

Making qualitative data reusable

Ricarda Braukmann & Maaïke Verburg

DCC Spring Training Days 2024

23 May 2024

Programme

13:00 - 13:30 Welcome & Introduction Round

13:30 - 14:00 Introduction to Making Qualitative Data Reusable

14:00 - 14:30 Planning a study with reuse in mind

14:30 - 14:45 Break

14:45 - 15:00 Open tools for processing and analysis

15:00 - 15:30 How to publish data for reuse

15:30 - 16:00 Q&A & Wrap up



Welcome & Introduction Round

DANS

Dutch national centre of expertise & repository for research data

Offers **data publishing and archiving** services for individual researchers and institutes

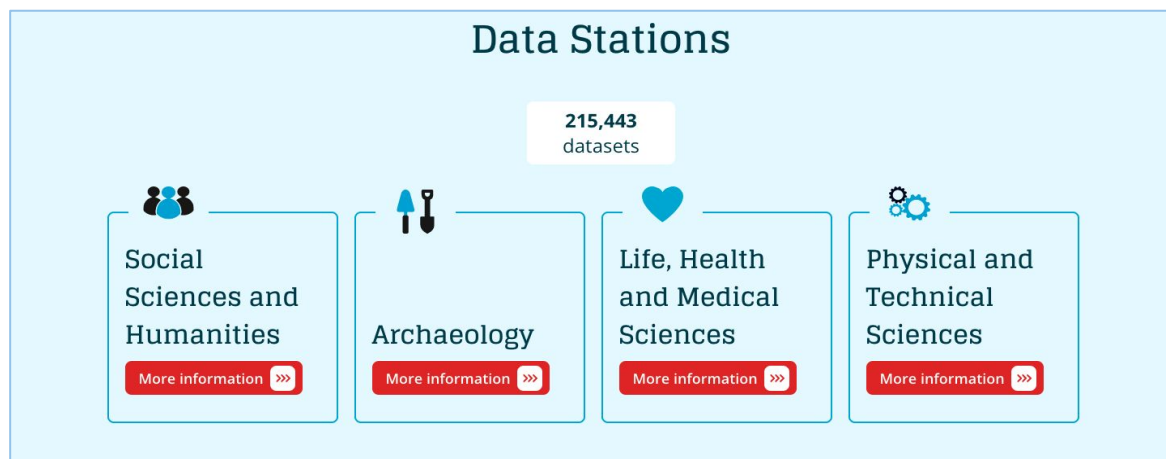


Offers **advice** and **training** on Research Data Management (**RDM**), **FAIR data** and **Open Science**



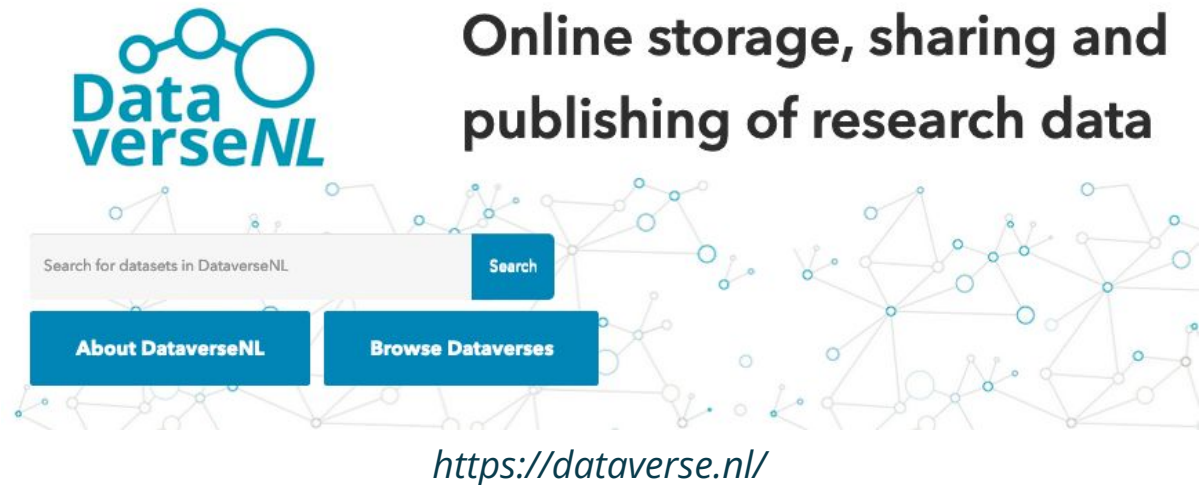
DANS Data Station

- Domain-specific data publishing and archiving service for researchers
- Making use of Dataverse technology
- DANS manages the technical infrastructure
- DANS manages the curation of the data
- Long term storage in our Vault
- CoreTrustSeal certified



DataverseNL

- Repository service for institutions (archiving & publishing)
- Making use of Dataverse technology
- DANS manages the technical infrastructure
- Institutes manage the curation of the data
- Long term storage in our Vault
- CTS certification can be done by institutes individually



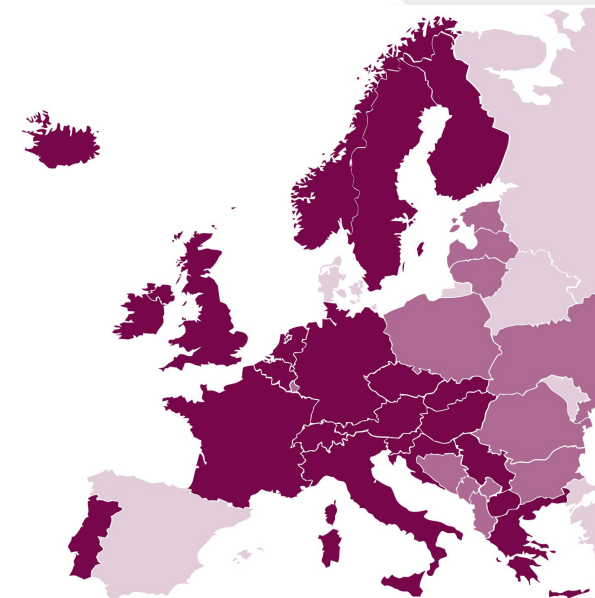
Consortium of **E**uropean **S**ocial **S**cience **D**ata **A**rchives

CESSDA provides large-scale, integrated and sustainable data services to the social sciences community, including

- A Data Discovery Portal
- Data Management Training
- Controlled Vocabulary Services



DANS is the Service Provider for the Netherlands



<https://www.cessda.eu/>



Making Qualitative Data Reusable

As open as possible as closed as necessary

Following the Open Science principles you want to publish **Open Access** where possible

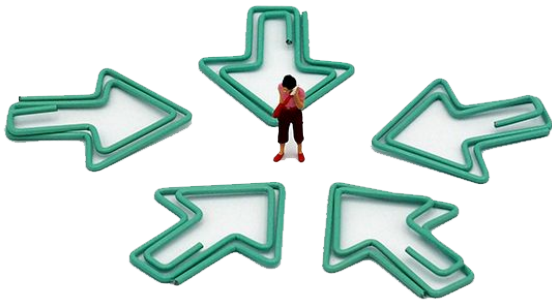


As open as possible as closed as necessary

Following the Open Science principles you want to publish **Open Access** where possible

Why would that not be possible?

- Personal data
- Sensitive information



When working with personal data

Common recommendations

- Gather **informed consent** (to share data) beforehand
- **Minimize** the amount of personal data you collect
- **De-identify** your data
 - Remove personal information
 - Anonymize or pseudonymise data



The challenges of qualitative data



Qualitative data refers to information that is not gathered in numerical form, but rather describes qualities or characteristics.

The challenges of qualitative data

Data which is hard to anonymise or pseudonymise!

- You lose contextual information
- Qualitative aspects are what you want to study

A lot of existing guidance applies to *quantitative* data

→ More guidance needed!



Qualitative data refers to information that is not gathered in numerical form, but rather describes qualities or characteristics.

Archiving and publishing of qualitative data

The screenshot displays the DANS Data Station Social Sciences and Humanities website. The browser address bar shows the URL: `ssh.datastations.nl/dataverse/root;jsessionid=7bcf54ec0dc00424d6b95d83a9ef?q=&fq1=d...`. The website header includes the DANS logo and navigation links: About, User Guide, Support, and Log In.

The main content area features the title "DANS Data Station Social Sciences and Humanities" and a description: "This Data Station allows you to deposit and search for data within the field of SSH." Below this, there is a "Metrics" section showing "149,059 Downloads" and buttons for "Contact" and "Share".

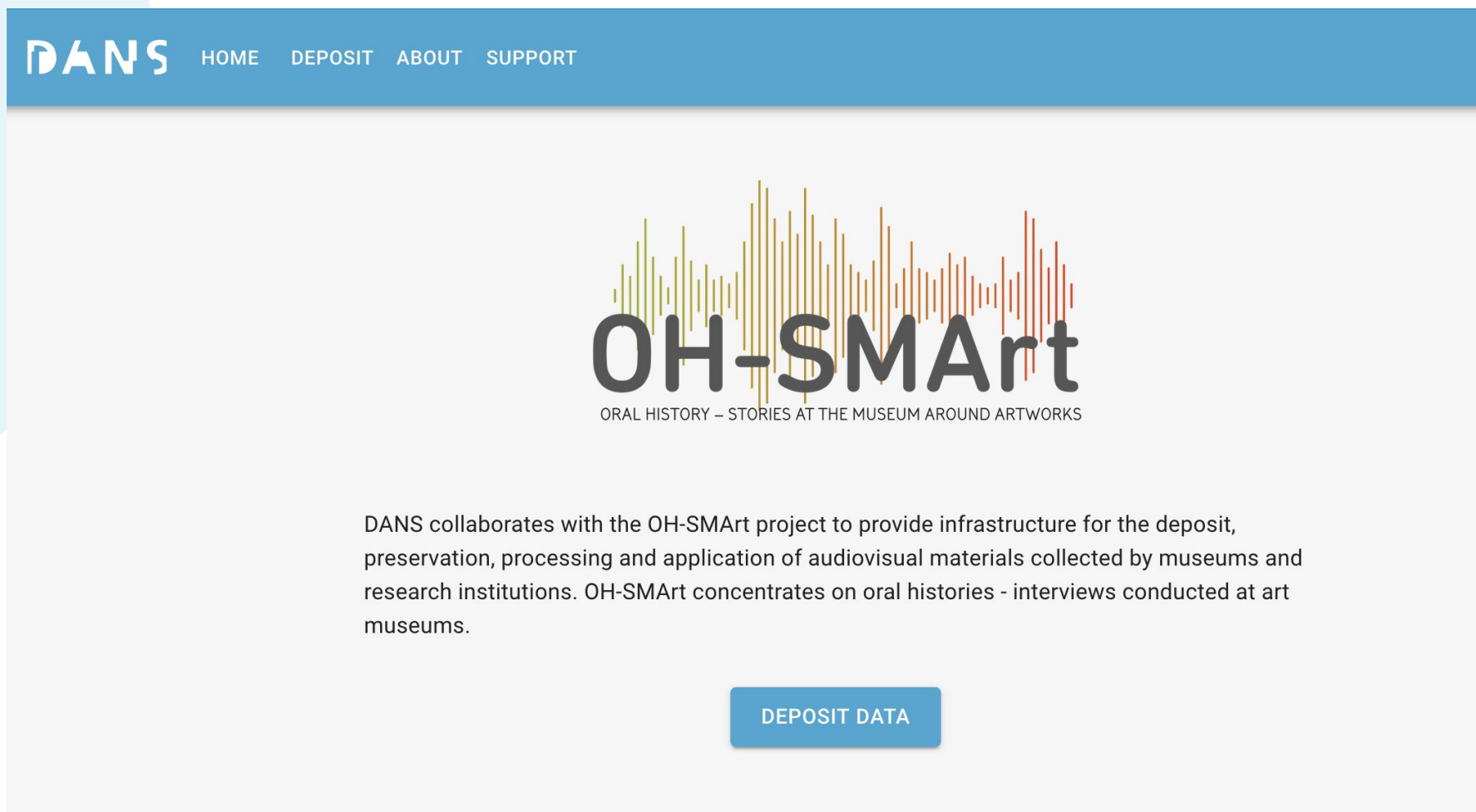
A search bar is present with the placeholder text "Search this dataverse..." and a search icon. To the right of the search bar is a link to "Advanced Search" and a button labeled "+ Add Data".

The search results are displayed in a list format. On the left side, there are filters for "Dataverses (0)", "Datasets (3,092)", and "Files (0)". Below these filters, there is a "Publication Year" section with a list of years and their corresponding counts: 2024 (39), 2022 (3), 2021 (1), 2020 (11), and 2019 (16). A "More..." link is available below the year list. At the bottom, there is a "Subject" filter.

The search results are titled "Collection: Oral History" and show "1 to 10 of 3,092 Results". The first result is "Friends in a Cold Climate: Neath Port Talbot-2", dated Mar 27, 2024. It includes a thumbnail image of a book cover and a description: "de Jager MA, E. J., 2024, 'Friends in a Cold Climate: Neath Port Talbot-2', [https://doi.org/10.17026/SS/WSNPUK](\"https://doi.org/10.17026/SS/WSNPUK\"), DANS Data Station Social Sciences and Humanities, V1". The description continues: "Former councillor Derek Vaughan grew up in Aberfan, a small village in South Wales known for a tragic event in 1966 when a coal tip collapsed onto a school, causing many deaths. Despite this tragedy, the speaker remembers the village as having a strong sense of community. However...".

The second result is "Project Ongekend Bijzonder, Amsterdam, interview 51", dated Mar 14, 2024.

Support tailored to qualitative data (e.g. interviews)



The screenshot shows the OH-SMART website. At the top is a blue navigation bar with the DANS logo and links for HOME, DEPOSIT, ABOUT, and SUPPORT. The main content area has a light gray background. In the center is the OH-SMART logo, which features a stylized waveform above the text 'OH-SMART' and the tagline 'ORAL HISTORY – STORIES AT THE MUSEUM AROUND ARTWORKS' below it. Below the logo, a paragraph of text describes the collaboration between DANS and the OH-SMART project. At the bottom center is a blue button with the text 'DEPOSIT DATA'.

DANS HOME DEPOSIT ABOUT SUPPORT

OH-SMART

ORAL HISTORY – STORIES AT THE MUSEUM AROUND ARTWORKS

DANS collaborates with the OH-SMART project to provide infrastructure for the deposit, preservation, processing and application of audiovisual materials collected by museums and research institutions. OH-SMART concentrates on oral histories - interviews conducted at art museums.

DEPOSIT DATA

ohsmart.datastations.nl

Making qualitative data reusable - Report

<https://doi.org/10.5281/zenodo.8160880>



Table of contents

About this guidebook	4
Definition of qualitative data	4
The importance of making qualitative data reusable	5
The challenges in making qualitative data reusable	5
Goals of this guidebook	6
Planning your project	7
Preparing informed consent	8
Organising and documenting qualitative data	9
Processing your data	10
De-identification of personal data	11
Archiving and Publishing qualitative data	12
Documentation and metadata	12
File formats	13
Decision tree for data reuse	14
Archive open access	16
Archiving with restricted access	16
Secure environment	17
The CaRe & DaRe solution: Decentralised reanalysis	17
Publish metadata only	18
Final thoughts	20
Additional resources	21
References	22

Planning Your Project

- Create a Data Management Plan (DMP) with reusability in mind
- Prepare your Informed Consents form well to include data sharing
- Consider best practices on file and folder naming to manage your data

Processing Your Data

- Consider if you can use open source software to process your data
- Work with open file formats wherever possible
- Consider the best techniques to de-identify your dataset

Archiving and Publishing Your Data

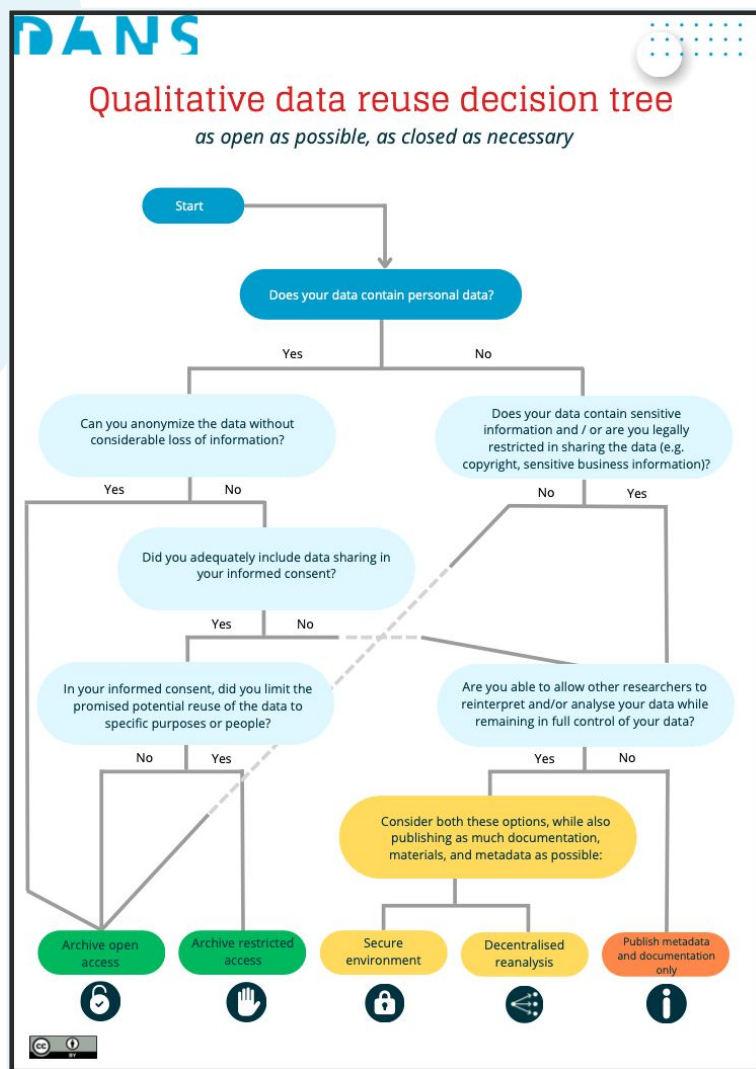
- Select a Trustworthy Digital Repository (TDR)
- Follow the TDR's advice on data documentation and metadata
- Use open file formats wherever possible

→ **Decide** whether your data can be published **open** or under **restricted access**



Making qualitative data reusable - Decision Tree

<https://doi.org/10.5281/zenodo.8160890>





Time For Questions





Planning a study with reuse in mind

Qualitative dataset

- To apply what you learn directly to data
- Use a dataset you brought yourself, or:
 1. AM Damen, 2022, "Palliatief Landelijk Onderzoek Eerstelijns Geestelijke verzorging (PLOEG) deelproject 3: 'Integratie GV eerste lijn vanuit 3 multidisciplinaire praktijken'",
<https://doi.org/10.17026/dans-xat-kj3c>, DANS Data Station Life Sciences, V2
 2. Stigter, Sanneke; Hatem, Amr; Mavroudis, Orestis, 2024, "Interview: My Bundgaard on Robert Rauschenberg's Mud Muse",
<https://doi.org/10.17026/SS/8RLLMN>, DANS Data Station Social Sciences and Humanities, V1



Informed consent

- Not only for participation in the study, but also for the archiving and sharing of the data
 - Consider in advance what data sharing you want to facilitate!
 - Purposes or role of reuser
 - Which parts of the data for which use (granular consent)
 - Anonymisation
- Sensitive data → GDPR
 - Art. 7 > Conditions for Informed Consent
 - Ch. 3 > Rights of the data subject

You can obtain consent for sharing non-anonymised data, but it is still not always recommended to do so

Art. 7 GDPR Conditions for consent

1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.
2. If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.
3. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent.
4. When assessing whether consent is freely given, utmost account shall be taken of whether, *inter alia*, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.

Templates



Data Management Expert Guide

ambiguous, specific and by a clear affirmative action that signifies agreement to the processing of personal data.

Written informed consent process (including online surveys)

Description

File download

Any relevant advertising or recruiting material (poster, email text, social media advert)

> Template poster advert

Template information sheet

> Par
info
tem



Template written consent form

> Written consent form template

Template information sheet for online research

For online tasks only, where there is no face-to-face contact with human participants

> Template information sheet for online research

Examples and templates

Note that all examples below assume that they are preceded by sufficiently specified information.

Template in Qualtrics

Examples CESSDA

Example sentences



Utrecht University

⊖ Click to see examples of consent forms

⊕ UK Data Archive

⊕ MRC Cognition and Brain Sciences Unit - University of Cambridge

⊕ FORS (Swiss Centre o

TEMPLATE 1: Participant Information/Opening Statement

Key points to include	Suggested text
<ol style="list-style-type: none">Level (eg: Masters, PhD, research) purpose, potential outcomes and implications of the study.The role of TU Delft and any third parties including funding bodyWho participants are (eg: children, experts, students in a dependent role to the researcher)What exactly what they are being asked to doWhat if any Personal Data (Personally Identifiable Information and/or Personally Identifiable Research Data) will be collected, and how it will be used, published and	<p>You are being invited to participate in a research study titled [Name of your research]. This study is being done by [Name of Researcher(s)] from the TU Delft [include also any collaborating partners including internship provider and/or funding body].</p> <p>The purpose of this research study is [provide participants with a short statement about the research], and will take you approximately [xx] minutes to complete. The data will be used for [provide list of intended uses, including publication, application and teaching]. We will be asking you to [provide summary of what kinds of questions or tasks participants will be faced with].</p> <p>As with any online activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by [be clear on whether the survey is completely anonymous, and/or whether IP addresses or other Personal Data will be collected. If so describe how you will safely store data, how confidentiality will be secured and how it will be anonymised].</p> <p>[mention Open data specifically if applicable]</p> <p>Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions. [Include also clarification on whether data can be removed within a given timescale. This will not be possible where surveys are completely anonymous]</p> <p>[Provide contact details for corresponding and Responsible Researcher]</p> <p>[If participants are agreeing to this Opening Statement by clicking through to an (anonymous) online survey, this should also be clear in the Opening Statement.]</p>

Note: the TUD Human Research Ethics Committee should not be included as a contact and does not deal with participant complaints.

[University of Oxford](#) | [CESSDA DMEG](#) |
[Utrecht University \(2\)](#) | [TU Delft](#)

Deidentification of data

- To increase possibilities of data sharing and reuse
- **Anonymous** - data cannot be re-identified and is no longer considered personal/sensitive
- **Pseudonymised** - re-identification of the data remains possible (e.g., due to key that links person to data)
- Planning for data reuse
 - data minimisation
 - creating anonymisation pipeline
 - granular consent

Guides, tips & tricks: [FSD](#) | [UKDA](#) | [CESSDA](#) | [Erasmus University](#)

Mini-exercise

Consider your dataset

- What data does/did the study generate?
- What questions and/or information would you put on your consent form to allow maximum data reuse?
- Consider granular consent, the exact data that would be shared, the purposes, etc.

File formats

- Balance between **open** formats and proprietary **community standards**

Open	Proprietary community-standard
<ul style="list-style-type: none">+ long-term sustainability (migration)+ accessible to all without software installation- if converted from proprietary format, information could be lost	<ul style="list-style-type: none">+ generally used and understood in the community+ no risk of information loss due to conversion- risk of poor long-term sustainability (no migration, software updates, obsolescence)- not/less accessible to other potential future users

- Preferred file formats ([DANS](#))
- Depositing the same data in multiple formats

Mini-exercise

Consider your dataset

- Which file formats are used in the deposit? Are they preferred (for that repository)?
- Could you replace/add other file formats to improve reusability of the dataset?
- Consider discipline-specific standards, long term preservation, and reuse outside of the original domain



Time For Questions





Break





Open tools for processing and analysis



•

•

•

•

•

•

•

•

•

A handful of packages for speech recognition exist on PyPI. A few of them include:

A handful of packages for speech recognition exist on PyPI. A few of them include:

- A handful of packages for speech recognition exist on PyPI. A few of them include:

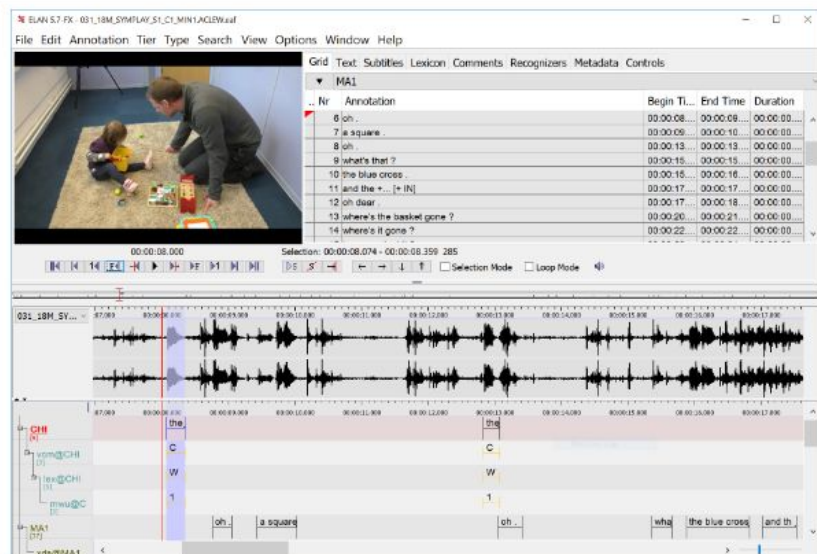
B
I
28
↺
➡
↻

Open tools for annotation

ELAN - free annotation tool for audio and video recordings developed by The Language Archive from the Max Planck Institute for Psycholinguistics. Open GPL3 license.

ELAN is an annotation tool for audio and video recordings.

Screenshot 1

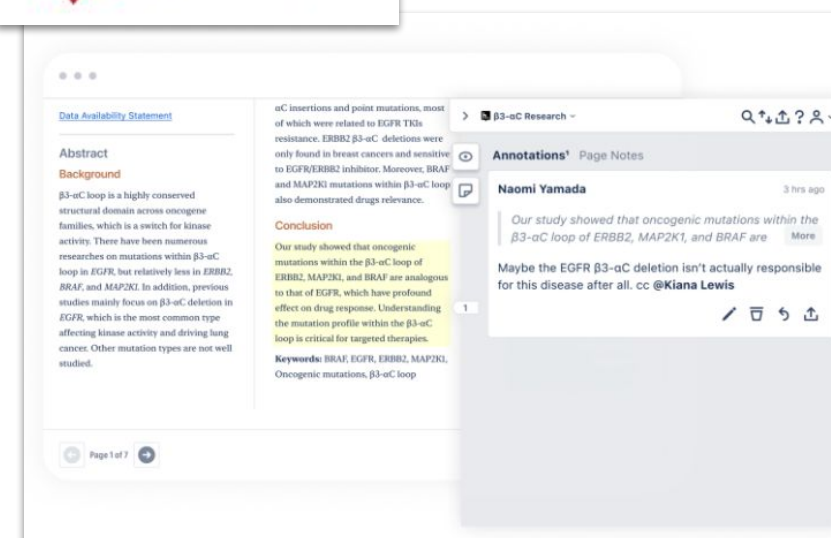


A sample from the [ACLEW project](#).

Taguette - Free and open-source tool to support manual processing of text. Work locally for data security.



Hypothes.is or **ATI** - for annotating literature / publications in pdf format. Can be collaborative.



Open tools for anonymisation/pseudonymisation

[anonymoUUs](#) - Python package, MIT license.
Runs through entire file tree to replace information with pseudo-IDs.

[UKDA anonymisation helper](#) - MS Word add-on. Does not change data, but highlights identifiable information to be considered manually.

[Textwash](#) and **FAMTAFOS** - Automatic identification and replacement of identifiable data. GPL3 license.



Utrecht
University

anonymoUUs

anonymoUUs is a Python package for replacing identifiable strings in multiple files and folders at once. It can be used to pseudonymise data files and therefore contributes to protecting personal data.



UK Data Service

Tools for anonymisation

Our [text anonymisation helper tool](#) can help you find disclosive information to remove or pseudonymise in qualitative data files. The tool does not anonymise or make changes to data, but uses MS Word macros to find and highlight numbers and words starting with capital letters in text. Numbers and capitalised words are often disclosive, e.g. as names, companies, birth dates, addresses, educational institutions and countries.

Textwash

Textwash is an automated text anonymisation tool written in Python. The tool can be used to anonymise unstructured text data. To achieve this, Textwash identifies and extracts personally-identifiable information (e.g., names, dates) from text and replaces the identified entities with a generic identifier (e.g., Jane Doe is replaced with PERSON_FIRSTNAME_1 PERSON_LASTNAME_1).



Let's put them to the test!

We created a mini experiment*

- We recorded a short audio fragment
- Script with various kinds of identifying information



A: Welcome **Maaïke**, nice that you are here.

B: Nice to see you too **Ricarda**, happy to be here.

A: So what we want to do today is test the anonymisation tool **Textwash** or **FAMTAFOS** that was developed by **Bennett Kleinberg** and colleagues.

B: Yes, we want to try ourselves if we can de-identify all the identifying information in this little audiosegment we are recording. It will serve as a test in our workshop "**Making Qualitative Data Reusable**" that we will hold during the **DCC Spring Training Days** on **May 23rd 2024**.

A: Right so to already throw in some identifying details - right now it is **Tuesday 23 April 2024 11:00**.

Step 1 From audio to text

- We searched in the CLARIAH Tools Registry

The screenshot displays the Clariah Tools Registry interface. At the top, there's a navigation bar with the Clariah logo and the word 'Tools'. Below this, a search bar and filters are visible. A 'Table of Contents' sidebar on the left lists various tools. The main content area features two tool cards. The first card is for 'Alpino Webservice 2.4', developed by Gerjan van Noord and Maarten van Gompel. It includes a description of the tool as a dependency parser for Dutch, its creators' affiliations, and links to the website and source code. The second card is for 'AlpinoGraph 1.0.5', developed by Peter Kleiweg. It describes the tool as a syntactic search engine for Dutch corpora, its affiliations, and links to the source code. Both cards show a 'technology readiness' bar and a 'repo status' indicator.

Clariah Tools

Toggle sidebar Search All tools Services only Table view SPARQL

Table of Contents Filters

- Alpino Webservice
- AlpinoGraph
- Automatic Speech Recognition Service
- Automatic Transcription of Dutch Speech Recordings
- FCS Aggregator
- Brieven als Buit search
- Corpus Hedendaags Nederlands
- OpenSoNaR
- CLARIAH Tools
- RU-Cesar
- e-WALD
- e-WBD
- e-WGD
- e-WLD
- FLAT: the FoLIA Linguistic Annotation Tool
- Piveling
- ForcedAlignment2
- Frog Webservice
- Grapheme to Phoneme converter
- Glem
- GrETEL 4
- I-Analyzer
- Golden Agents | lenticularis.org
- Lenticular Lens
- Lenticular Lens
- CLARIAH Media Suite
- Dataset Register OpenAPI
- Network of Terms GraphQL API
- Network of Terms Reconciliation API
- PaQu
- SASTA
- SHEBANQ
- SHEBANQ
- T-scan
- Ucto Webservice
- Service to tokenize, lemmatize, pos-tag and dependency parse using udpipe

Here you find all tools (i.e. software and software services) developed in the CLARIAH project, as well as some tools from predecessors and sister projects. Our tools are designed for researchers and developers in the Humanities and Social Sciences. Not all tools are suitable for all audiences and not all tools are mature and stable, this information should be clearly indicated for each tool, so you can make an informed judgement whether a tool might be suitable for you.

This list is automatically harvested from the tool producers and providers themselves, and updated daily.

Are you a CLARIAH developer and is your tool not included in the index yet or do you have questions or comments on the metadata? Please read our [contribution guidelines](#)

Alpino Webservice 2.4

Gerjan van Noord (backend), Maarten van Gompel (webservice)
Rijksuniversiteit Groningen (backend), Radboud Universiteit Nijmegen (webservice)
KNAW Humanities Cluster & CLST, Radboud University

Alpino is a dependency parser for Dutch, developed in the context of the PIONIER Project Algorithms for Linguistic Processing, developed by Gerjan van Noord at the University of Groningen. You can upload either tokenised or untokenised files (which will be automatically tokenised for you using ucto), the output will consist of a zip file containing XML files, one for each sentence in the input document. [\[view more\]](#)

Internet > WWW/HTTP > WSGI > Application Text Processing > Linguistic

dependency parsing folia linguistics nlp syntax

Created: 2015-09-08 Modified: 2023-11-01

AlpinoGraph 1.0.5

Peter Kleiweg
Computationele Taalkunde, Faculteit der Letteren, Rijksuniversiteit Groningen, Groningen University
Computationele Taalkunde, Faculteit der Letteren, Rijksuniversiteit Groningen, Groningen University

AlpinoGraph is een tool om syntactisch geannoteerde corpora te doorzoeken. De tool maakt gebruik van AgensGraph. AgensGraph combineert databasetechnologie (PostgreSQL) en Cypher, de standaard zoektaal voor grafen. De zoek-queries die je in AlpinoGraph kunt gebruiken zijn daarom een mix van SQL en Cypher. Daar voegt AlpinoGraph nog enkele extra uitbreidingen aan toe, zoals een eenvoudig maar handig systeem van macro's, en visualisatie van de resultaten. [\[view more\]](#)

Linguistics nwo:ComputationalLinguisticsandPhilology Software for humanities Structural Analysis

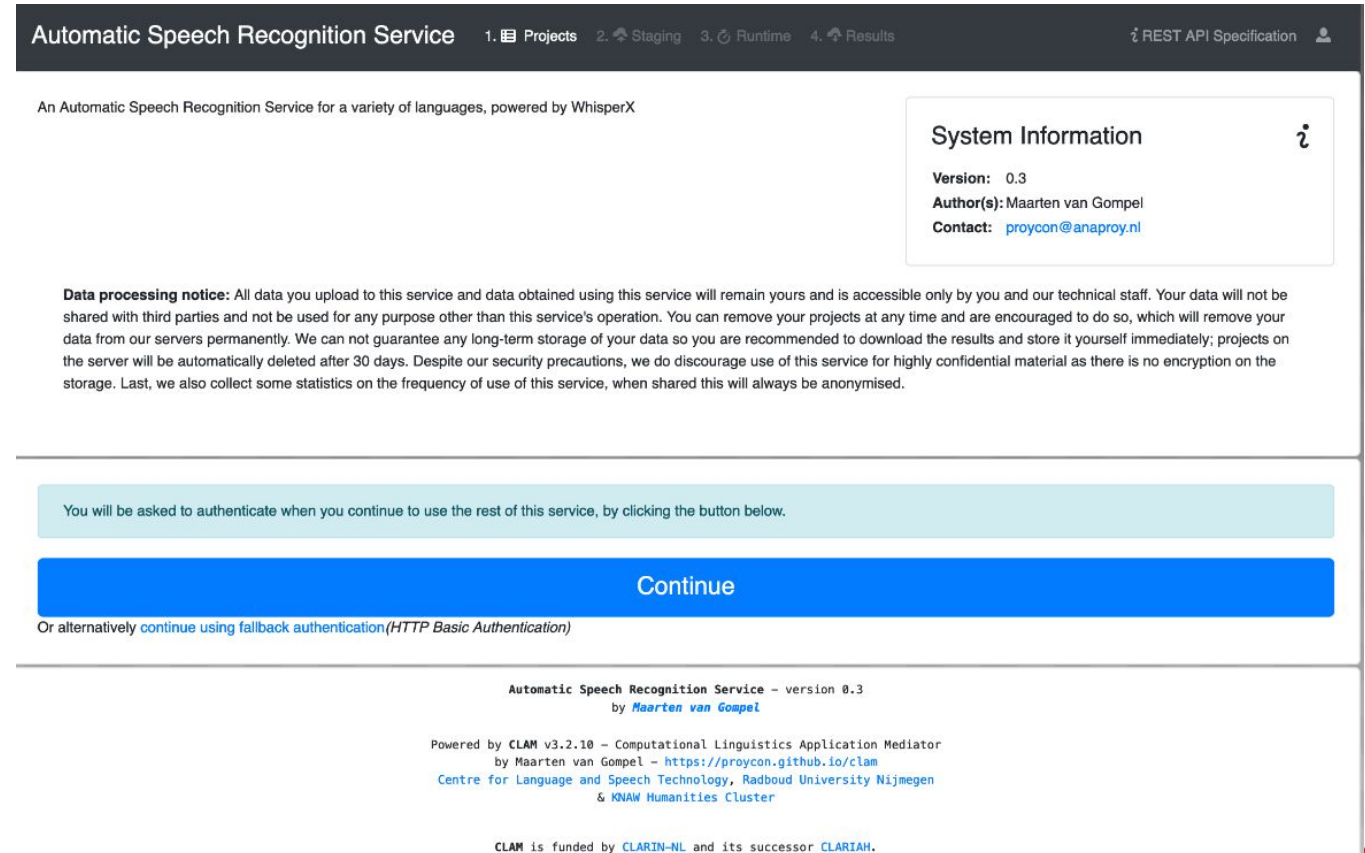
Alpino Cypher Dependency parsing SP0D: Syntactic profiler of Dutch UD: Universal Dependencies

Created: 2020-03-25 Modified: 2024-04-24

Step 1 From audio to text - Results

- We found and tried out two tools
- Audio file formats can be an issue!
- Quality of the output varied

→ Automatic Speech Recognition Service



The screenshot shows the web interface of the Automatic Speech Recognition Service. At the top, a dark navigation bar contains the service name and a progress indicator with four steps: 1. Projects, 2. Staging, 3. Runtime, and 4. Results (which is currently selected). A link to the REST API Specification is also present. Below the navigation bar, a subtitle reads 'An Automatic Speech Recognition Service for a variety of languages, powered by WhisperX'. On the right, a 'System Information' box displays the version (0.3), author (Maarten van Gompel), and contact email (proycon@anaproy.nl). A 'Data processing notice' paragraph explains that user data remains on the server for 30 days and is not shared. A light blue box informs users that authentication is required for the REST API. A large blue 'Continue' button is prominently displayed. Below the button, a link offers an alternative authentication method using fallback authentication (HTTP Basic Authentication). The footer section provides version information (0.3 by Maarten van Gompel), mentions the service is powered by CLAM v3.2.10, and lists the affiliations: Centre for Language and Speech Technology, Radboud University Nijmegen, and KNAW Humanities Cluster. It also states that the service is funded by CLARIN-NL and its successor CLARIAH.

Automatic Speech Recognition Service 1. Projects 2. Staging 3. Runtime 4. Results [REST API Specification](#)

An Automatic Speech Recognition Service for a variety of languages, powered by WhisperX

System Information ⓘ

Version: 0.3
Author(s): Maarten van Gompel
Contact: proycon@anaproy.nl

Data processing notice: All data you upload to this service and data obtained using this service will remain yours and is accessible only by you and our technical staff. Your data will not be shared with third parties and not be used for any purpose other than this service's operation. You can remove your projects at any time and are encouraged to do so, which will remove your data from our servers permanently. We can not guarantee any long-term storage of your data so you are recommended to download the results and store it yourself immediately; projects on the server will be automatically deleted after 30 days. Despite our security precautions, we do discourage use of this service for highly confidential material as there is no encryption on the storage. Last, we also collect some statistics on the frequency of use of this service, when shared this will always be anonymised.

You will be asked to authenticate when you continue to use the rest of this service, by clicking the button below.

Continue

Or alternatively continue using [fallback authentication](#) (HTTP Basic Authentication)

Automatic Speech Recognition Service - version 0.3
by [Maarten van Gompel](#)

Powered by CLAM v3.2.10 - Computational Linguistics Application Mediator
by Maarten van Gompel - <https://proycon.github.io/clam>
Centre for Language and Speech Technology, Radboud University Nijmegen
& KNAW Humanities Cluster

CLAM is funded by CLARIN-NL and its successor CLARIAH.

<https://github.com/opensource-spraakherkenning-nl/asrservice>
<https://tools.clariah.nl/asrservice/0.3/>

Step 1 From audio to text - Resulting text file

A: Welcome Maaïke, nice that you are here.

B: Nice to see you too Ricarda, happy to be here.

A: So what we want to do today is test the anonymisation tool Textwash or FAMTAFOS that was developed by Bennett Kleinberg and colleagues.

B: Yes, we want to try ourselves if we can de-identify all the identifying information in this little audio segment we are recording. It will serve as a test in our workshop "Making Qualitative Data Reusable" that we will hold during the DCC Spring Training Days on May 23rd 2024.

A: Right so to already throw in some identifying details - right now it is Tuesday 23 April 2024 11:00.



Welcome, Maaïke.

Nice that you're here.

Nice to see you, Ricarda.

Happy to be here.

So what we want to do today is test the anonymization tool Textwash or Fantafoss that was developed by Bennett, Kleinberg and colleagues.

Yes, we want to try ourselves if we can de-identify all the identifying information in this little audio segment that we are recording.

It will serve as a test in our workshop, making qualitative data reusable, that we will hold during the DCC Spring Training Days on May 23rd, 2024.

Right.

So to already throw in some identifying details, right now it is Tuesday, 23rd of April, 2024, 11 o'clock.

Step 2 De-identification - UKDA Tool

- UKDA Text anonymisation Tool
- Word plugin that identifies capital letters
- Easy to install, well described documentation
- Needs MS Word (txt files need to be imported into word).
- Does not work on Mac for us
- Still lot's of manual work to do the de-identification
- Language independent theoretically; works on the logic of capital letters for certain words

Step 2 De-identification - UKDA Tool - Results

A: Welcome Maaike, nice that you are here.

B: Nice to see you too Ricarda, happy to be here.

A: So what we want to do today is test the anonymisation tool Textwash or FAMTAFOS that was developed by Bennett Kleinberg and colleagues.

B: Yes, we want to try ourselves if we can de-identify all the identifying information in this little audiosegment we are recording. It will serve as a test in our workshop "Making Qualitative Data Reusable" that we will hold during the DCC Spring Training Days on May 23rd 2024.

A: Right so to already throw in some identifying details - right now it is Tuesday 23 April 2024 11:00.



A: Welcome, Maaike. Nice that you're here.

B: Nice to see you, Ricarda. Happy to be here.

A: So what we want to do today is test the anonymization tool Textwash or FAMTAFOS that was developed by Bennett Kleinberg and colleagues.

B: Yes, we want to try ourselves if we can de-identify all the identifying information in this little audio segment that we are recording. It will serve as a test in our workshop, making qualitative data reusable, that we will hold during the DCC Spring Training Days on May 23rd, 2024.

A: Right. So to already throw in some identifying details, right now it is Tuesday, 23rd of April, 2024, 11 o'clock.

Step 2 De-identification - Textwash

- Textwash
- Python based tool that replaces identifying information automatically
- At the moment, difficult to run without help if you have limited time
 - We asked help from the author Bennett Kleinberg who ran the analysis for us (Thank you!!!)
 - However: they are constantly improving the Tool and its usability!
- Platform independent (runs on Python)
- Performs the de-identification for you
- Available models for English and Dutch

<https://github.com/ben-aaron188/textwash>

Step 2 De-identification - Textwash - Results

A: Welcome Maaïke, nice that you are here.

B: Nice to see you too Ricarda, happy to be here.

A: So what we want to do today is test the anonymisation tool Textwash or FAMTAFOS that was developed by Bennett Kleinberg and colleagues.

B: Yes, we want to try ourselves if we can de-identify all the identifying information in this little audio segment we are recording. It will serve as a test in our workshop "Making Qualitative Data Reusable" that we will hold during the DCC Spring Training Days on May 23rd 2024.

A: Right so to already throw in some identifying details - right now it is Tuesday 23 April 2024 11:00.



A: Welcome, PERSON_FIRSTNAME_2. Nice that you're here.

B: Nice to see you, PERSON_FIRSTNAME_3. Happy to be here.

A: So what we want to do today is test the anonymization tool OTHER_1 or OTHER_2 that was developed by PERSON_FIRSTNAME_1 PERSON_LASTNAME_1 and colleagues.

B: Yes, we want to try ourselves if we can de-identify all the identifying information in this little audio segment that we are recording. It will serve as a test in our workshop, making qualitative data reusable, that we will hold during the OTHER_3 OTHER_4 Training Days on DATE_1 DATE_2, NUMERIC_1.

A: Right. So to already throw in some identifying details, right now it is DATE_5, DATE_2 DATE_3 DATE_1, NUMERIC_1, TIME_1 o'clock.



Time For Questions





How to publish data for reuse

Selecting a repository

- Archive data so it is securely stored for the long term
- Publish (meta)data so the dataset can be found and reused

→ Selecting a Trustworthy Digital Repository

Selecting a repository

- Archive data so it is securely stored for the long term
- Publish (meta)data so the dataset can be found and reused

→ Selecting a Trustworthy Digital Repository



Selecting a repository

- Archive data so it is securely stored for the long term
- Publish (meta)data so the dataset can be found and reused

→ Selecting a Trustworthy Digital Repository

Selecting a repository

- Archive data so it is securely stored for the long term
- Publish (meta)data so the dataset can be found and reused

→ Selecting a Trustworthy Digital Repository

- Your institute may have regulations or recommendations

Data Station Social Sciences and Humanities



Data
deponeren

Deponeren >>>

[Hulp nodig bij deponeren?](#)



Data zoeken

Typ hier je zoekopdracht in...

Zoek

[Hulp nodig bij zoeken?](#)

The logo for Data verseNL, featuring a stylized network of three blue circles connected by lines above the text "Data verseNL" in a bold, blue, sans-serif font.

Preparing your dataset

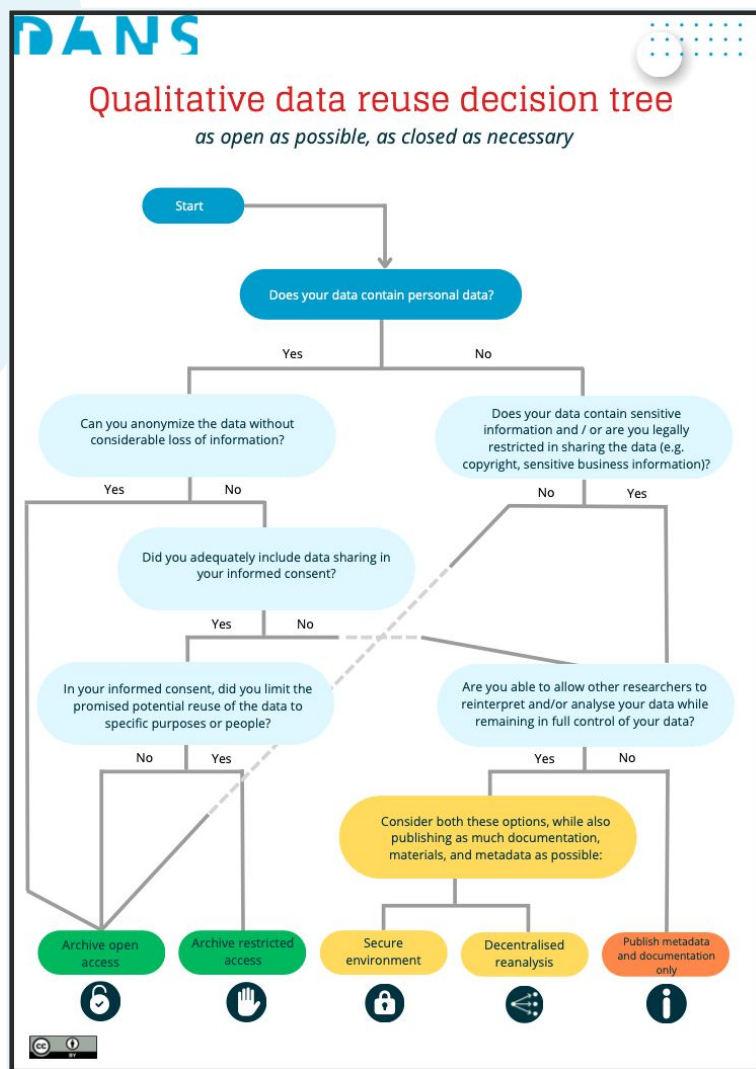
- Prepare your dataset according to the repository guidelines
 - Metadata and Documentation
 - Preferred file formats
- Available guidance (for qualitative data)
 - CESSDA - Data Management Expert Guide - [Documentation and Metadata](#)
 - DANS - [Deposit manual](#)
 - DANS - [Preferred file formats](#)
 - FSD - [Processing qualitative data files](#)

→ **Decide** whether your data can be published **open** or under **restricted access**

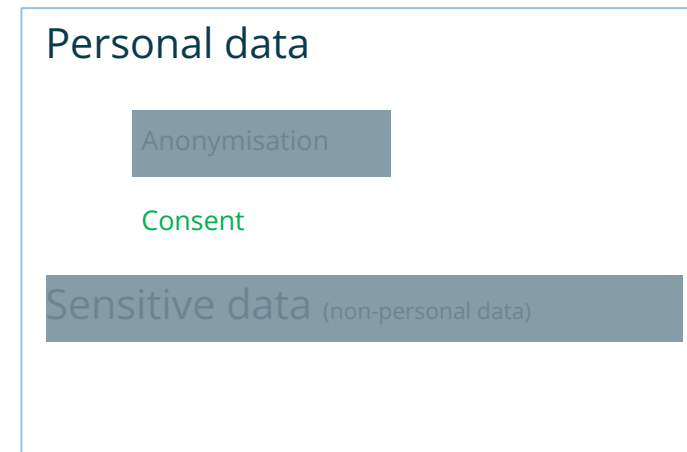
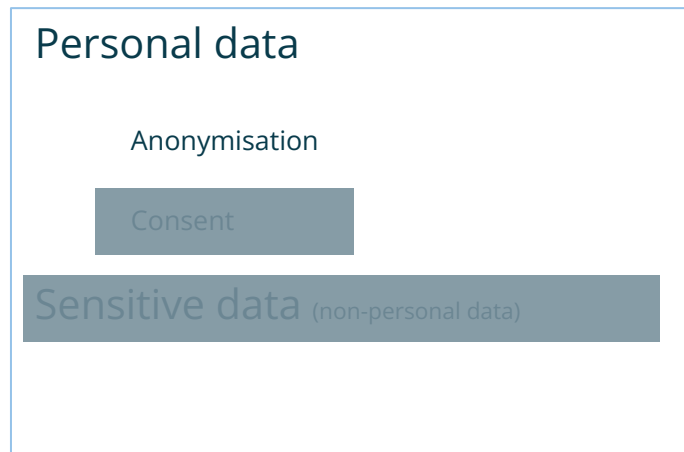
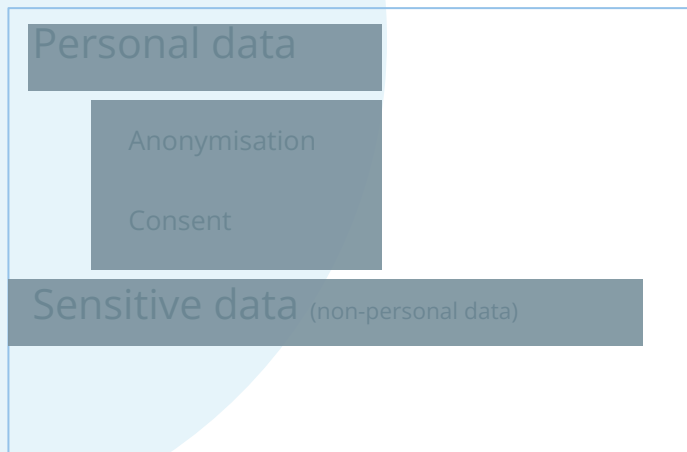


Making qualitative data reusable - Decision Tree

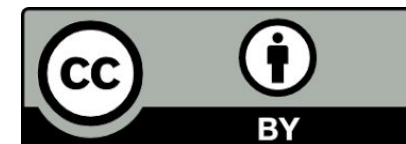
<https://doi.org/10.5281/zenodo.8160890>



Decision tree



Archive open
access



Decision tree



Archive
restricted access

Personal data

Anonymisation

Consent

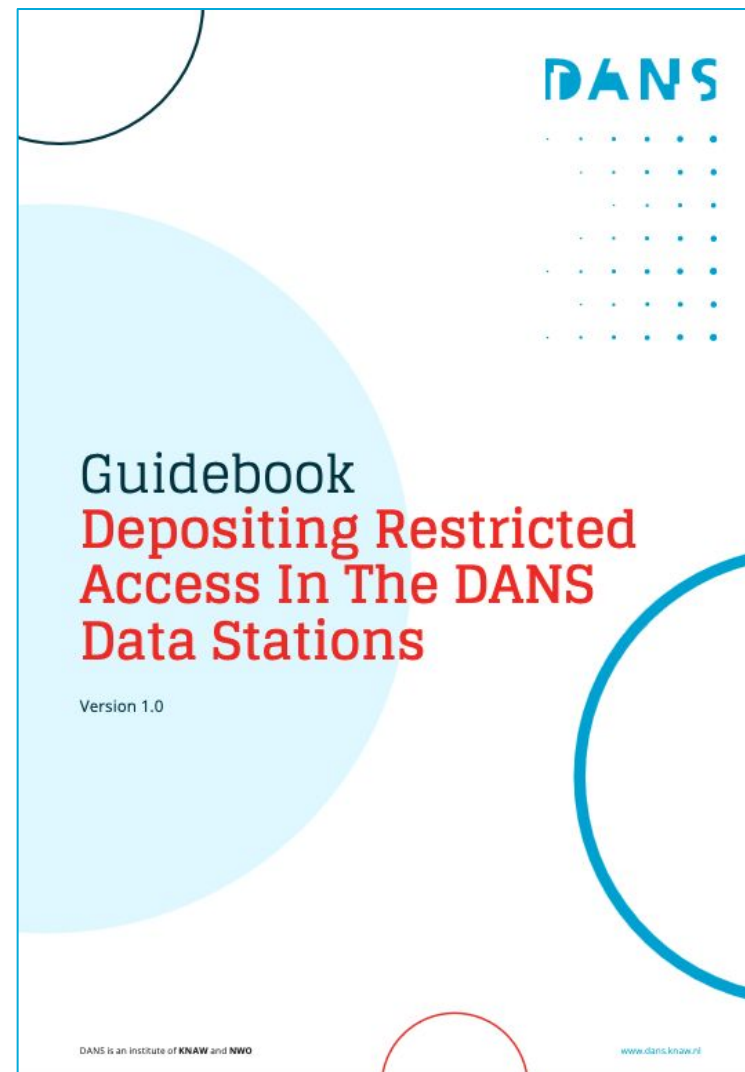
Sensitive data (non-personal data)

Decision tree

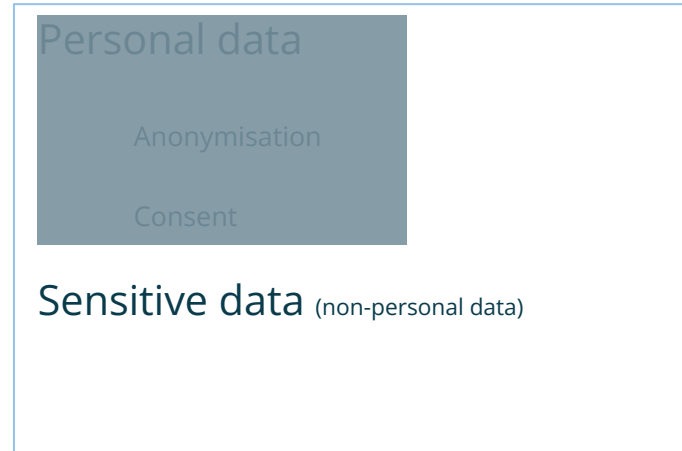
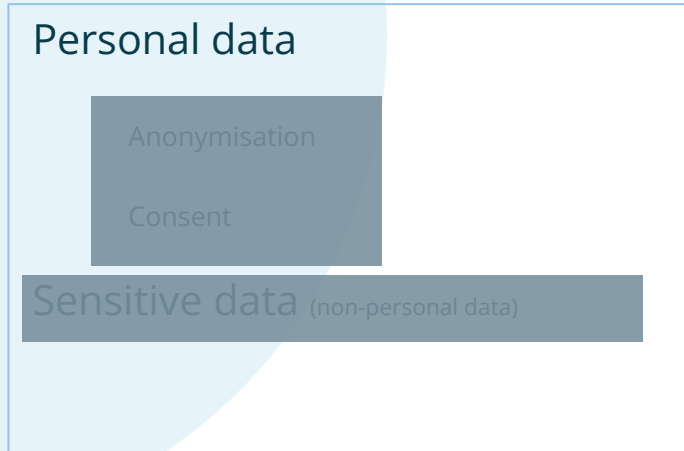
New guidebook
Braukmann, R., Verburg, M., & Mahabier,
W. (2024). Guidebook - Depositing
Restricted Access In The DANS Data
Stations (1.0). Zenodo.
<https://doi.org/10.5281/zenodo.10887484>



Archive
restricted access



Decision tree



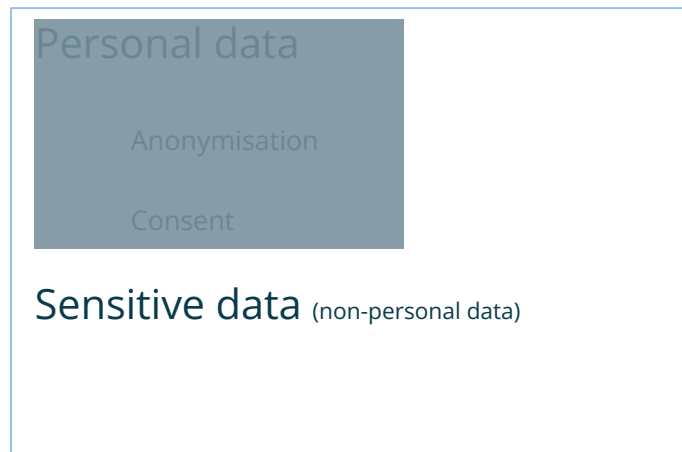
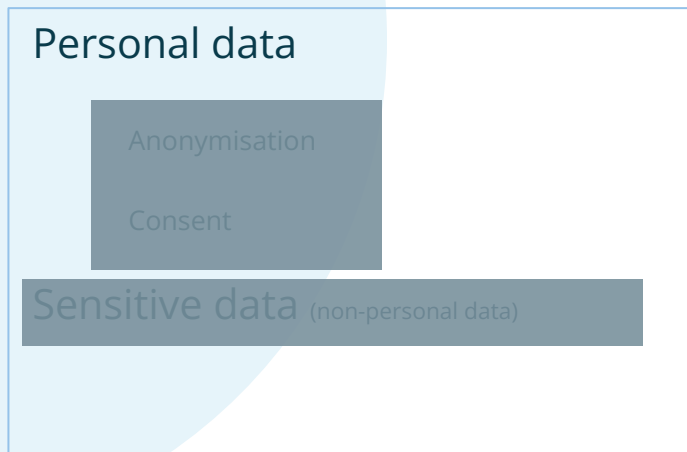
If you are able to allow other researchers to reinterpret and/or analyse your data while remaining in full control of your data



Secure environment

Publish project documentation and metadata about your data to make it findable.

Decision tree



If you are able to allow other researchers to reinterpret and/or analyse your data while remaining in full control of your data



Decentralised
reanalysis

<https://www.nwo.nl/en/projects/40621eb014>

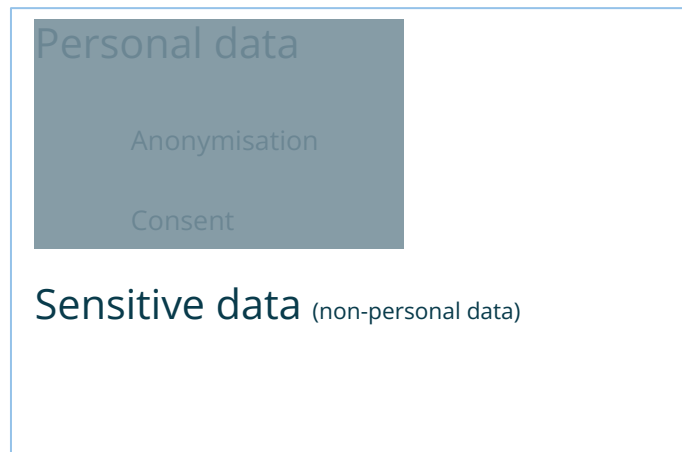
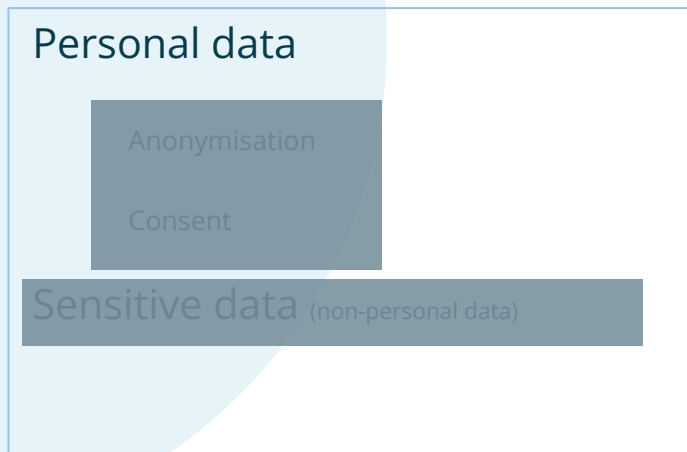
Research & results > Innovating Methods for Open Science in Qualitative Management Research (OPEN-QUAL)

Innovating Methods for Open Science in
Qualitative Management Research (OPEN-
QUAL)

Publish project documentation and metadata about your data to make it findable.

<https://doi.org/10.5281/zenodo.8160890>

Decision tree



If you are NOT able to allow other researchers to reinterpret and/or analyse your data while remaining in full control of your data

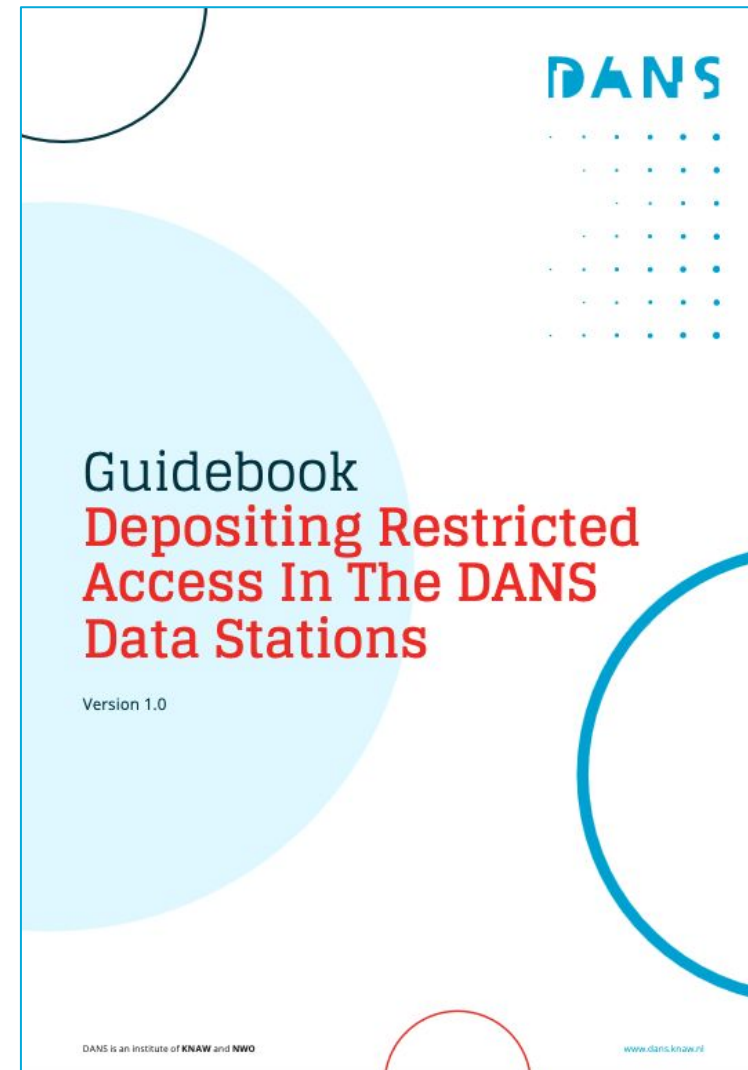


Publish
metadata only

Publish project documentation and metadata about your data to make it findable.

Guidebook on Restricted Access

- Recently published to provide additional guidance
- Focus on depositing in the DANS Data Station SSH
 - Access categories and licences we offer
 - Metadata fields around access information
- Does include general guidance
 - Tips on selecting the right access for your dataset
 - Tips on what information to document
 - Template for a Data Access Protocol



<https://doi.org/10.5281/zenodo.10887484>

Prepare a Data Access Protocol

- DAP template takes you through all the questions to consider when restricting access
 - Do you allow
 - use beyond research?
 - commercial use?
 - educational use?
 - use outside of Europe?
 - Do you require a motivation for reuse? How is it evaluated?
 - Who is responsible to evaluate requests? How is this organized now and in the future?
- Preparing this well makes it easier for you to manage access
- Researchers who want to reuse data can better evaluate if they are allowed to
-

Download DAP template:
<https://doi.org/10.5281/zenodo.10887571>



Q&A & Wrap Up



Contributors

Making Qualitative Data Reusable

- Widia Mahabier - DANS
- Jetze Touber – DANS
- Samantha Willemsen – DANS
- Hans Berends - VU
- Eric Haynes – VU
- Kacana Khadjavi Pour – VU
- Fleur Deken - VU

Questions?

Open Hour Social Sciences and Humanities

Q&A on Open Science, data storage
and Research Data Management

Every Monday
10:00 - 11:00 CET

DANS



Thank you for your attention

Visit our website www.dans.knaw.nl

And follow us online




Mastodon @DANS_knaw_nwo



LinkedIn @DANS



X @DANS_knaw_nwo



*Please fill out
the evaluation
form*

