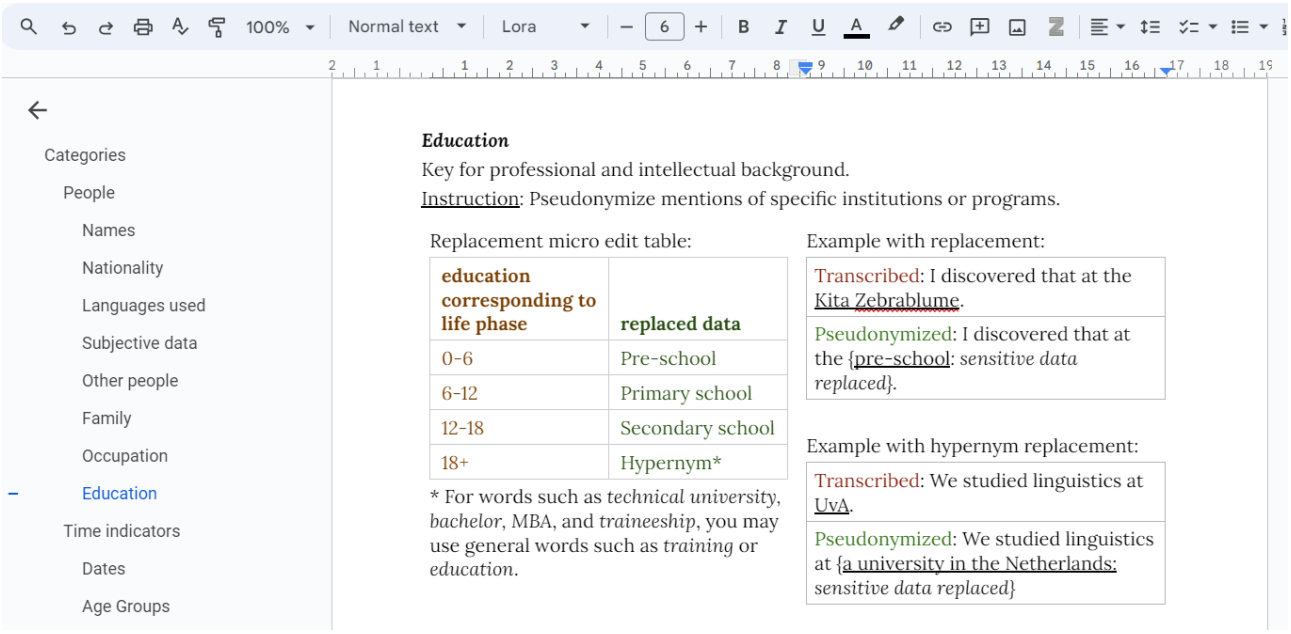


Figure 4: screenshot of the pseudonymization scheme

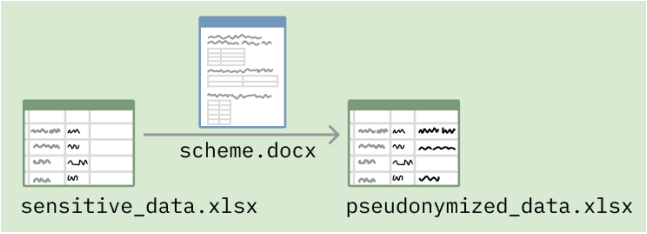
Step 3: Reflect

Files: - updated working diary
- sensitive data spreadsheet

- Some less common types of sensitive information were encountered. We discussed how these relate to our pseudonymization strategy.
- Specific place names and language code-switching required using search engines (e.g., Google) to assess their sensitivity.
- We updated the pseudonymization spreadsheet based on a commonly agreed scheme to ensure a cohesive course of action.

Time estimate:
Phase 2 tasks and meetings: 20 hours
Cumulative time into the project: 49 hours

Phase 3: Develop and implement the pseudonymization scheme



Step 1: Develop a consistent methodology

Files: - updated working diary
- pseudonymization scheme
- pseudonymization spreadsheet

- We identified frequent types of sensitive data and established a consistent approach to pseudonymization.
- We created a set of rules to ensure uniform application and conducted a collaborative exercise to practice pseudonymization, treating it as a practical quiz.
- In a new document we described thoroughly the scheme used for clarity and coherence.
- Tried out this pseudonymization methodology on a sample of rows.
- We added a column in our spreadsheet to label **the types of edits for pseudonymization**.

	A	B	C	D	E	F	G	H	I
1	Interview	Speaker	Time stamp	Type of sensitive i	Sensitive Info, original text	Pseudonymized	Type pseud.	Keywords	notes
2	1	Lila	0:00:43	année d'arrivée	je suis arrivée en France à l'âge de neuf ans à cause de la guerre dans mon pay	not changed	not changed		
3	1	Lila	0:00:43	biographie	donc on a été catapulté en France	not changed	not changed		
4	1	Lila	0:00:43	famille	ma mère bon elle était danceuse	ma mère bon elle était {dans le domaine artistique: sensitive d	replaced		
5	1	Lila	0:00:43	famille	elle a appris un peu le français par rapport à ses études de théâtre	elle a appris un peu le français par rapport à ses études {sensiti	deleted		
6	1	Lila	0:00:43	année d'arrivée	donc on est arrivé en France quand moi j'avais neuf ans	not changed	not changed		
7	1	Lila	0:01:50	langue maternelle	ce passage de ma langue maternelle qui était le serbo-croate	not changed	not changed		

Figure 5: screenshot of the pseudonymization spreadsheet with extra column for the types of edits

Examples of pseudonymized data

1. Place of origin and dates (here arrival in Berlin/German)

- *Transcribed version*: "The first time I discovered Berlin was when I moved with my parents from Mainz in 2002."
- *Pseudonymized version*: "The first time that I discovered Berlin was when I moved with my parents from {a small town in West Germany: sensitive data replaced} in {early 2000s: sensitive data reformulated}."

In this invented example, the details that could be used to identify the interviewee include their place of origin and the specifics of their move to Berlin. To preserve the context provided by this information, **we substitute it with a broader category instead of removing it entirely**.

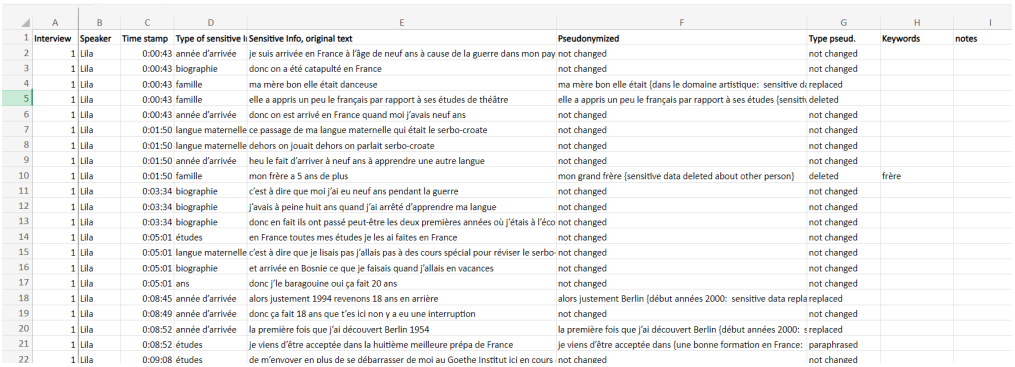
2. Names, occupation, and other biographical information

- *Transcribed version*: "My good friend Marion has been working at Meta since 2023. During this time she got married to Leon and they moved to Québec."
- *Pseudonymized version*: "My good friend {Mary: sensitive data replaced} has been working at {a tech company: sensitive data reformulated} since {early 2020s: sensitive data reformulated}. During this time she got married to {Alex: sensitive data replaced} and they moved to {large province of Canada: sensitive data reformulated}."

Step 2: Finalize pseudonymization

`Files: - pseudonymization scheme`
`- pseudonymization spreadsheet`

- A small batch of rows was pseudonymized which led to exceptions and special cases that needed to be addressed.
- Fill the column with the type of edits used for pseudonymization (replacing, paraphrasing, deleting).
- For instances where the information was deemed not too sensitive, the corresponding rows were intentionally left blank.
- We edited all the rows with sensitive information.
- We double-checked the edited rows, and proofread for spelling.
- The pseudonymization in the provided example mitigates the risk of identifying the interviewee by **altering or generalizing the specific details that could be used to recognize the individual in question**.

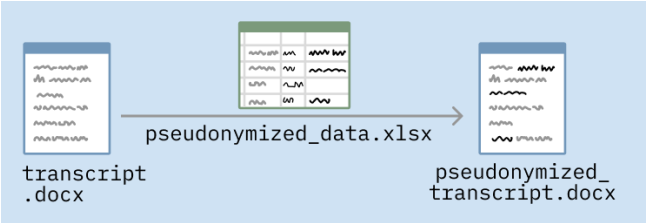


	A	B	C	D	E	F	G	H	I
1	Interview	Speaker	Time stamp	Type of sensitive I	Sensitive Info, original text	Pseudonymized	Type pseud.	Keywords	notes
2	1	Lila	0:00:43	année d'arrivée	je suis arrivée en France à l'âge de neuf ans à cause de la guerre dans mon pay	not changed	not changed		
3	1	Lila	0:00:43	biographie	donc on a été catapulté en France	not changed	not changed		
4	1	Lila	0:00:43	famille	ma mère bon elle était danceuse	ma mère bon elle était {dans le domaine artistique: sensitive d	replaced		
5	1	Lila	0:00:43	famille	elle a appris un peu le français par rapport à ses études de théâtre	elle a appris un peu le français par rapport à ses études {sensiti	deleted		
6	1	Lila	0:00:43	année d'arrivée	donc on est arrivé en France quand moi j'avais neuf ans	not changed	not changed		
7	1	Lila	0:01:50	langue maternelle	ce passage de ma langue maternelle qui était le serbo-croate	not changed	not changed		
8	1	Lila	0:01:50	langue maternelle	dehors on jouait dehors on parlait serbo-croate	not changed	not changed		
9	1	Lila	0:01:50	année d'arrivée	heu le fait d'arriver à neuf ans à apprendre une autre langue	not changed	not changed		
10	1	Lila	0:01:50	famille	mon frère a 5 ans de plus	mon grand frère {sensitive data deleted about other person}	deleted	frère	
11	1	Lila	0:03:34	biographie	c'est à dire que moi j'ai eu neuf ans pendant la guerre	not changed	not changed		
12	1	Lila	0:03:34	biographie	j'avais à peine huit ans quand j'ai arrêté d'apprendre ma langue	not changed	not changed		
13	1	Lila	0:03:34	biographie	donc en fait ils ont passé peut-être les deux premières années où j'étais à l'éco	not changed	not changed		
14	1	Lila	0:05:01	études	en France toutes mes études je les ai faites en France	not changed	not changed		
15	1	Lila	0:05:01	langue maternelle	c'est à dire que je lisais pas j'allais pas à des cours spécial pour réviser le serbo-	not changed	not changed		
16	1	Lila	0:05:01	biographie	et arrivée en Bosnie ce que je faisais quand j'allais en vacances	not changed	not changed		
17	1	Lila	0:05:01	ans	donc j'le baragouine oui ça fait 20 ans	not changed	not changed		
18	1	Lila	0:08:45	année d'arrivée	alors justement 1994 revenons 18 ans en arrière	alors justement Berlin {début années 2000: sensitive data repla	replaced		
19	1	Lila	0:08:49	année d'arrivée	donc ça fait 18 ans que t'es ici non y a eu une interruption	not changed	not changed		
20	1	Lila	0:08:52	année d'arrivée	la première fois que j'ai découvert Berlin 1954	la première fois que j'ai découvert Berlin {début années 2000: s	replaced		
21	1	Lila	0:08:52	études	je viens d'être acceptée dans la huitième meilleure prépa de France	je viens d'être acceptée dans {une bonne formation en France:	paraphrased		
22	1	Lila	0:09:08	études	de m'envoyer en plus de se débarrasser de moi au Goethe Institut ici en cours	not changed	not changed		

Figure 6: screenshot of the updated pseudonymization spreadsheet

Time estimate:
Phase 3 tasks and meetings: 30 hours
Cumulative time into the project: 79 hours

Phase 4: Reformate transcriptions and make final files



Step 1: Implement and automate

`Files:` - original transcript
- pseudonymized spreadsheet
- automation scripts

- We started by **manually implementing the pseudonymization rows** to the different interview documents.
- During this process, it became apparent that the volume of changes required was significant, leading to the decision to **develop a Python script** for automation (https://github.com/iconolocode/pseudonymization-interviews-going-open).
- Address any issues that arise due to module limitations and character encoding incompatibilities.
- An error log was also created to document and manage any issues that arose during this process.
- This script successfully automated a large portion of the work, though manual review was still necessary for 75 sentences.

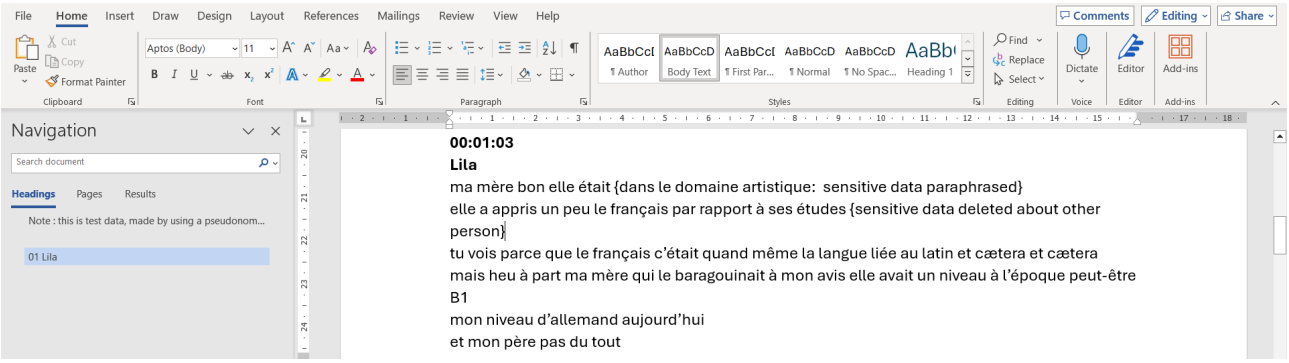


Figure 7: screenshot of the transcript with the pseudonymization edits

Step 2: Check and finalize transcripts

`File:` error log

- Manually apply the edits from the error log to the documents, ensuring accuracy.
- Perform necessary formatting, such as italicizing foreign words.
- Review and correct any remaining untranscribed sections by listening to the audio again and if needed pseudonymize them.

- Split the finalized main transcript into individual files for each interview.

Step 3: Pseudonymize metadata

Files: - pseudonymization spreadsheet
- pseudonymization keyfile
- pseudonymization metadata
- transcription metadata

- Ensure that the other files that describe the projects, such as the transcription metadata table, which includes information on each interview, is fully pseudonymized before public release.
- Make a copy of the pseudonymization spreadsheet and remove the instances that are not considered to reveal sensitive information. This finalized spreadsheet is a keyfile, it contains only the changes that were implemented to produce the pseudonymized interview transcripts.
- Make a copy of the keyfile and delete the column containing the original transcribed sentence, this can be published along with the metadata.

	A	B	C	D	E	F	G
1	Interview	Speaker	Time stamp	Type of sensitive I	Sensitive Info, original text	Pseudonymized	Type pseud.
2	1	Lila	0:00:43	famille	ma mère bon elle était danceuse	ma mère bon elle était {dans le domaine artistique: sensitive d	replaced
3	1	Lila	0:00:43	famille	elle a appris un peu le français par rapport à ses études de théâtre	elle a appris un peu le français par rapport à ses études {sensiti	deleted
4	1	Lila	0:01:50	famille	mon frère a 5 ans de plus	mon grand frère {sensitive data deleted about other person}	deleted
5	1	Lila	0:08:45	année d'arrivée	alors justement 1994 revenons 18 ans en arrière	alors justement Berlin {début années 2000: sensitive data repla	replaced
6	1	Lila	0:08:52	année d'arrivée	la première fois que j'ai découvert Berlin 1954	la première fois que j'ai découvert Berlin {début années 2000: s	replaced
7	1	Lila	0:08:52	études	je viens d'être acceptée dans la huitième meilleure prépa de France	je viens d'être acceptée dans {une bonne formation en France:	paraphrased
8	1	Lila	0:09:08	études	de m'envoyer en plus de se débarrasser de moi au Goethe Institut ici en cours	not changed	not changed

Figure 8: screenshot of the keyfile table

	A	B	C	D	E	F	G
1	Interview	Speaker	Time stamp	Type of sensitive I	Pseudonymized	Type pseud.	
2	1	Lila	0:00:43	famille	ma mère bon elle était {dans le domaine artistique: sensitive d	replaced	
3	1	Lila	0:00:43	famille	elle a appris un peu le français par rapport à ses études {sensiti	deleted	
4	1	Lila	0:01:50	famille	mon grand frère {sensitive data deleted about other person}	deleted	
5	1	Lila	0:08:45	année d'arrivée	alors justement Berlin {début années 2000: sensitive data repla	replaced	
6	1	Lila	0:08:52	année d'arrivée	la première fois que j'ai découvert Berlin {début années 2000: s	replaced	
7	1	Lila	0:08:52	études	je viens d'être acceptée dans {une bonne formation en France:	paraphrased	
8	1	Lila	0:09:08	études	not changed	not changed	

Figure 9: screenshot of the pseudonymized metadata table

Time estimate:
Phase 4 tasks and meetings: 16 hours
Cumulative time into the project: 95 hours

Phase 5: Document, archive, publish

Step 1: Document

`File: workflow document`

- Optional: Prepare a comprehensive documentation of the pseudonymization process, including methods, standards, and decisions.
- This step was facilitated by the diary we set up in phase 1, step 1.

Step 2: Ask for informed consent to make the (pseudonymized) data open

Disclaimer: Arguably, this phase usually takes place *before* data collection, i.e. before the interview starts. However, at the start of the project, making the data open was not taken into consideration. The informed consent gathered back then thus remained limited to data treatment and analysis by the researcher.

- Contact the participants to seek their consent for making the pseudonymized data publicly available.
- Provide clear information about how their data will be used, the measures taken to protect their identity, and any potential risks or benefits.

Step 3: Archive

`Files: folder with all the files`

- Securely store all files, including pseudonymized data, spreadsheets, and documentation, in an organized and protected archive.
  - *`Original Interview Transcript`* > should not be released
  - *`Working Diary`* > can/should be released (if no original examples)
  - *`Sensitive data spreadsheet`* > should not be released
  - *`Pseudonymization spreadsheet`* > should not be released
  - *`Pseudonymization scheme`* > can/should be released (no original examples)
  - *`Pseudonymization data keyfile`* > should not be released
  - *`Pseudonymization metadata`* > can/should be released
  - *`Transcription metadata`* > can/should be released (if pseudonymized)
  - *`Pseudonymized Interview Transcription`* > can/should be released
  - *`Automation scripts with a ReadMe`* > can/should be released
  - *`Error log`* > should not be released (contains original sentences)

Step 4: Publish

- Disseminate the finalized, pseudonymized data and findings according to relevant guidelines and permissions.

Time estimate:
Phase 5 tasks: 5 hours
Cumulative time into the project: 100 hours

During all phases:
Administrative tasks and meetings for the project: 30 hours

Ideas for platforms and open repositories specialized in interview data

- **Dataverse NL**: https://dataverse.nl/, online storage, sharing and publishing of research data
- **Open Science Framework**: https://osf.io/
- **CLARIN Virtual Language Observatory**: https://vlo.clarin.eu/, not a data repository in the narrow sense, but an interface to explore language resources and tools
- **Leiden Language Data**: https://leiland.lucdh.nl/, a searchable catalogue containing basic metadata information about linguistic data collected by researchers at Leiden University
- **The Language Archive:** https://archive.mpi.nl/tla/
- **Specific for French**: https://www.ortolang.fr/en/home/, a platform of linguistic tools and resources for an optimized treatment of the French language

References cited

Allen, Liz, Alison O'Connell & Veronique Kiermer. 2019. How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing* 32(1). 71–74.

Boellstorff, Tom, Bonnie Nardi, Celia Pearce & T. L. Taylor. 2012. *Ethnography and Virtual Worlds: A Handbook of Method*. Princeton University Press.

Campbell, Rebecca, McKenzie Javorka, Jasmine Engleton, Kathryn Fishwick, Katie Gregory & Rachael Goodman-Williams. 2023. Open-Science Guidance for Qualitative Research: An Empirically Validated Approach for De-Identifying Sensitive Narrative Data. *Advances in Methods and Practices in Psychological Science*. 6(4). 1–17. https://doi.org/10.1177/25152459231205832.

DeLacey, Hannah. 2024. Pseudonymizing Data. Presented at the Centre for Digital Scholarship: Summer Training Week, Leiden University. https://www.digitalscholarshipleiden.nl/articles/cds-summer-training-week-2024. (2 December, 2024).

Saunders, Benjamin, Jenny Kitzinger & Celia Kitzinger. 2015. Anonymising interview data: challenges and compromise in practice. *Qualitative Research* 15(5). 616–632.

Truan, Naomi. 2024. Whose language counts? Native speakerism and monolingual bias in language ideological research: Challenges and directions for further research. *European Journal of Applied Linguistics* 12(1). 34–53. https://doi.org/10.1515/eujal-2024-0006.

Truan, Naomi. forthcoming. Becoming a Speaker of German as an L1 French Speaker: Elite Multilingualism as a Means of Distinction in a Globalized World. *International Journal of Multilingualism*.

Truan, Naomi, Sophie Granger & Jo Lychnara. 2024. Interviews Going Open! Pseudonymization Strategies: Protecting Data While Preserving Utility.