Anonymisation for data sharing in practice

Train the Trainers Workshop

Vedran Halamić and Marijana Glavica, CROSSDA
Jan Dalsten Sørensen and Mads Thilsing-Engholm, DNA

04 October 2022

cessda.eu   @CESSDA_Data

DOI: 10.5281/zenodo.7139099

Anonymisation of Data for Reuse

at the Danish National Archives

Jan Dalsten Sørensen and Mads Thilsing-Engholm

cessda.eu

@CESSDA_Data

The Danish National Archives

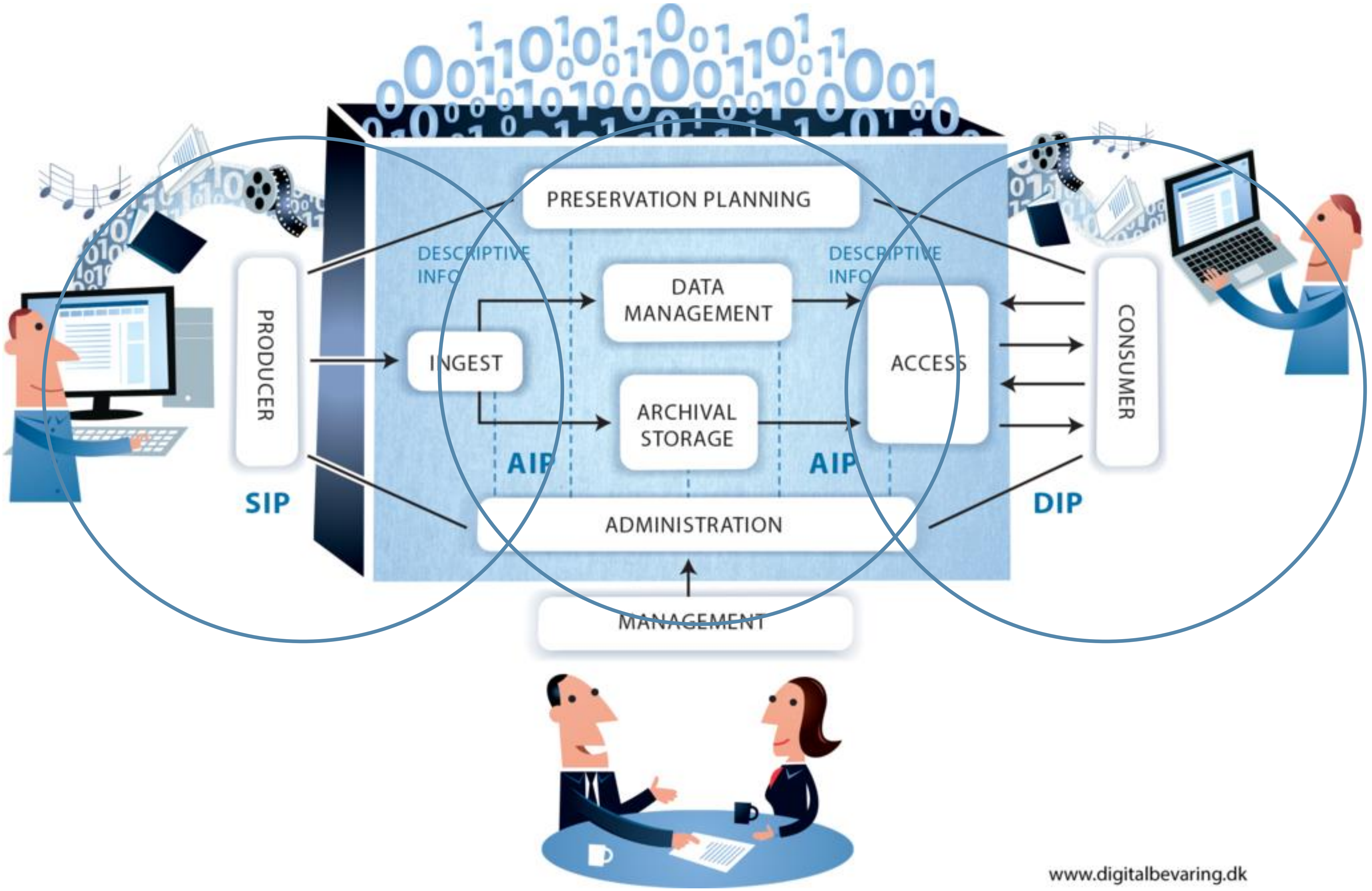
The DNA is the National Archives of Denmark for

- All state authorities and institutions
- All 5 regions (primarily health care)
- Approximately half of the municipalities (about 48 out of 98 municipalities)
- Publically funded research (e.g. universitites)
- Private records' creators, including research who donate their data voluntarily

Our purpose, as stated in the Archives' Act:

- To preserve records of historical value or that serve as important administrative or legal documentation
- To make sure that records creators can dispose of records that are not worthy of preservation
- To make records accessible for citizens authorities, including for research
- To guide users in their use of records
- To perform research and disseminate the results of the research

The OAIS Model



Legislation and other rules

- The Archives' Act
- GDPR
- Possible specific access restrictions set by the donor

GDPR

- Article 89:

1. Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.

GDPR

- Article 89, ctd.

2. Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.

GDPR

- Unless you fully anonymize data, i.e. destroy the possibility of reversing the anonymisation process, you still need to take GDPR into account, e.g. by ensuring that you have a legal basis for the processing of personal data, and that you have appropriate safeguards in place (technical and organisational measures).

- We can dramatically reduce the risks to the rights and freedoms of the data subjects by anonymizing/pseudonymizing, but only full anonymization (destruction of personal information) makes GDPR inapplicable in this context.

The data in question

What kind of data does the National Archives provide to researchers?

- Primarily quantitative data
  - Gathered for the purpose of:
    - Scientific research
      - Primarily survey data, especially (but not exclusively) health & social sciences.
      - Approx. 3000 data units
    - Public administration at state and local level
      - Public records, registers etc.
      - Approx. 6000 data units
      - Many digital documents, of course, but they will not be covered by this presentation
- A large percentage contain direct identifiers
- Almost all contain indirect identifiers
- Find it at https://digidata.rigsarkivet.dk/

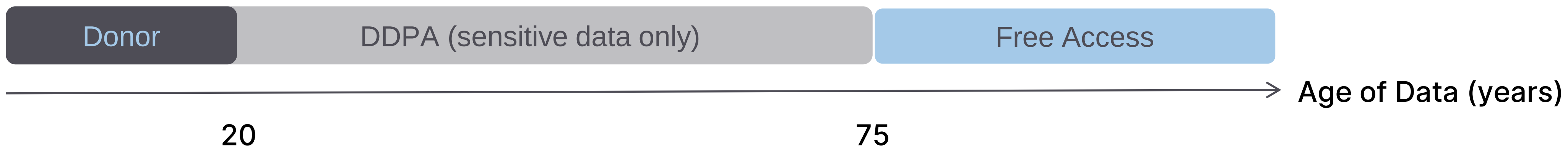
Can you access sensitive data?

- YES, if you have the required permissions

- The rules (The Danish Archival Act):

  - **All Data:**
    20 years: Permission from Data Donor

  - **Sensitive data:**
  - 75 years: Permission from the Danish Data Protection Agency (DDPA)



Donor

DDPA (sensitive data only)

Free Access

Age of Data (years)

20

75

BUT it’s not easy getting the permissions needed

(nor should it be)

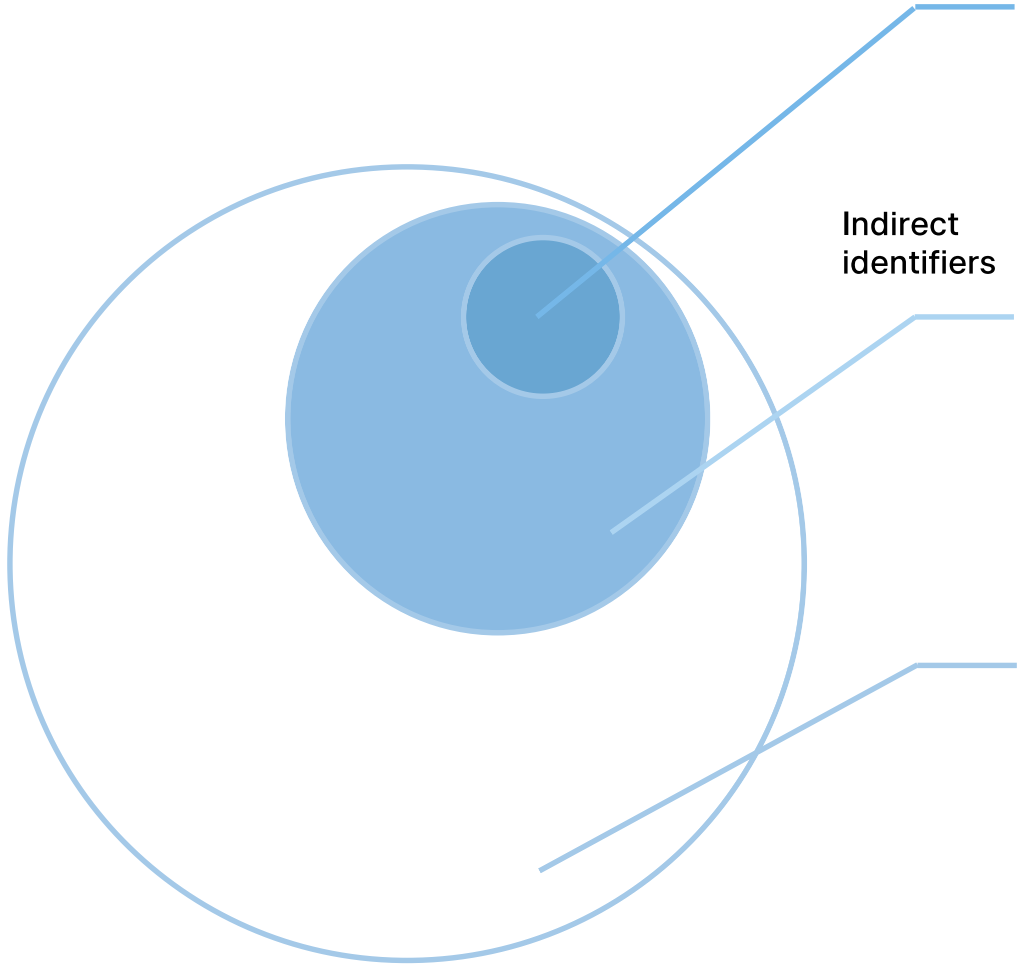
- Only for serious and genuine – primarily scientific – purposes
- Access to the full data – with sensitive data included – is often not necessary for the purpose of analysis

This is where anonymisation comes in to play

- Balancing data security and analytical value
- Keeping ‘the good stuff’ without compromising individuals

Anonymisation

What do we remove from data?



Direct identifiers

Always removed:
- Civil registration number
- Name
- Email, phone number, address
- Sequential number

Indirect identifiers

Sometimes removed:
- Job title
- Municipality
- Date and place of birth
- Nationality
- Religion
- Postal code
- Education
- String (free text)
- etc.

Never removed:
- Non identifying variables

Depends on:
- The sensitivity of the research Issue
  - More sensitive Issues → stricter anonymisation
  - Abuse, disease, sexual orientation etc.
- The size of specific observations
  - Small counts → stricter anonymization
- The size and composition of the researched Population
  - Smaller and specific Population → stricter anonymisation
  - Minorities, school children etc.
- The presence of other information in data
  - Does the combination of info enable easy identification?
    → stricter anonymisation
- The age of the data
  - Newer data → stricter anonymisation

‘It depends’ - An example

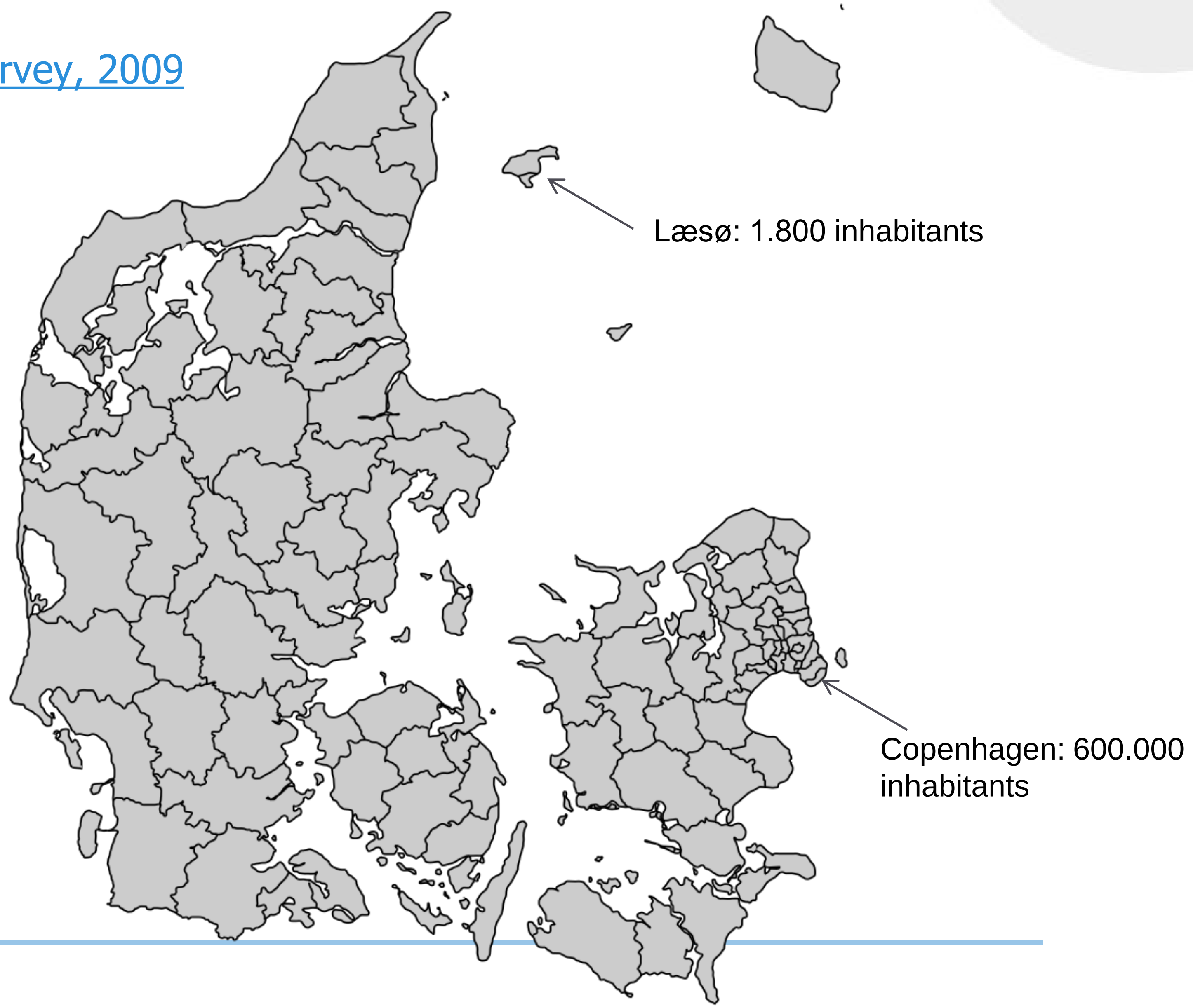
- Normally we don’t remove variables on **gender**, **age** and **region** when anonymising
  - Important variables for analysis
  - In most cases not enough to identify individuals

- BUT we did remove them in the case of this Study:
  Evaluation of the policy governance model in the five regions - Politician survey, 2009

- Why?:
  - There are only a few hundred regional politicians in Denmark (123 responded in the survey)
    - Too easy to identify individuals due to small population (gender + region might be enough)
  - Relatively new Data

- Why not?:
  - The topic was not overly sensitive
  - Loss of analytical value

- On the balance caution prevails



The process

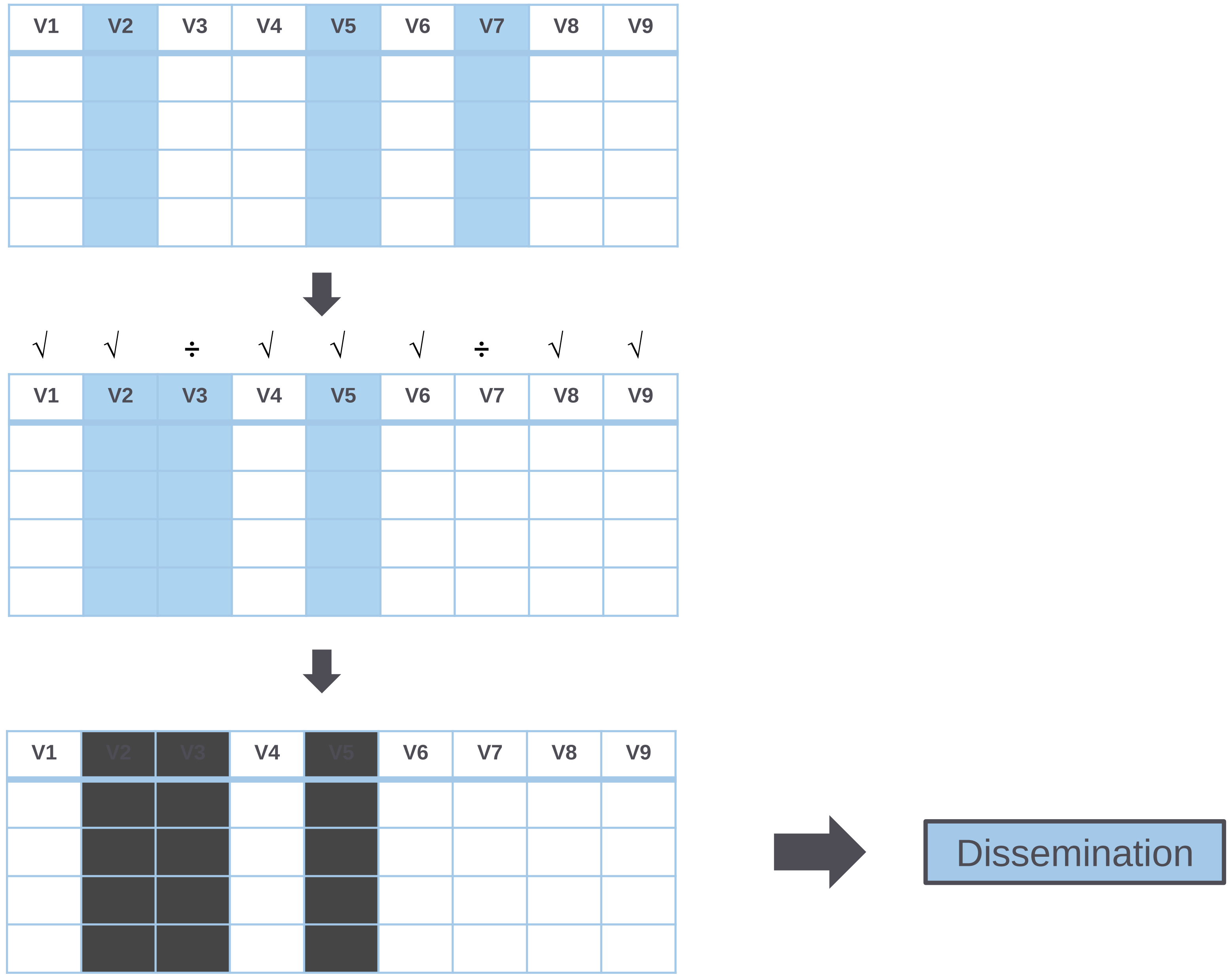
- Anonymisation on demand
    - When an order comes in for a dataset (the first time only)
    - When we make a dataset available for download

- A fairly manual process
    - No business case (yet) for applying tools, since it’s relatively small scale

- Double screening of the data by two archivists
    - Minimising the risk of errors
    - Identifying edge cases
    - Taking into account that there is some judgement involved, too

- Producing a separate anonymized version of the data for dissemination
    - Keeping the original data intact

The process in detail



Archivist 1:
Identifying sensible variables
(in this case V2, V5 and V7)

V1	V2	V3	V4	V5	V6	V7	V8	V9

Archivist 2:
Verifying the first assessment
and flagging necessary changes
(V3 and V7)

√	√	÷	√	√	√	÷	√	√
V1	V2	V3	V4	V5	V6	V7	V8	V9

Archivist 1:
Producing and disseminating the
anonymised dataset

V1	[redacted]	[redacted]	V4	[redacted]	V6	V7	V8	V9
	[redacted]	[redacted]		[redacted]				
	[redacted]	[redacted]		[redacted]				
	[redacted]	[redacted]		[redacted]				
	[redacted]	[redacted]		[redacted]				

Administrative Data

- Not collected for scientific purposes

- But great potential for scientific use
  - Examples:
    - The Conscription Examination Database.
    - Municipal Care Systems
    - Ship logs (paper records) → climate change models

- Principle: Data minimisation
  - Only the necessary information is given
  - Often only very specific information requested

Administrative data, example

- Example: The Game yield register
    - How many deer, pheasants etc. was shot in a given year/month at a given region?
    - Can give insight into development/decline of wildlife in Denmark
    - The first version was anonymous, BUT the next version also contains the **personal identification number** of the Hunter
      (Insert big red flag here)
    - Removing this does in most cases not reduce the analytical value of the data

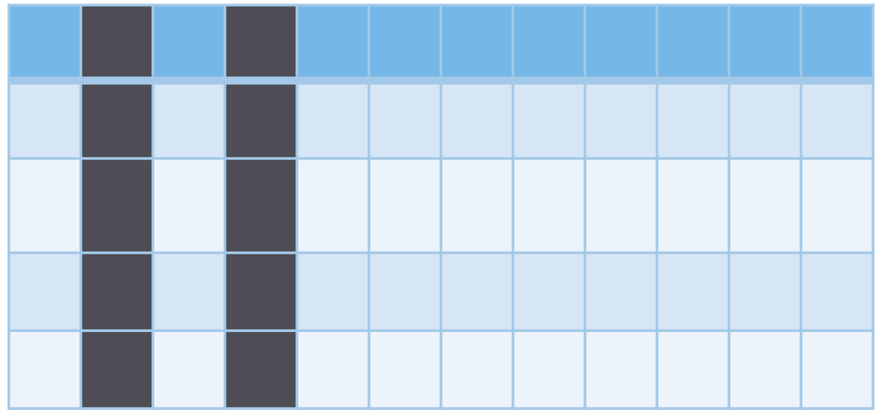
Process for administrative data

- Talk to the user in order to identify their needs

Different data, different approaches

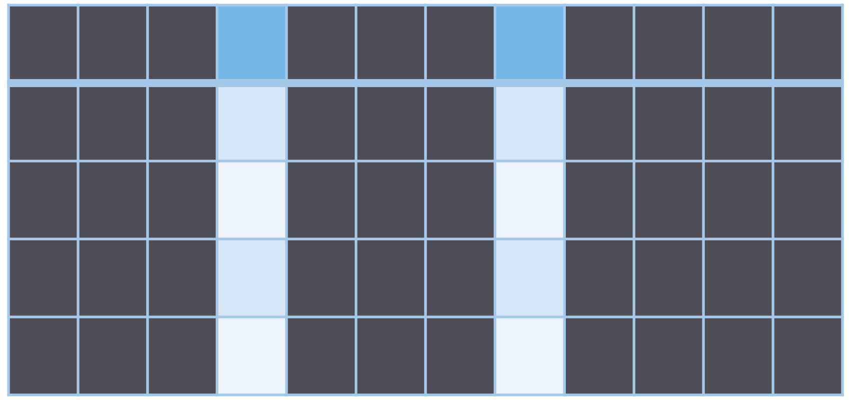
As a general rule (does not always apply) we take different approaches to survey data and administrative data

Survey data:



As much of the data as possible is disseminated

Administrative data:



Only the specific necessary information from data is disseminated

Structured metadata is our friend

- The Data in the Danish National Archives is documented with extensive and structured metadata
    - On Unit level:
        - What is the Data about?
        - Who provided it?
        - What time period?
        - Topics and content
    - On Data Table level:
        - Descriptions of each variable and value in the dataset
- Making it easier to:
    - Identify relevant data
    - Using the data
    - AND anonymising the data

A few pieces of advice

1. Try to get as adequate and complete metadata from the donor at the time of submission, and keep the need for anonymisation in mind. The better the metadata, the better are your possibilities to understand data and pick out the relevant variables.
2. Always remember to pay close attention to metadata before anonymising. Get to know the data before determining how to proceed.
3. Don’t trust your memory, and remember that data sets can change from one submission to the next.
4. Know your legislation. In our case, both the Archives Act and the GDPR. Do you know what legislation pertains to your work?
5. Two are stronger than one. Have, if possible, two archivists look at the data before determining how to anonymize.
6. Save, if possible and relevant, the anonymized data sets. They often come in handy afterwards.

Questions

Train the Trainer Workshop: Basic Concepts of Anonymisation

Welcome

Vedran Halamić / CROSSDA

04 October 2022

cessda.eu @CESSDA_Data

Programme

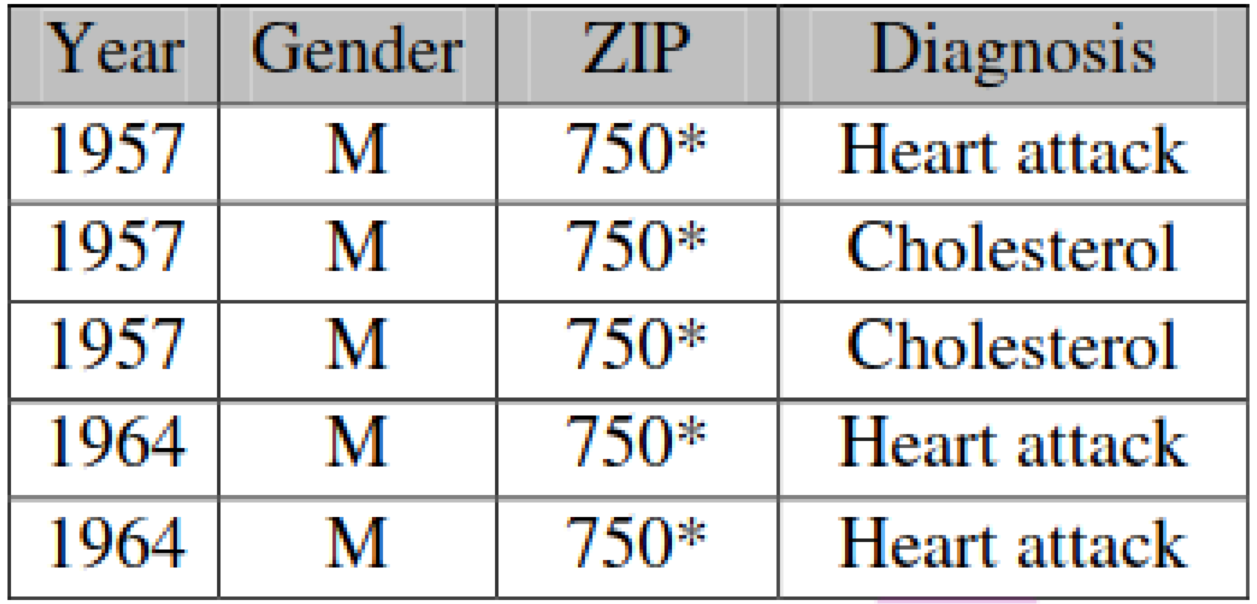
- Basic terminology and logic of anonymisation
- K-anonymity
- Data environment and dataset properties
- sdcMicro

Basic terminology and logic of anonymisation

- “The Anonymisation Decision-Making Framework: European Practitioners’ Guide” (Elliot, Mackey & O’Hara, 2020)
  - **Data situation audit -> Risk analysis and control -> Impact management**
- “A complex process to transform identifiable data into non-identifiable (anonymous) data. This usually requires that identifiers be removed, obscured, aggregated and/or altered in some way. In may also involve restrictions on the data environment.” (Elliot, Mackey & O’Hara, 2020)
- Direct and Indirect identifiers (https://dmeg.cessda.eu/Data-Management-Expert-Guide/5.-Protect/Anonymisation)
- Absolute, formal, statistical or functional anonymisation?

k-anonymity

- Aims to prevent a data subject from being singled out by grouping them with, at least, k other individuals. (EU Opinion 05/2014, 2014)
- Does not prevent any type of inference attack
- l-diversity and t-closeness



Year	Gender	ZIP	Diagnosis
1957	M	750*	Heart attack
1957	M	750*	Cholesterol
1957	M	750*	Cholesterol
1964	M	750*	Heart attack
1964	M	750*	Heart attack

Example from "Opinion 05/2014 on Anonymisation Techniques" by Data protection Working Party (https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf)

Data environment and Dataset properties

- “A context for an item of data” (Elliot et al., 2018)
- Data environment core features : Other data, Agents, Governance processes and Infrastructure (Elliot, Mackey & O’Hara, 2020)
- Environment based solutions for anonymisation
- Dataset properties: data quality, age of data, hierarchical data, longitudinal data

sdcMicro

- Open source R package (https://github.com/sdcTools/sdcMicro)
- Graphical user interface (almost no coding required)
- Includes various risk estimation methods
- Helpful documentation (https://sdcappdocs.readthedocs.io/_/downloads/en/latest/pdf/)

References

- Benschop, T. (2021) sdcMicro GUI manual documentation.
- Benschop, T., Machingauta, C., & Welch, M. (2019). Statistical disclosure control: A practice guide.
- CESSDA Training Team (2017 - 2022). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. Retrieved from https://dmeg.cessda.eu/
- Elliot, M., Mackey, E., & O'Hara, K. (2020). The anonymisation decision-making framework 2nd Edition: European practitioners' guide.
- Elliot, M., Mackey, E., O’Hara, K & Tudor, C. (2016). *The anonymisation decision-making framework*. UKAN Publications.
- Elliot, M., O'hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, *34*(2), 204-221.
- EU Data Protection Working Party (2022, October 3) European Commission. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Thank you

We'll take a quick break and then continue with sdcMicro.

cessda.eu

@CESSDA_Data