

u^b

b

**UNIVERSITÄT
BERN**

u^b

Data in disguise

Different tools for anonymising your research data

Christine Krebs, Jennifer Morger
Data Stewards, Open Science Team

Research Skills@vonRoll

Content

- Introduction: Why and how to anonymise your research data
- Presentation of two anonymisation tools:
 - scdMicro for tabular data
 - QualiAnon for textual data
- Take home messages

Definitions / Why

Personal Data:

FADP Art 5 a.: any information relating to an identified or identifiable natural person

FADP Art. 39 a.: the data are anonymised as soon as the purpose of processing permits

Anonymised data is not subject to data protection regulations anymore

From: [SR 235.1 - Federal Act of 25 September 2020 on D... | Fedlex](#)

Definitions

Anonymisation:

HRA Art. 3 i.: Anonymised *biological material and anonymised health-related data* means biological material and health-related data which cannot (without disproportionate effort) be traced to a specific person

From [SR 810.30 - Federal Act of 30 September 2011 on ... | Fedlex](#)

u^b How to pseudonymise

Code list

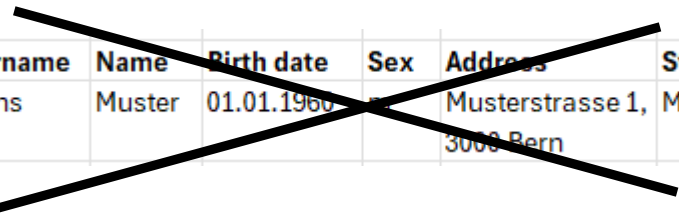
Surname	Name	Birth date	Sex	Address	Study ID
Hans	Muster	01.01.1960	m	Musterstrasse 1, 3000 Bern	MM4m73

Pseudonymised/coded data

Study ID	Sex	Diagnosis	RT_1	RT_2	RT_3
MM4m73	m	MDD	0.782	0.485	0.862

How to anonymise

No Code list



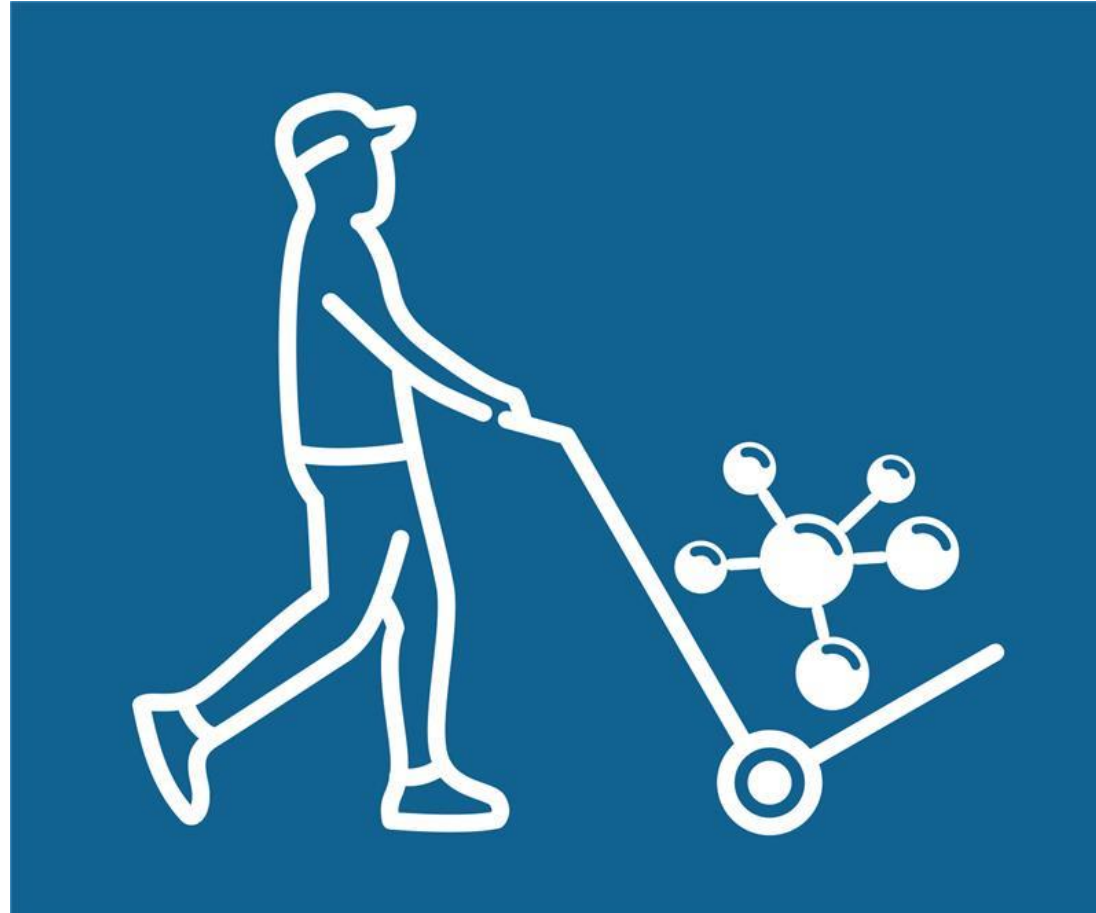
Surname	Name	Birth date	Sex	Address	Study ID
Hans	Muster	01.01.1960	m	Musterstrasse 1, 3000 Bern	MM4m73

Anonymised data

Study ID	Sex	Diagnosis	RT_1	RT_2	RT_3
*****	m	MDD	0.782	0.485	0.862

u^b

Questions?



Direct and indirect identifiers

Direct identifiers -> Should be removed / coded

Indirect identifiers -> «key variables», will be anonymised

Surname	Name	Sex	Birth date	AHV_Nr	ZIP	RT_1	RT_2	RT_3
Hans	Muster	m	01.01.1960	756.1234.5678.90	3000	0.782	0.485	0.862

- Statistical Disclosure Control (SDC) = microdata anonymisation
- Open Source, package for "r"
- Webbased GUI can be launched (sdcApp())
- Various examples and guides are available online
- Developed by Matthias Templ et al.

sdcMicro

Reference

📖 README

sdcMicro

R-CMD-check **passing** CRAN **5.7.8** downloads **681/month** mentioned in **awesome**

sdcMicro is an R-package to anonymize microdata. Most functionalities of the package are also available via an interactive shiny-based graphical user interface.

The online documentation can also be found at sdctools.github.io/sdcMicro.

Templ M, Kowarik A, Meindl B (2015). “Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro.” *Journal of Statistical Software*, **67**(4), 1–36. [doi:10.18637/jss.v067.i04](https://doi.org/10.18637/jss.v067.i04).

sdcMicro GUI

Step 1: import data to "r" or GUI

The screenshot displays the sdcMicro GUI interface. At the top, a navigation bar includes the title 'sdcMicro GUI' and several menu items: 'About/Help', 'Microdata' (which is currently selected), 'Anonymize', 'Risk/Utility', 'Export Data', 'Reproducibility', and 'Undo'. On the left side, there is a sidebar titled 'Select data source' containing a list of options: 'Testdata/internal data' (highlighted with a blue bar), 'R-dataset (.rdata)', 'SPSS-file (.sav)', 'SAS-file (.sas7dat)', 'CSV-file (.csv, .txt)', and 'STATA-file (.dta)'. The main content area is titled 'Uploading microdata' and contains the instruction 'Load the dataset to be anonymized.' Below this, a light gray box prompts the user to 'Select a test dataset or any object in your current workspace'. This prompt is followed by a dropdown menu that currently shows 'testdata'. A blue 'Load data' button is positioned below the dropdown menu.

sdcMicro GUI

Step 2: explore data



sdcMicro GUI

Step 2: define key variables

sdcMicro GUI About/Help Microdata **Anonymize** Risk/Utility Export Data Reproducibility Undo

Anonymize

Select variables and set parameters to create the SDC problem.

Select variables ⓘ

Variable name	Type	Key variables	Weight	Hierarchical identifier	PRAM	Delete	Number of levels	Number of missing
urbrur	integer	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2	0
roof	integer	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	5	0
walls	integer	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	3	0
water	integer	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	8	0
electcon	integer	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	3	0
relat	factor	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	9	0
sex	factor	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	2	0
age	numeric	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	88	0
hhcivil	factor	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4	0
expend	integer	<input checked="" type="radio"/> No <input type="radio"/> Cat. <input type="radio"/> Cont.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4580	0

Error: No categorical key variables selected

Set additional parameters

Parameter 'alpha' ⓘ

Parameter 'seed' ⓘ

Explore variables

urbrur (integer)

Number of levels including missing (NA): 3

Step 3: check summary

sdcmicro GUI[About/Help](#)[Microdata](#)[Anonymize](#)[Risk/Utility](#)[Export Data](#)[Reproducibility](#)[Undo](#)

View/Analyze existing sdcProblem
[Show summary](#)

[Explore variables](#)
[Add linked variables](#)
[Create new IDs](#)

Anonymize categorical variables
[Recoding](#)
[k-Anonymity](#)
[PRAM \(simple\)](#)
[PRAM \(expert\)](#)
[Suppress values with high risks](#)

Anonymize numerical variables
[Top/bottom coding](#)
[Microaggregation](#)
[Adding noise](#)
[Rank swapping](#)

Summary of dataset and variable selection

The loaded dataset consists of **4580** records and **15** variables.

Categorical key variable(s): **relat sex hhcivil**
Numerical key variable(s): **age**

Computation time

The current computation time was ~ **4.29 seconds**.

Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
relat	9 (9)	508.889 (508.889)	1 (1)
sex	2 (2)	2290.000 (2290.000)	2284 (2284)
hhcivil	4 (4)	1145.000 (1145.000)	38 (38)

Risk measures for categorical key variables

We expect **38.00 (0.83%)** re-identifications in the population, as compared to **38.00 (0.83%)** re-identifications in the original data.

66 observations have a higher risk than the risk in the main part of the data, as compared to **66** observations in the original data. ⓘ

Step 4: anonymise data: suppression

sdcmicro GUI[About/Help](#)[Microdata](#)[Anonymize](#)[Risk/Utility](#)[Export Data](#)[Reproducibility](#)[Undo](#)

View/Analyze existing sdcProblem
[Show summary](#)
[Explore variables](#)
[Add linked variables](#)
[Create new IDs](#)

Anonymize categorical variables
[Recoding](#)
[k-Anonymity](#)
[PRAM \(simple\)](#)
[PRAM \(expert\)](#)
[Suppress values with high risks](#)

Anonymize numerical variables

Establish k-anonymity

k-anonymity will be established by suppressing or rather setting some values in the selected categorical key variables to **NA**.

By default, the key variables will be considered for suppression in the order of their number of distinct categories. A variable with many categories is less likely to have values suppressed than a variable with few categories. It is also possible to set the order by specifying an importance vector.

You may also decide to apply the procedure for all possible subsets of key variables. This is useful, if you have many key variables and can reduce computation time. You can set a different value for the parameter **k** for each size of subsets.

Do you want to apply the method for each group defined by the selected variable? ⓘ

no stratification

Do you want to modify importance of key variables for suppression? ⓘ

☒ No ☐ Yes

Tip - The total number of suppressions is likely to increase by specifying an importance vector. Specifying an importance vector can affect the computation time.

Apply k-anonymity to subsets of key variables? ⓘ

☒ No ☐ Yes

Variable selection

Variable name	Type	Additional suppressions by local suppression algorithm
relat	cat. key variable	1
sex	cat. key variable	6
age	num. key variable	

Additional parameters

sdcMicro

Step 4: anonymise data: recode

sdcMicro GUI

About/Help
Microdata
Anonymize
Risk/Utility
Export Data
Reproducibility
Undo

View/Analyze existing sdcProblem

Show summary

Explore variables

Add linked variables

Create new IDs

Anonymize categorical variables

Recoding

k-Anonymity

PRAM (simple)

PRAM (expert)

Supress values with high risks

Anonymize numerical variables

Top/bottom coding

Microaggregation

Adding noise

Rank swapping

Reset SDC problem

Recode categorical key variables

To reduce risk, it is often useful to combine the levels of categorical key variables into a new, combined category. You need to select a categorical key variable and then choose two or more levels, which you want to combine. Once this has been done, a new label for the new category can be assigned.

Note: If you only select only one level, you can rename the selected value.

Choose factor variable

relat

Select levels to recode/combine

Level	Frequency
1	1000
2	800
3	2600
4	50
5	100
6	50
7	100
8	50
9	50
NA	50

Variable selection

Variable name	Type	Additional suppressions by local suppression algorithm
relat	cat. key variable	0
sex	cat. key variable	17
hhcivil	cat. key variable	0
age	num. key variable	

Additional parameters

Parameter	Value
number of records	4580
alpha	1
random seed	0

k-anonymity

k-anonymity	Modified data	Original data
2-anonymity	4 (0.09%)	9 (0.20%)
3-anonymity	6 (0.13%)	17 (0.37%)
5-anonymity	22 (0.48%)	33 (0.72%)

17

April 29, 2025, Bern

University Library of Bern

Step 5: check success of anonymisation

sdcmicro GUI[About/Help](#)[Microdata](#)[Anonymize](#)[Risk/Utility](#)[Export Data](#)[Reproducibility](#)[Undo](#)

View/Analyze existing sdcProblem
[Show summary](#)
Explore variables
Add linked variables
Create new IDs
Anonymize categorical variables
Recoding
k-Anonymity
PRAM (simple)
PRAM (expert)
Suppress values with high risks
Anonymize numerical variables
Top/bottom coding
Microaggregation
Adding noise

We expect **31.00** (**0.68%**) re-identifications in the population, as compared to **38.00** (**0.83%**) re-identifications in the original data.

53 observations have a higher risk than the risk in the main part of the data, as compared to **66** observations in the original data. ⓘ

Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	4 (0.087%)	9 (0.197%)
3-anonymity	6 (0.131%)	17 (0.371%)
5-anonymity	22 (0.480%)	33 (0.721%)

Information on local suppression

Key variable	Number of suppressions	Total missing values (NA) before applying local suppression	Total missing values (NA) after applying local suppression
relat	0 (0.000%)	0 (0.000%)	0 (0.000%)
sex	17 (0.371%)	0 (0.000%)	17 (0.371%)
hhcivil	0 (0.000%)	0 (0.000%)	0 (0.000%)

What to consider

- Make sure, that your session protocol and new dataset are stored
- With the GUI you can mostly undo only the last step
- Sample weights: if you have no sample weights, use k-anonymity
- Export the script from the GUI to be able to replicate your anonymisation steps

Qualitative Text Data Anonymization

- Unstructured or semi-structured data (e.g., interview transcripts, focus group notes)
- Requires context-sensitive anonymization, such as pseudonymization, redaction, and substitution
- Aims to preserve the meaning and context of the data for research purposes
- Involves identifying personal identifiers and sensitive content in textual data (names, locations, etc.)

QualiAnon

Reference

QualiAnon

Qualiservice tool for anonymizing text data



Tom Nicolai, Kati Mozygemba, Susanne Kretzer, Betina Hollstein, Egor Gordeev (2025): QualiAnon - Qualiservice tool for anonymizing text data (version 1.5.0). Qualiservice. University of Bremen. Software available at: <https://github.com/pangaea-data-publisher/qualianon>

QualiAnon

- Semi-automated tool for anonymizing/pseudonymizing text data
- Tailored for qualitative data such as interviews and focus group transcripts.
- Supports the concept of “Flexible Anonymization”
- Open source; available for download on GitHub
- Developed by the Research Data Center Qualiservice

QualiAnon

Step 1: install and access the tool

📖 README

Qualiservice tool for anonymizing text data



<https://www.qualiservice.org>

If you have any questions regarding the use of QualiAnon for your research project, please send us an email to:

qualianon@uni-bremen.de

Documentation

- User manual: https://docs.google.com/document/d/1fLLYvsgXjh_p9p_E1fhikkIPQb19VUiltbRgaWQoD-M/edit?usp=sharing

Installation + Requirements

[Wiki - Installation instructions](#)

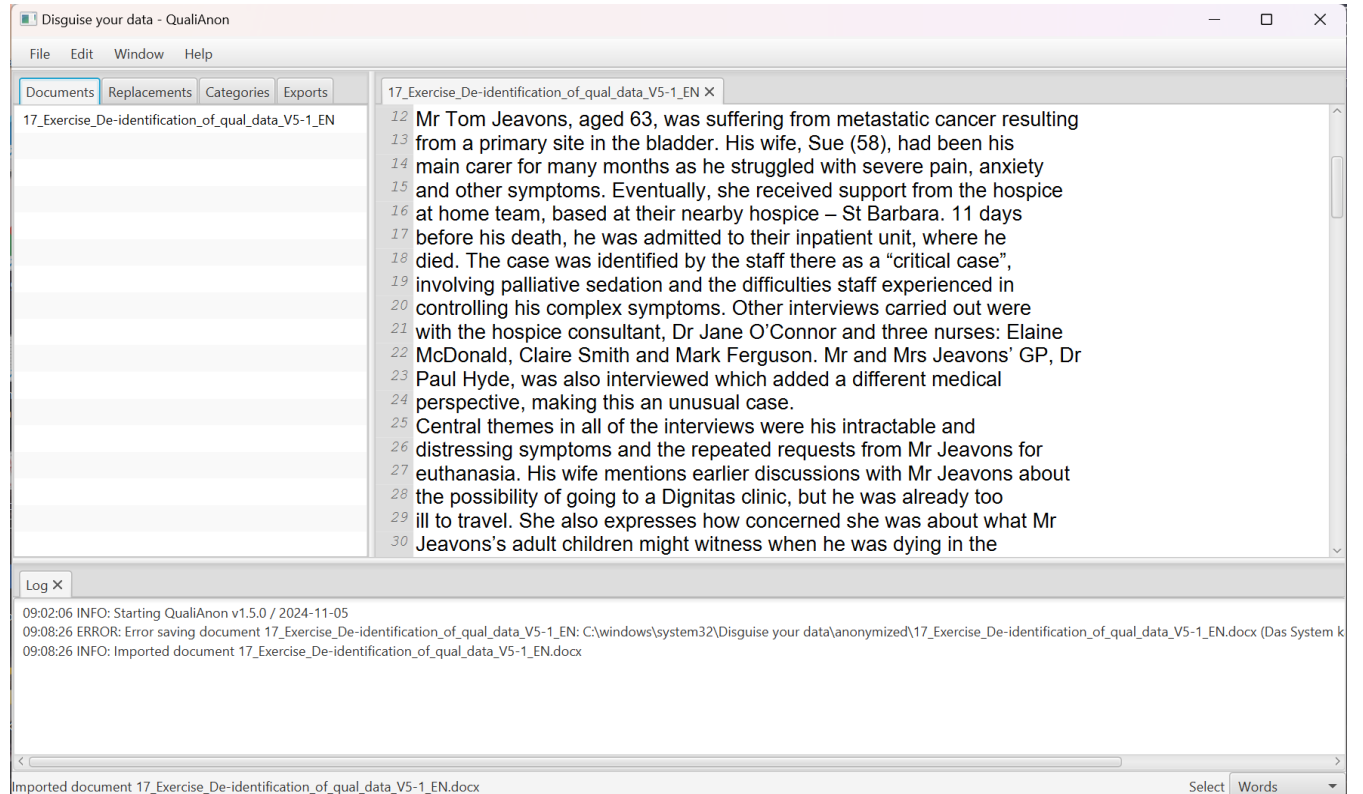
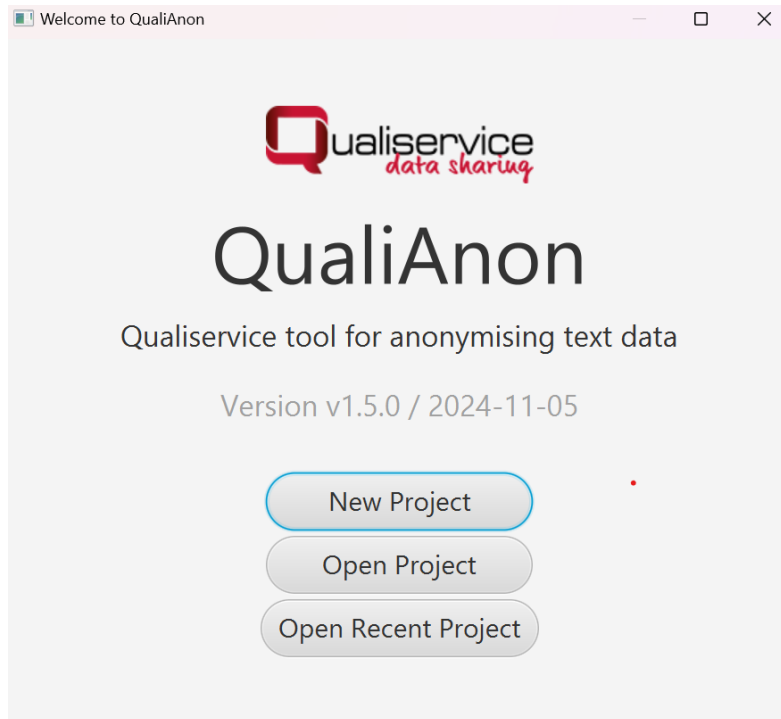
Download

Download the latest release from the "Release" section at <https://github.com/pangaea-data-publisher/qualianon/releases>

Citation

QualiAnon

Step 2: New Project & import documents

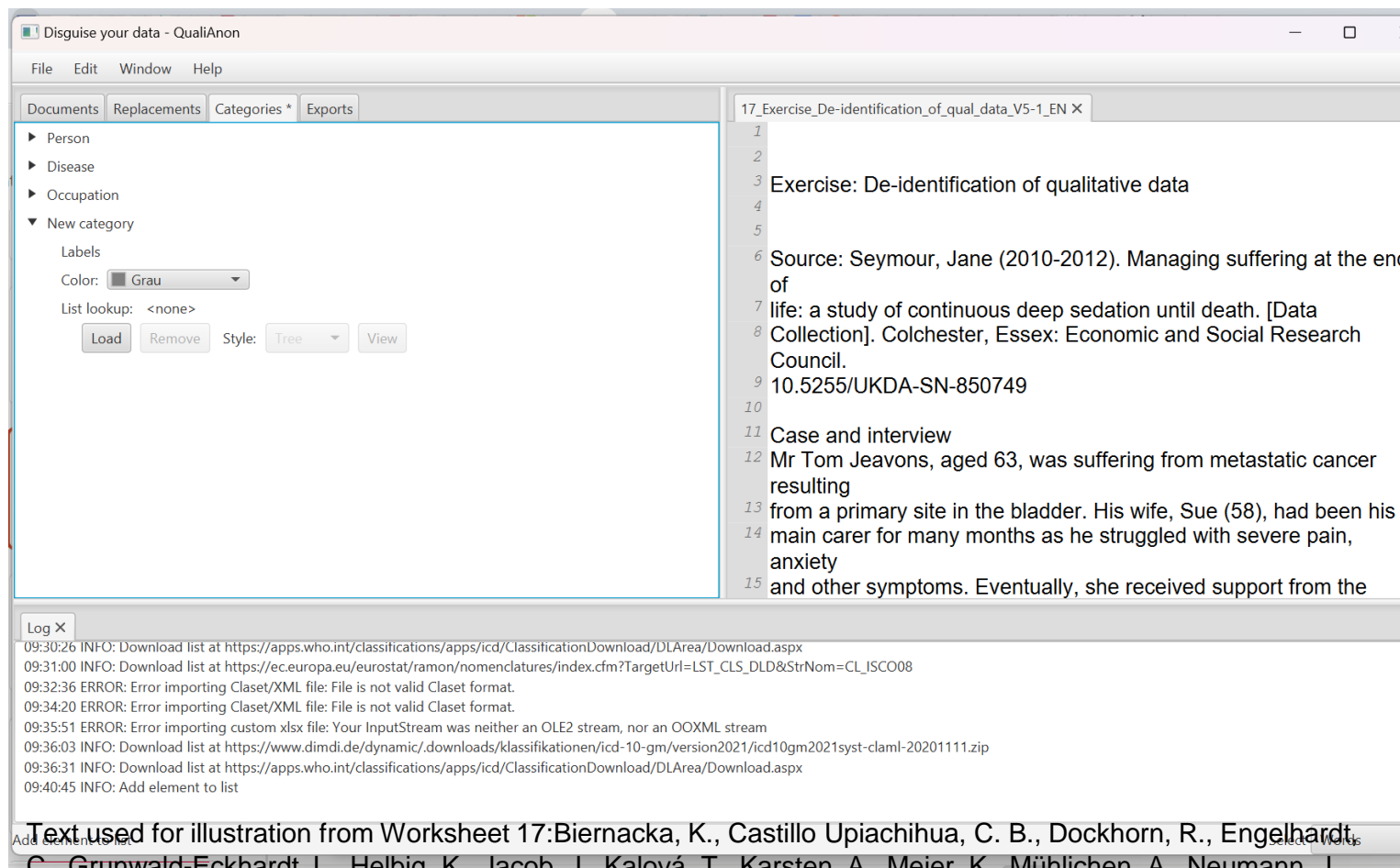


Screenshots: Text used for illustrating workflow: Worksheet 17 from the Train-the-Trainer Concept:

Biernacka, K., Castillo Upiachihua, C. B., Dockhorn, R., Engelhardt, C., Grunwald-Eckhardt, L., Helbig, K., Jacob, J., Kalová, T., Karsten, A., Meier, K., Mühlichen, A., Neumann, J., Petersen, B., Scherreiks, P., Slowig, B., Trautwein-Bruns, U., Wilbrandt, J., & Wiljes, C. (2024). Train-the-Trainer Concept on Research Data Management (5.1) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.13927614>

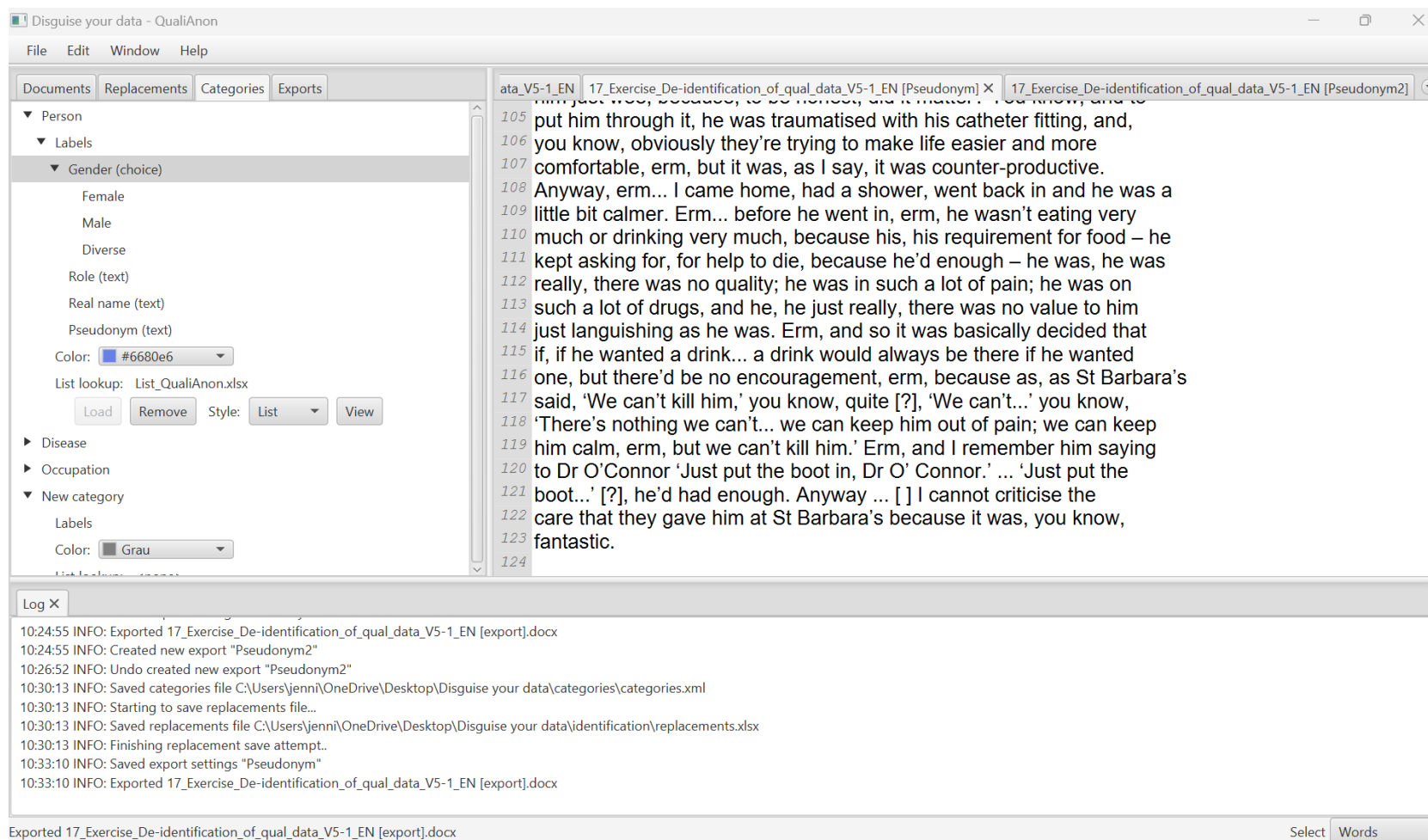
QualiAnon

Step 3: define categories

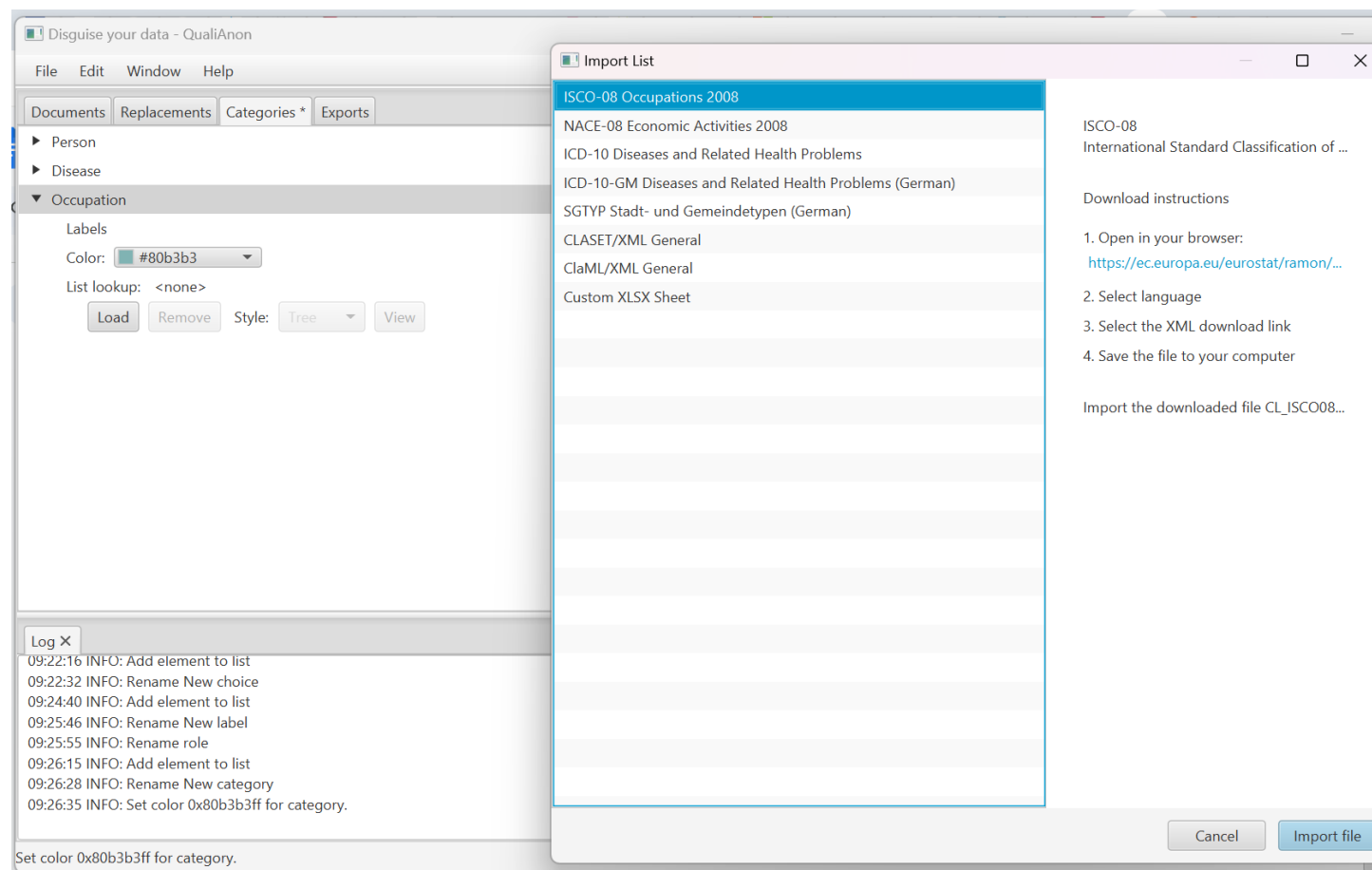


Text used for illustration from Worksheet 17: Biernacka, K., Castillo Upiachihua, C. B., Dockhorn, R., Engelhardt, C., Grunwald-Eckhardt, L., Helbig, K., Jacob, J., Kalová, T., Karsten, A., Meier, K., Mühlichen, A., Neumann, J., Petersen, B., Scherreiks, P., Slowig, B., Trautwein-Bruns, U., Wilbrandt, J., & Wiljes, C. (2024). Train-the-Trainer Concept on Research Data Management (5.1) [Computer software].

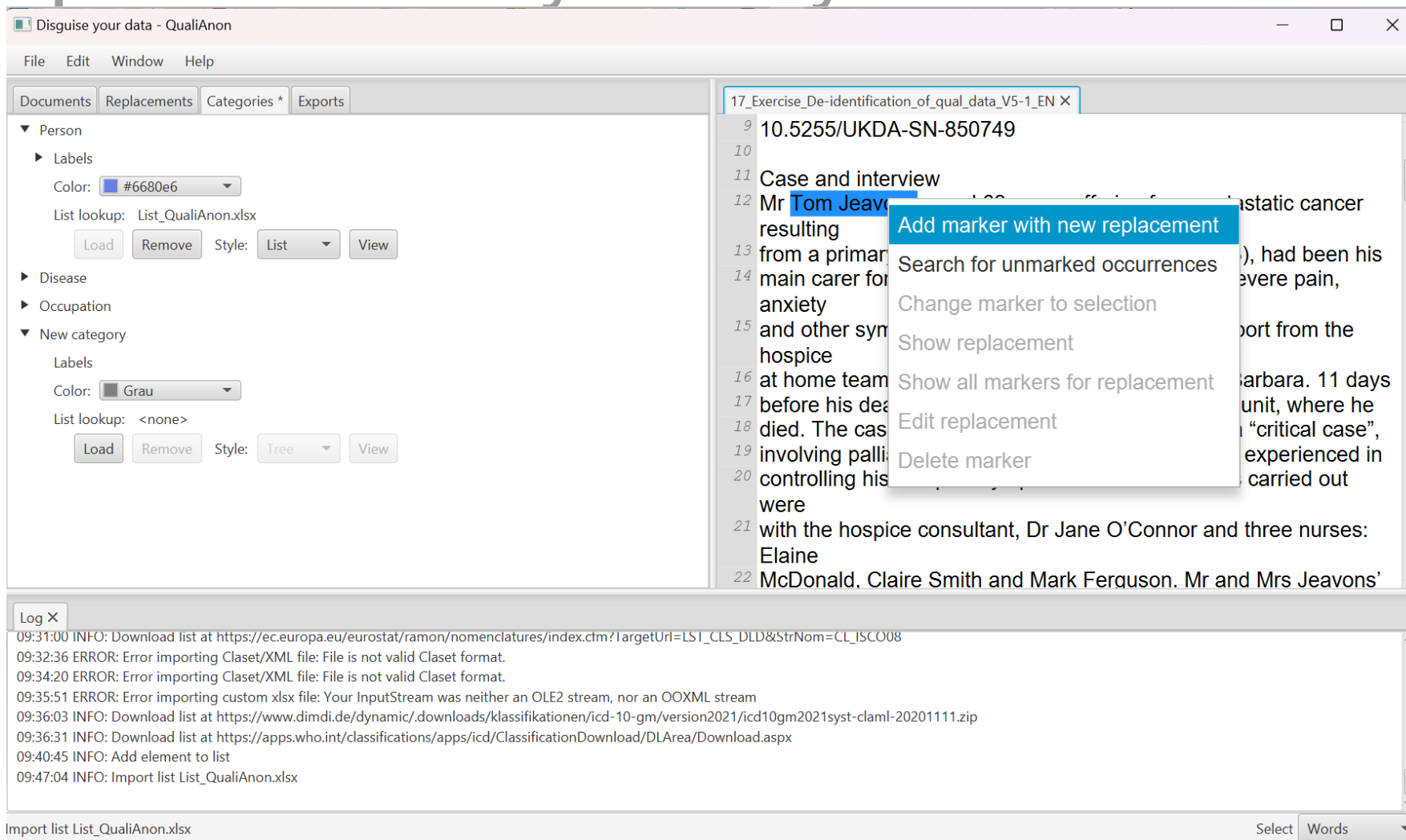
Step 3: define categories - lists



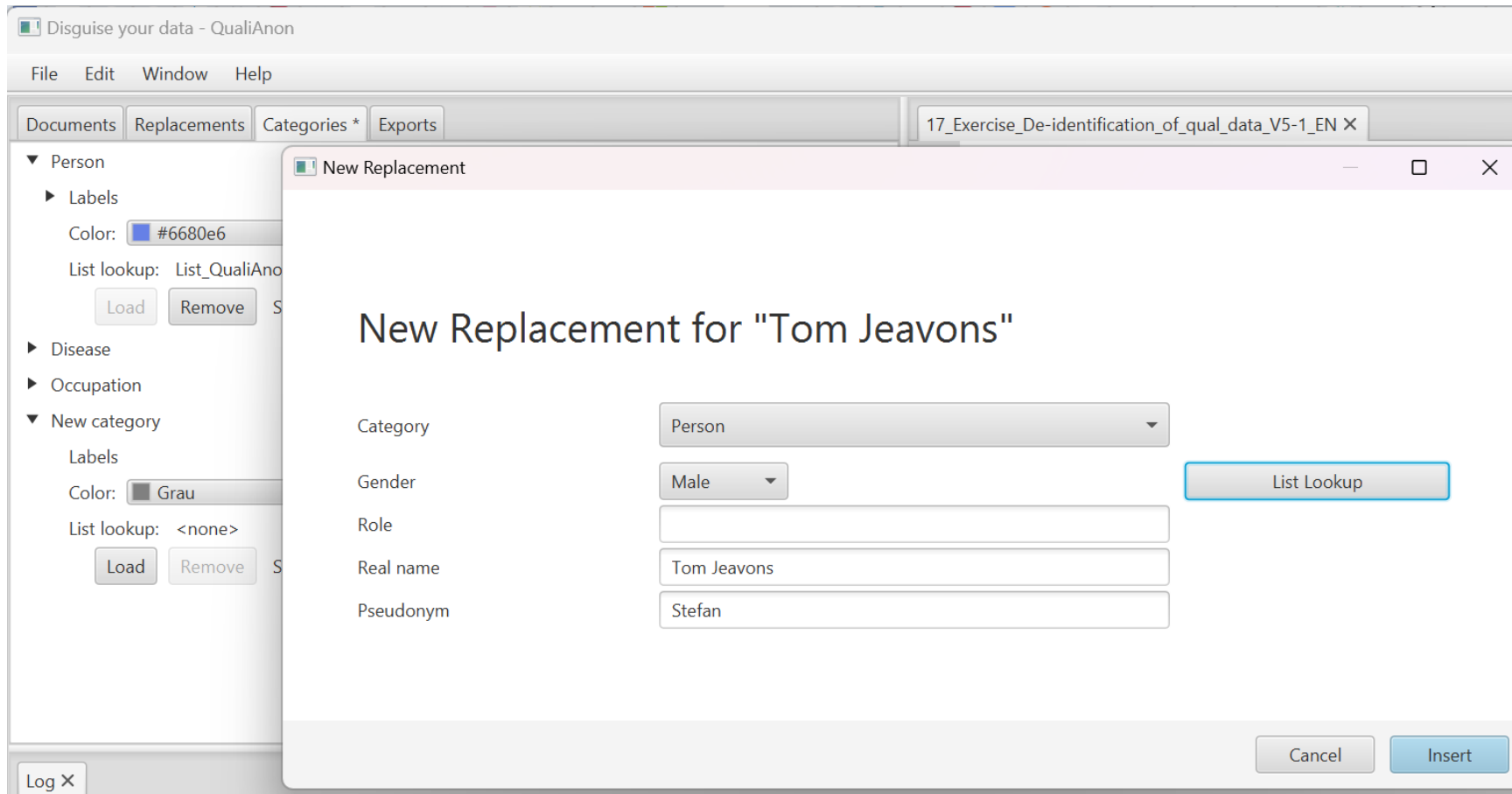
Step 3: define categories - lists



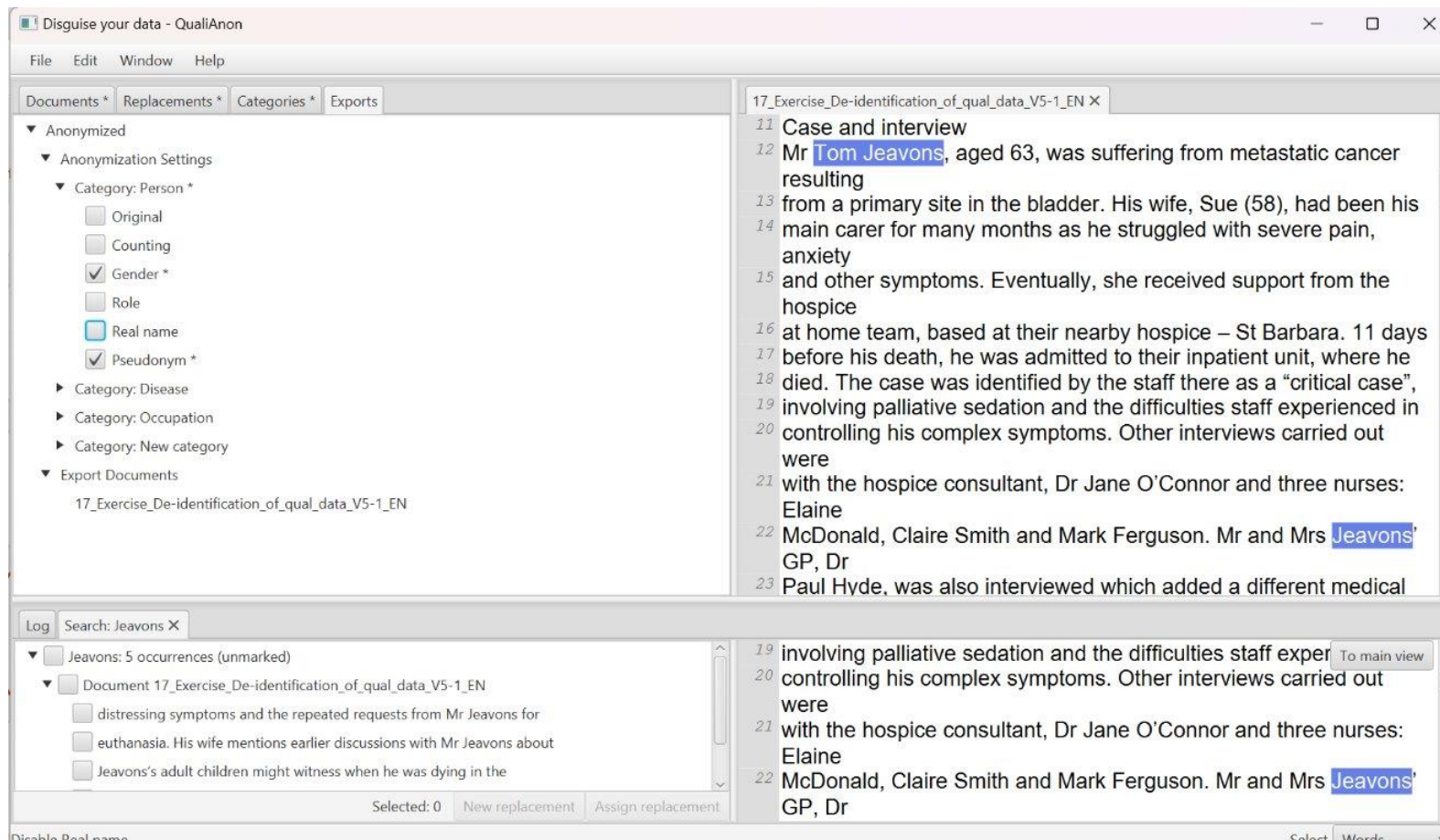
Step 4: manually anonymise data



Step 4: manually anonymise data

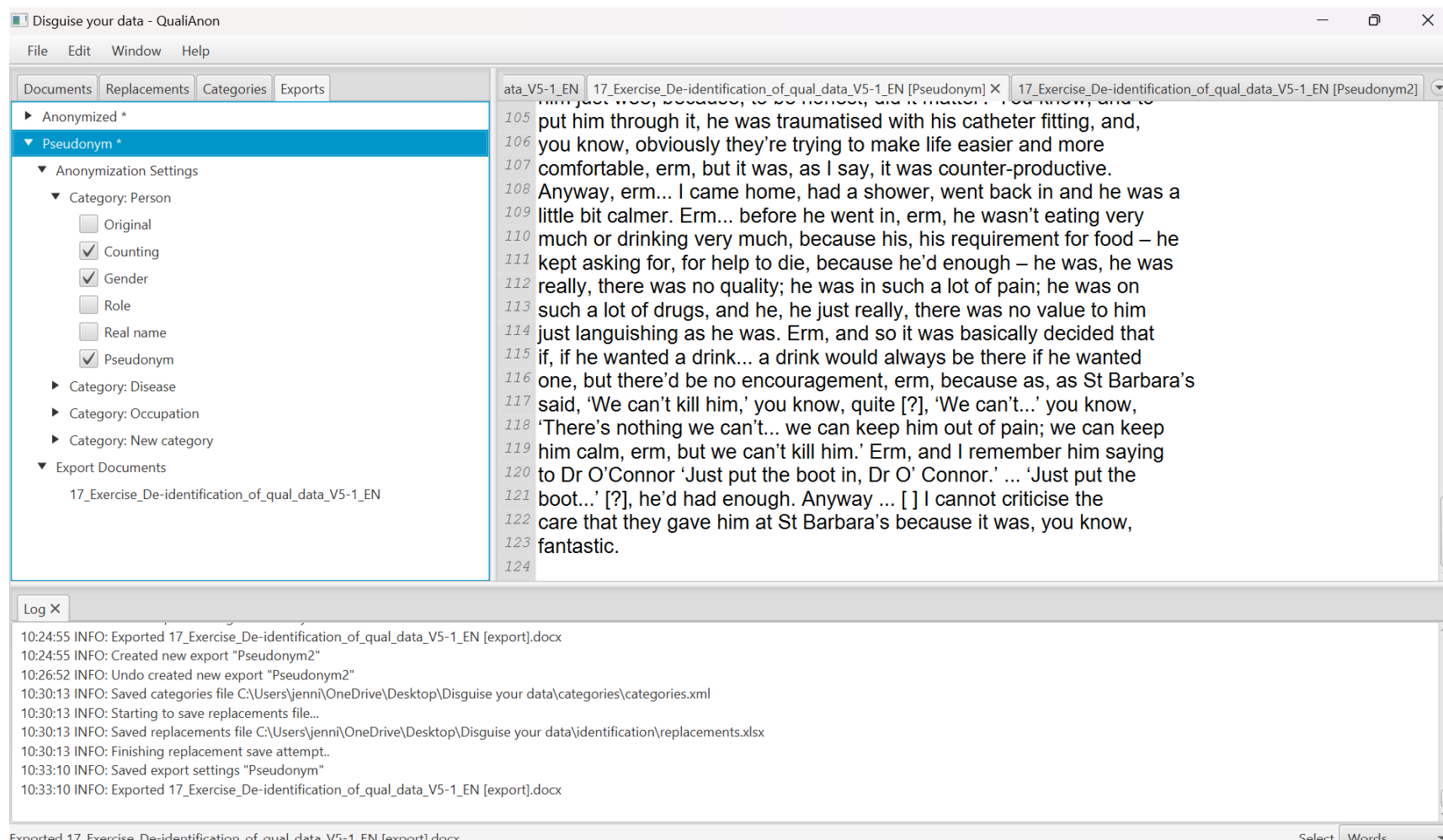


Step 4: manually anonymise data



QualiAnon

Step 5: export



Step 5: export

(11) Case and interview

(12) Mr [Person 1 | Gender: Male | Pseudonym: Stefan], aged 63, was suffering from metastatic cancer resulting

(13) from a primary site in the bladder. His wife, Sue (58), had been his

(14) main carer for many months as he struggled with severe pain, anxiety

(15) and other symptoms. Eventually, she received support from the hospice

(16) at home team, based at their nearby hospice – St Barbara. 11 days

(17) before his death, he was admitted to their inpatient unit, where he

(18) died. The case was identified by the staff there as a “critical case”,















(19) involving palliative sedation and the difficulties staff experienced in

(20) controlling his complex symptoms. Other interviews carried out were

(21) with the hospice consultant, Dr Jane O’Connor and three nurses: Elaine

(22) McDonald, Claire Smith and Mark Ferguson. Mr and Mrs [Person 2 | Gender: Male | Pseudonym: Stefan]’ GP, Dr

Step 5: export - folders

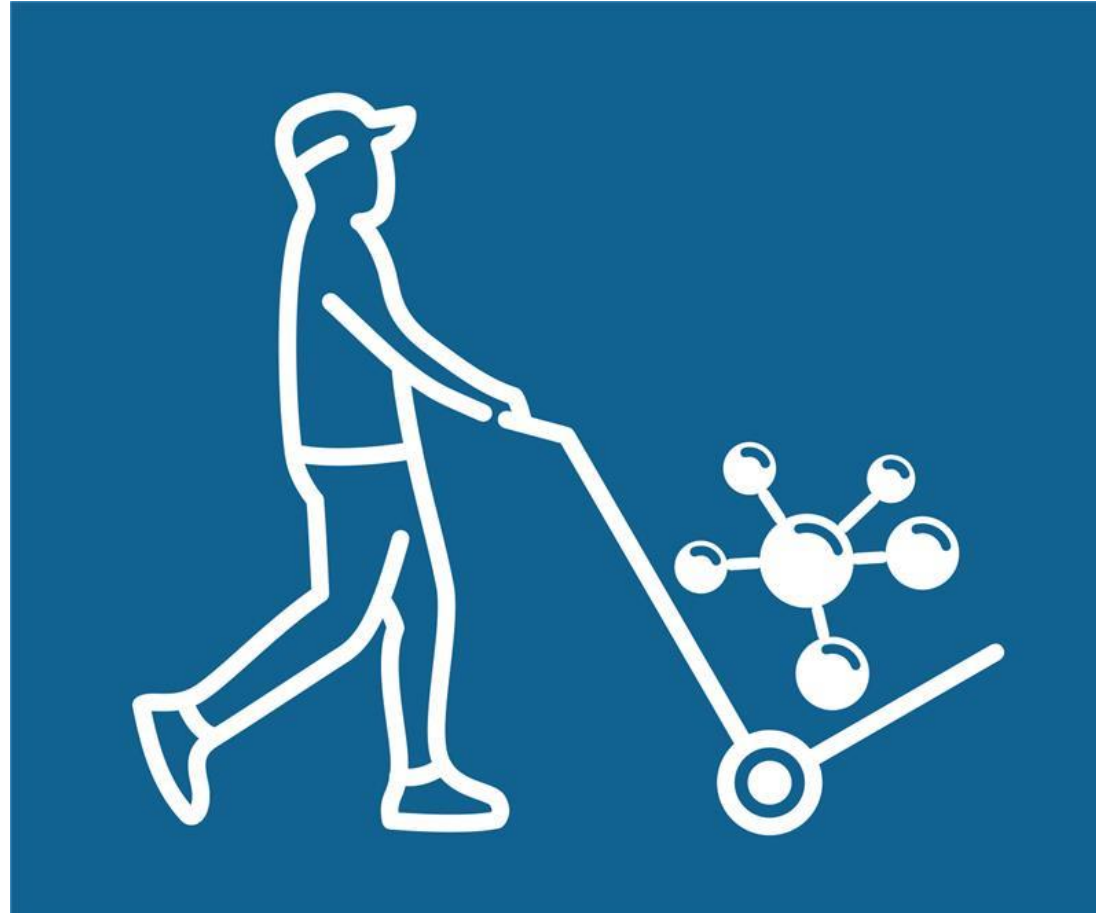
 anonymized		29.04.2025 10:21	Dateiordner	
 backup		29.04.2025 10:30	Dateiordner	
 categories		29.04.2025 10:30	Dateiordner	
 exports		29.04.2025 10:26	Dateiordner	
 identification		29.04.2025 10:30	Dateiordner	
 trash		29.04.2025 10:26	Dateiordner	
 auditfile		29.04.2025 10:33	Textdokument	69 KB

u^b Take home messages

- Choose the right tool for your data type
- Invest time in Learning
- Understand the Trade-Off
- Conduct a Risk Assessment (see [FADP Art 5 c.](#): sensitive personal data)

u^b

Questions?



Save the date - Research Data Day

Topic: Protect your data

When: 18 November, 9:15 – 15:00

Where: UniS

What: Input lecture and workshops on secure data handling

Target group: Researchers from all disciplines

Registration: Opening soon

Service and Support

Website: [Research Data Management Support](#)

Newsletter: <http://www.unibe.ch/ub/osnews>

E-Mail: researchdata@unibe.ch

u^b

Thank you!

Christine Krebs

E-Mail: christine.r.krebs@unibe.ch

Data Steward for Human Sciences



Jennifer Morger

E-Mail: jennifer.morger@unibe.ch

Data Steward for Business, Economics,
Social Science

