cessda eric
Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

Research Data Management

International Summer School in Uganda

Dr. Anja Perry

Oliver Watteler

gesis Leibniz Institute for the Social Sciences

Image: pixabay (CC-0)

cessda eric
Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

Composing the dataset structure

Dataset structure

Objective 1: Meaningful definition of variables

- Keeps the question/variable relationship
- Easily recognize & distinguish variables while working with the data

Objective 2: Compose a systematic dataset structure

- Easily follow question sequences and different question/variable groups
- Simplify and facilitate syntax processing

Versioning

To uniquely identify a specific dataset processed

Version of the dataset

- as part of the file name
- as part of the data, i.e. as a variable in the data matrix

Processing a versioning variable

- based on e.g. the date of processing or an international standard,

  ⇒ decide on versioning standard at the beginning

- use string variables not numerical one to avoid accidental re-coding

  ⇒ do not include special characters (punctuation)

Excursus: The DDI Versioning-Standard

DDI defines a standard of versioning datasets

major.minor.revision

1.0.0

- major position (starts with 1):
  major change, e.g. adding a new variable
- minor position: (starts with 0):
  relevant changes, e.g. re-coding a variable
- revision position (starts with 0):
  smaller changes, e.g. correcting typos in labels

ID-Variable(s)

- At least one unambiguous identifier for each single observation
- More than one identifier, e.g. for different contexts like
  - e.g. time points of measurements
  - e.g. geographical units like countries
- Multiple identifiers in a dataset can be combined
  - identifiers can include various information
  - care about re-identification

Coding of ID variable

- Use string variables not numerical one to avoid accidental re-coding

  ⇒ do not include special characters (punctuation)
- Each identifier should
  - be composed similarly (consistency)
  - have the same length, i.e. number of characters
- Place identifier variables at the beginning of dataset, i.e. as initial variable(s)

Variable Relationship

Simple Question - Variable relationship

- **Question:** Gender (m/f)
- **Variable:** D2002 (1/2)

Variable definitions depending on the question construct

- **Question:** What do you think about each of our political parties? Please rate it on a scale from 0 to 10, where 0 means strongly dislike and 10 means strongly like. The first party is [PARTY A]. Using the same scale, where would you place [PARTY B]?
- **Variables:** D3011_A (LIKE-DISLIKE PARTY A), D3011_B (LIKE-DISLIKE PARTY B) (both on scale 0 to 10)

Standardized demographic variables

- **Question:** Household Income
- **Requirement to consider:** unit of measurement / standard code (here: sample quintiles)
- **Variable:** D2020 (Values from '1. lowest household income quintile' to '5. Highest household income quintile')

Conventions for Variable Names, Values & Labels

Variable names: 4 options

- Names ascending numbered — V1, V2 …
- Names with question or item number — V1q1, V2q2a, V3q2b, …
- Mnemotechnical names — inc, edu, …
- Names with different elements — AUT_PRTY, …

Suggestion 1: Always edit the variable & value labels

Suggestion 2: Content of labels, names and values should be

- Short, meaningful, distinguishable (max. no. of characters is software dependent)
- No special characters, no spaces, rather: "_"

Avoid spaces in variable names

Check this blog post on different ways to avoid spaces:

https://medium.com/better-programming/string-case-styles-camel-pascal-snake-and-kebab-case-981407998841

Images: pixabay (CC-0)

Grouping of variables

Objectives

- Set clear order of variables to ease orientation in the dataset
- Keep its relationship to the questions in the questionnaire

Suggestions

- Group variables into groups according to formal criteria:
  - Administrative variable
  - Demographic variables
  - Question related-/core survey variables

- Integrate derived variables into this data set structure
  - Variable Position: right after or close to the referenced variable(s)

“Missing“ Values

Types of missing values

- Unit non-response
- Item-non-responses
- Missing by design (e.g. trend series, cross-national datasets)

General rules

- Missing values are assigned to each variable special codes ("Missing Values")
- The meaning is explained by a unique label
- Use of the highest numeric code, which is outside the respective valid range of values or apply coding by negative values (to delimit them from the positive valid values)

“Missing“ Values

Example A	**Example B**
If value "7" is part of the valid value range of the variable, "97" is encoded.	Coding with negative values
7 (resp. 97, 997) refused 8 (resp. 98, 998) don’t know 9 (or 99, 999) no answer 0 does not apply	-1 do not know -2 no answer -3 not applicable -4 not asked in survey

„Missing“ Values

Special coding rules when a question was not presented

- E.g. not asked in wave, or country

Category "does not apply" in filter follow-up questions

- In general, filtering condition only from valid value range of filter query
- Typically contains code "0" or a corresponding negative value
- Previously defined missing values in filter question typically set missing in subsequent questions
- When defining filter follow-up questions, the category “not applicable” should also clearly state which previous encoding(s) of the question and category(s) it refers to.

“Missing“ Values

Coding a filter-sequence relationship

Q17 Have you ever been unemployed?	Q18 (if respondent was unemployed): How long were you unemployed?
1 Yes (go to Q18) n = 210 2 No (go to Q19) n = 1060 9 Missing (go to Q19) n = 10	1 under one year n = 150 2 One year and longer n = 50 9 Missing n = 10 0 Not applicable (Q17 code 2 or 9) n = 1070

Image: pixabay (CC-0)

cessda eric
Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

Data Processing

First of all…

- Manifold possibilities for validation and checking data consistency
- Suggestions to support you in creating your own checking routines

- Ideally: source of inconsistency can be traced back
- Enables informed decisions
- Comprehensive documentation of data collection and data processing essential

Image: pixabay (CC-0)

DATA VALIDATION

Unit Non-Response

Research objects (drawn sample)

- refused participation
  ⇒ such cases can be deleted

- break off, i.e. objects refuse to continue during the interviewing
  ⇒ delete? consider whether useful for analysis

Image: pixabay (CC-0)

- Document such deletions!

Representativeness

Is the sample a random sample?

→ check methodological report

Is the data representative?

→ compare sample and population distributions

What to do if the sample distribution does not fit?

- Weighting variables can correct to some degree
- document construction and use of weights

COMMON INCONSISTENCIES AND DATA CONSISTENCY CHECKS

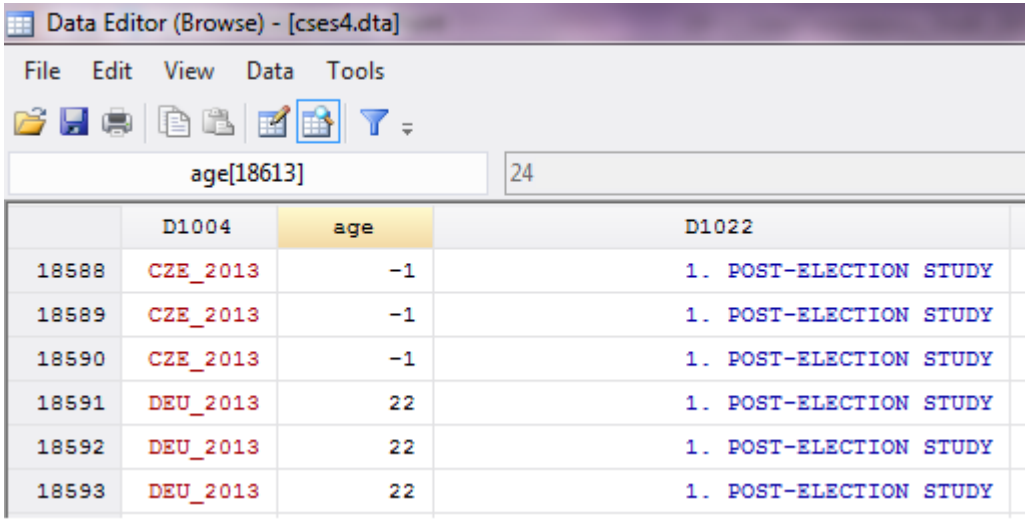
Missing Values for Certain Groups

Possible reasons?

- Processing: overwriting or deleting variables
- Where data for different groups collected seperately: data might not have been collected for some of them
- Certain groups might systematically refuse

Approaches

- Visual check: sort by group ID
- Syntax check: create list of groups with only missing values



	D1004	age	D1022
18588	CZE_2013	-1	1. POST-ELECTION STUDY
18589	CZE_2013	-1	1. POST-ELECTION STUDY
18590	CZE_2013	-1	1. POST-ELECTION STUDY
18591	DEU_2013	22	1. POST-ELECTION STUDY
18592	DEU_2013	22	1. POST-ELECTION STUDY
18593	DEU_2013	22	1. POST-ELECTION STUDY

Wild Codes and Unlabeled Values

Possible reasons?

- Likely processing error during data entry, recoding, labeling

Approaches

- Visual checks
  - Manual comparison of data & codebook/questionnaire
  - Apply value labels & check frequency distributions
- Syntax checks
  - List out of range values per variable
  - For string variables: syntax checking correct length of values

Consistent Application of Filters

Possible reasons?

- Filter instructions disregarded or applied incorrectly

Approaches

- Visual check: inspection of cross tabulation
- Syntax check: list observations taking on non-missing values for follow-up variables even though value of filter variable indicates follow-up should have not been asked

	SPOUSE: CURRENT EMPLOYMENT STATUS				
MARITAL STATUS	01. EMPLO	02. EMPLO	04. HELPI	99. MISSI	Total
1. MARRIED OR LIVING	13,700	1,789	295	9,710	25,494
2. WIDOWED	105	7	10	4,451	4,573
3. DIVORCED OR SEPARA	154	13	0	4,120	4,287
4. SINGLE, NEVER MARR	308	21	3	11,802	12,134
7. REFUSED	5	1	0	141	147
8. DON'T KNOW	5	0	0	36	41
9. MISSING	32	4	3	1,987	2,026
Total	14,309	1,835	311	32,247	48,702

Contradictory Answers

Possible reasons?

- Respondents misunderstood survey question or misreport for other reasons (social desirability etc.)
- Mistake in questionnaire, e.g. inconsistency across questionnaires of waves

Approaches

- Visual check: inspection of cross tabulation
- Syntax check: list observations taking values for certain variables that should not appear if specific response for another variable is given.

D21 NUMBER OF PERSONS IN HOUSEHOLD	D22 NUMBER OF PERSONS IN HOUSEHOLD UNDER THE AGE OF 18				
	1	2	3	4	Total
1	42	4	0	1	47
2	0	131	1	1	133
3	0	0	36	1	37
4	0	0	0	17	17
Total	42	135	37	20	234

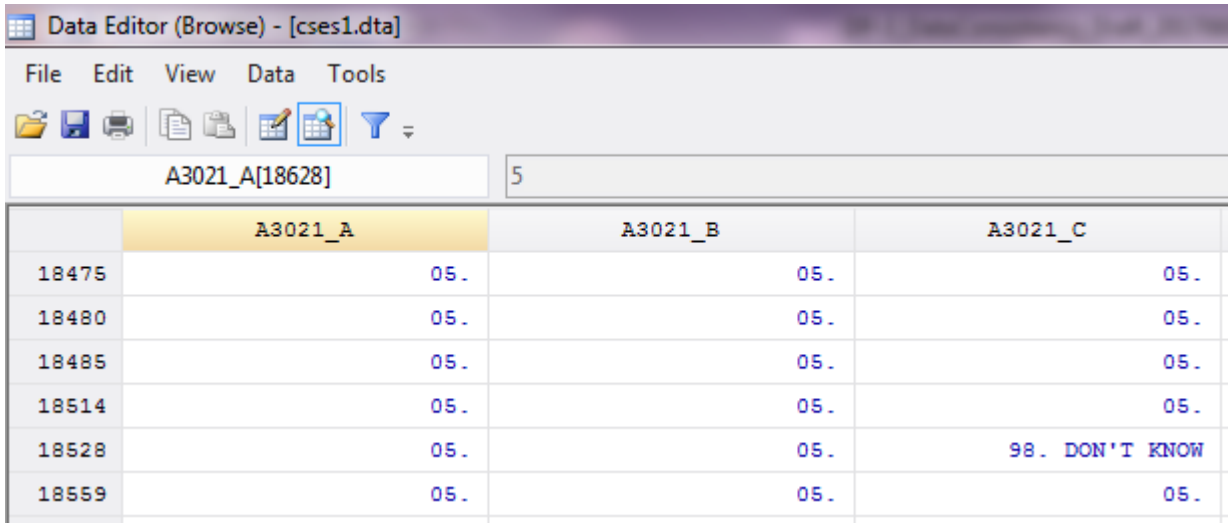
Systematic Answer Behavior

Possible reasons?

- Satisficing

Approaches

- Visual check: inspection of values across battery items per observation
- Syntax check: list observations for which answers are systematic



	A3021_A	A3021_B	A3021_C
18475	05.	05.	05.
18480	05.	05.	05.
18485	05.	05.	05.
18514	05.	05.	05.
18528	05.	05.	98. DON'T KNOW
18559	05.	05.	05.

Implausible Correlations

Possible reasons?

- Scales might have been reversed in questionnaire, during data entry or processing

Approaches

- Visual check: inspection of cross tabulation
- Syntax check: correlations with clear expectation about directionality

Check of Weighting Variables

Image: pixabay (CC-0)

- List system missing and zero values in weighting variables
- Inspect distribution and summary statistics of weighting variables
- Mean should usually equal 1

Duplicate cases

- List all duplicates in terms of ID variables

→ `Duplicates report`

Image: pixabay (CC-0)

HOW TO DEAL WITH DETECTED INCONSISTENCIES

Strategies

Definitely

Attempt to trace source with help of documentation (questionnaire, fieldwork report, methodological report, ideally: processing syntax)

Possibly

Contact data collection agency and/or subject matter experts

Considerations

- Are we certain this is an error?
- Are we able to trace back source of error?
- Can we correct values be inferred?
- Number of cases affected?

Image: pixabay (CC-0)

Ways of handling inconsistencies

Correct inconsistent values

- Certainty about error, source determined, correct values inferred

Set inconsistent values to missing

- Certainty about error, but not enough information for correction

Document inconsistent values (codebook or flag variable)

- Recommended in case of uncertainty
- Analyst can decide on basis of background info provided

Set up project rules for consistent handling & document them!

Example for highlighting inconsistencies

```
| There are some instances in which the number of persons
| in household is equal to or less than the number of persons
| under age 18. These data remained unchanged.
     ...
    |                          EQUAL         LESS
    |
    | AUSTRIA (2008)             17            1
    | BELARUS (2008)              2            0
    | CANADA (2008)               2            0
    | CZECH REPUBLIC (2010)       0            1
    ...
```

Example taken from CSES III: http://www.cses.org/.

Image: pixabay (CC-0)

Image: pixabay (CC-0)

cessda eric
Consortium of European Social Science Data Archives
European Research Infrastructure Consortium

Anonymisation

Anonymization

- Social science is concerned with personal data
- Need strategy to protect the identity of participants
  - Legal requirement by EU law
  - Ethical reasons to protect participants from harm and commercial interests

Image: pixabay (CC-0)

The legal requirement to anonymize

picture: pixabay (CC-0)

„processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information … to ensure that the personal data are not attributed to an identified or identifiable natural person“ (Art. 4(5) GDPR)

Two kinds of identifiers

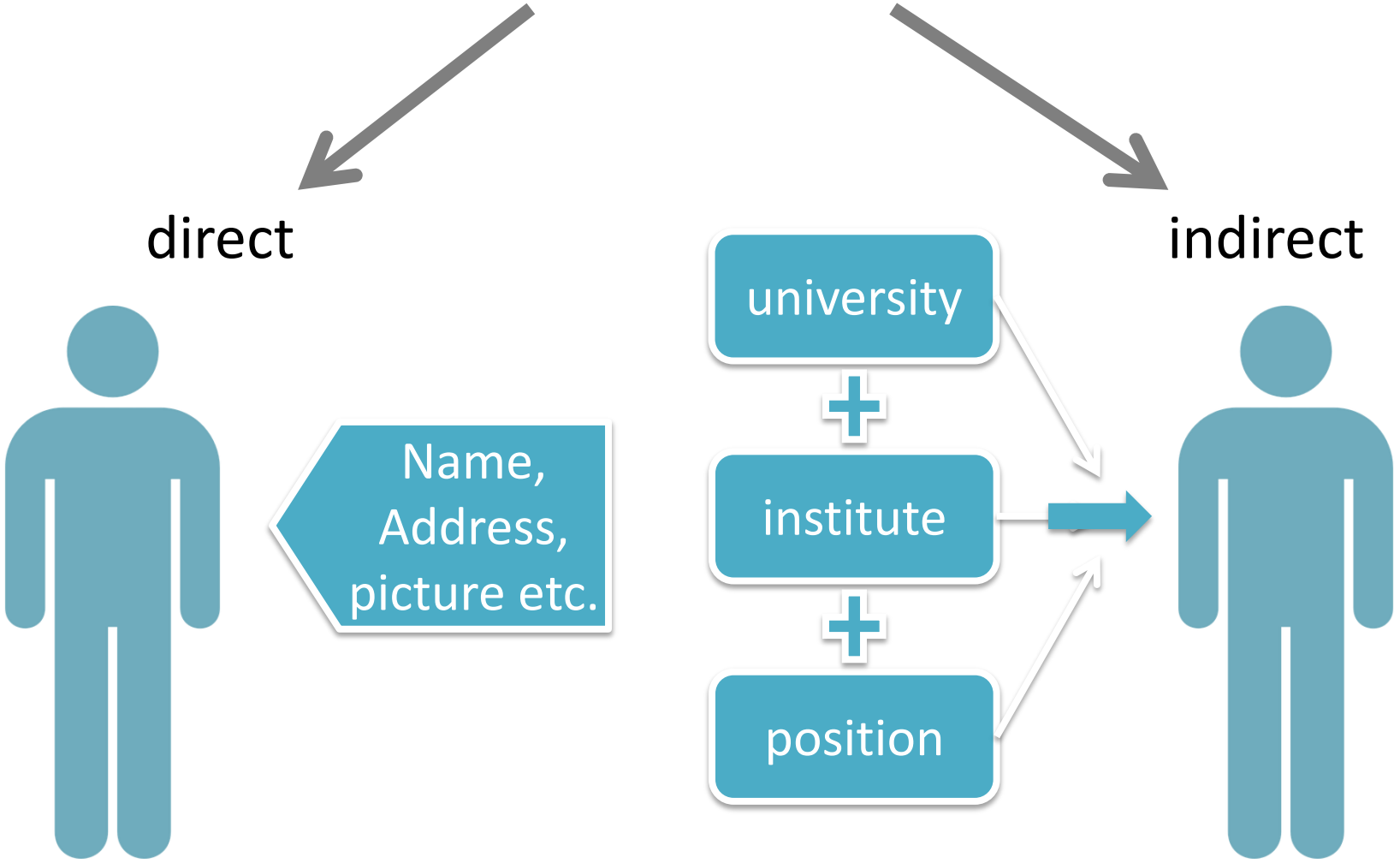


Image: pixabay (CC-0)

Anonymization is an early task

- **Plan before collecting data**
  - saves resource
  - enables a consistent anonymization process
  - yields better informed consent
- **Think about the data to collect**
  - take care of data protection laws
  - type of data affects anonymization strategy
  - discuss with your archive about keeping sensitive data on separate files

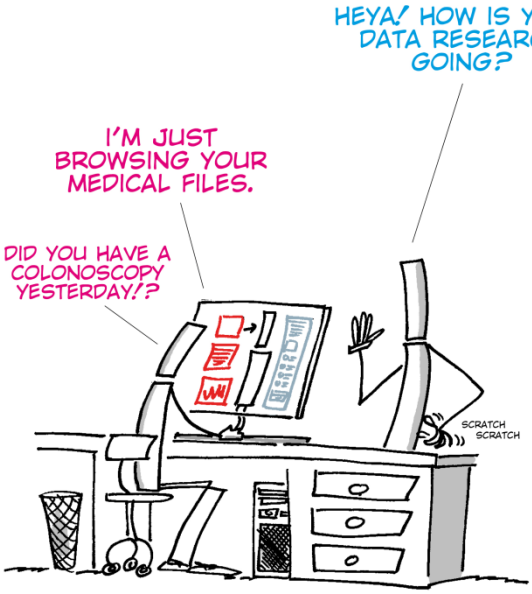
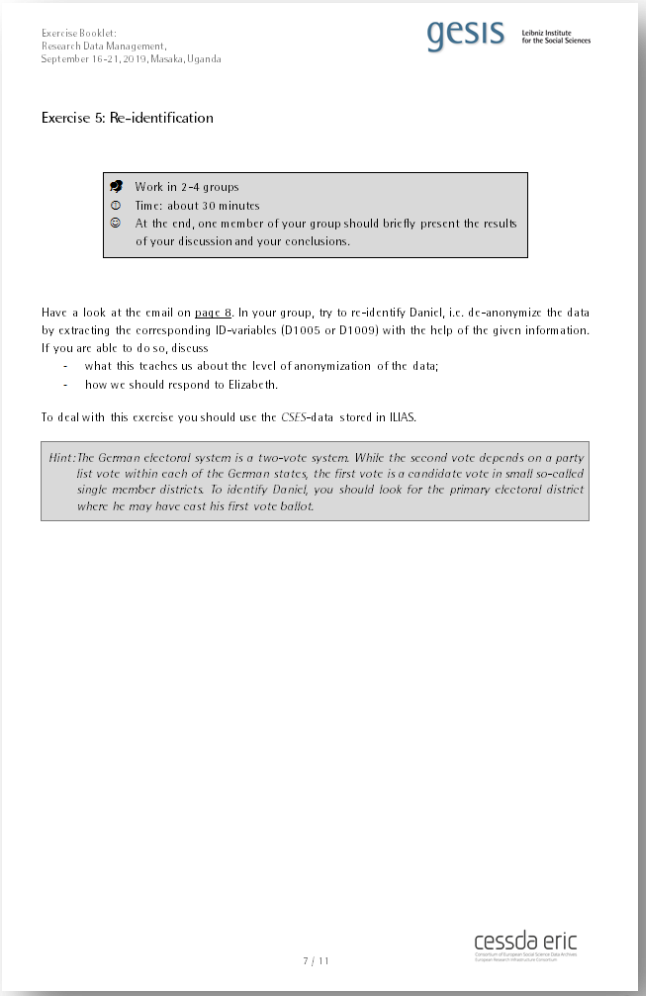


Image: P. Hochstenbach (CC-BY)

Exercise: Re-identification

- work in 5-6 groups
- time: about 30 minutes
- afterwards, we will commonly discuss your suggestions
- see Exercise-Booklet for details on Exercise 5



Exercise Booklet:
Research Data Management,
September 16-21, 2019, Masaka, Uganda

gesis Leibniz Institute for the Social Sciences

Exercise 5: Re-identification

- Work in 2-4 groups
- Time: about 30 minutes
- At the end, one member of your group should briefly present the results of your discussion and your conclusions.

Have a look at the email on page 8. In your group, try to re-identify Daniel, i.e. de-anonymize the data by extracting the corresponding ID-variables (D1005 or D1009) with the help of the given information. If you are able to do so, discuss

- what this teaches us about the level of anonymization of the data;
- how we should respond to Elizabeth.

To deal with this exercise you should use the *CSES*-data stored in ILIAS.

Hint: The German electoral system is a two-vote system. While the second vote depends on a party list vote within each of the German states, the first vote is a candidate vote in small so-called single member districts. To identify Daniel, you should look for the primary electoral district where he may have cast his first vote ballot.

7 / 11

cessda eric

Anonymization strategies

- Keep sensitive / personal data (such as contact information) in separate files
- Remove variables with sensitive data if it doesn't compromise the data
  - should you even need, i.e. measure, such variables?
  - can you keep them for restricted use?
- Use meaningful pseudonyms and replacements for identifiers, e.g.
  - "Masaka" ⇒ "medium-sized city in Uganda"
  - "John" ⇒ "Pupil 1"

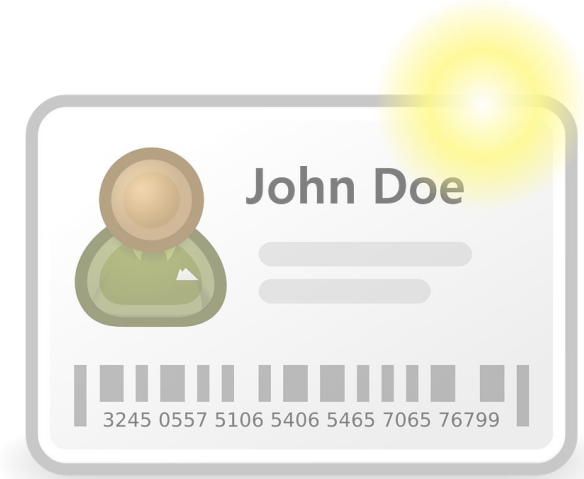


Image: pixabay (CC-0)

Anonymization strategies

- Restrict upper and lower ranges of variables, e.g. on high income earners
- Low-level aggregation of data
  - moving to a larger spatial unit, e.g. aggregate single streets into broader geographical units
  - transforming continuous into discrete variables, e.g. age groups instead of (continuous) age in years
- Document changes undertaken and flag anonymization, e.g. “Except for *Pupil 1*… […] Lots of difficult experiences in his life. *difficult familiar situation* […]”

Anonymizing audio / video files

- Anonymize audio and visual files by digital manipulation, e.g. voice alteration or image blurring
- But, digital manipulation is
  - labour intensive and expensive
  - may compromise data quality
- Better
  - obtain consent to use and share data unaltered
  - avoid collecting disclosing information during audio recordings

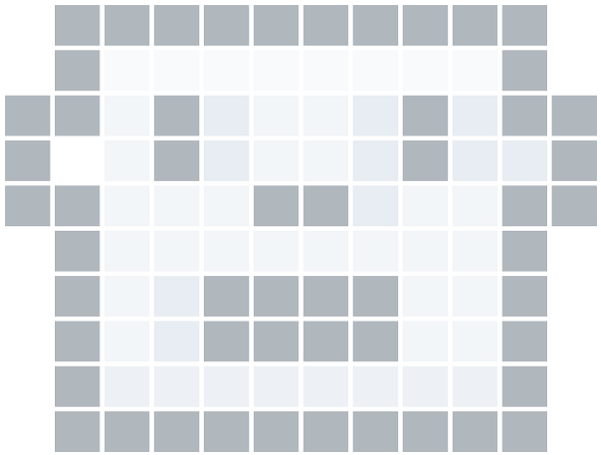


Bild: pixabay (CC-0)

Anonymization: What to avoid

- Over-anonymising text, which can distort data
- Removing information, rather use pseudonyms or replacements
- “Find and Replace”
- Inconsistency within research team and throughout project

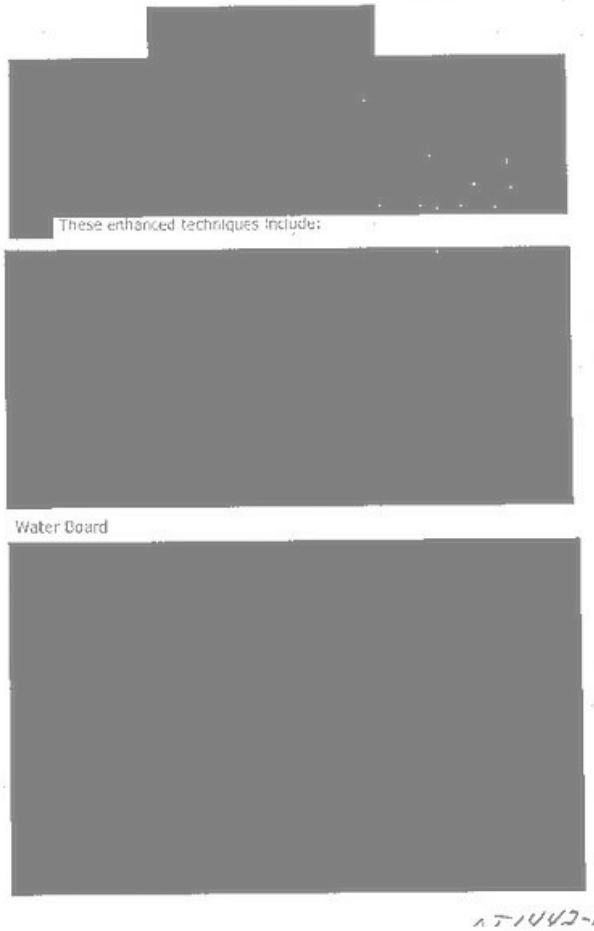


Image: pixabay (CC-0)

What to do if anonymization is impossible

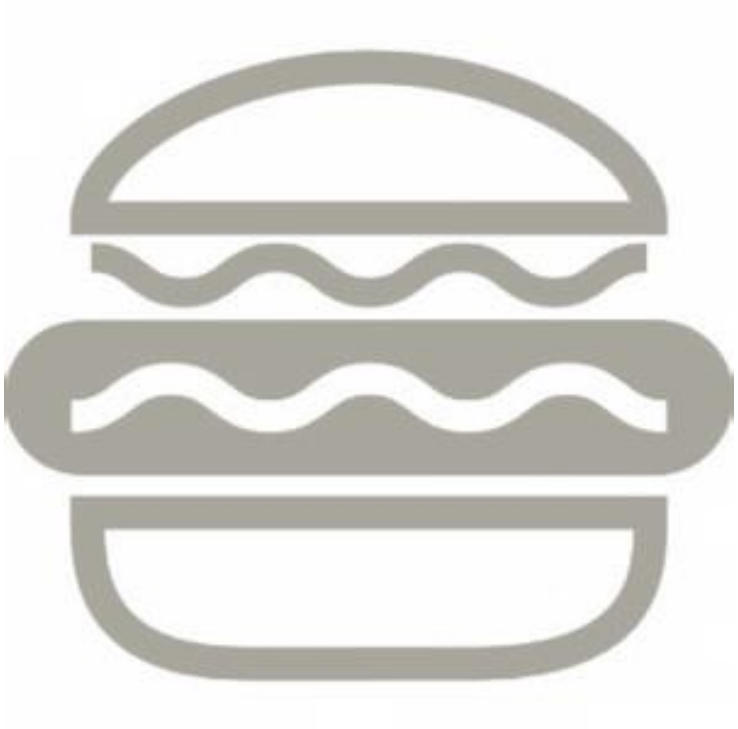
- Obtain informed consent for sharing non-anonymous data
- Control access to your data and regulate their reuse, e.g. through an archive
- Place confidential data under embargo for specified period

Image: pixabay (CC-0)

Image: pixabay (CC-0)

Thank you for your attention!

anja.perry@gesis.org
oliver.watteler@gesis.org



Adapt your DMP

Focus on the documentation of inconsistencies in the data

- how to deal with inconsistencies?
- how to document decision and steps undertaken to deal with inconsistencies?

Your task

Work with the training dataset. Please complete the following tasks:

- Check the dataset for ‘wild codes’. Are there any codes in the dataset that do not seem to have a substantial meaning?
- Check the dataset for ‘missing values’. Are there any missings in the dataset? Beware of codes that look like missing values but belong to valid values.
- Variables Q55A through Q55F belong to three pairs of questions. The first question of each pair holds a filter. Check if the filters were used correctly. Please consult the data codebook for question wording and interviewer instructions.
- Variable Q29A is about the meaning of ‘democracy’. It is the starting variable to a range of other variables concerning this issue. Please check Q29A and the following variable and see if you find anything that strikes you in the answering behavior of respondents. This task is rather philosophical or logical. There are no errors here but an important point for discussion.