

Ethics and Data Aggregation

Louise Bezuidenhout
Leiden University

ENVRI-Hub NEXT
3 September 2026



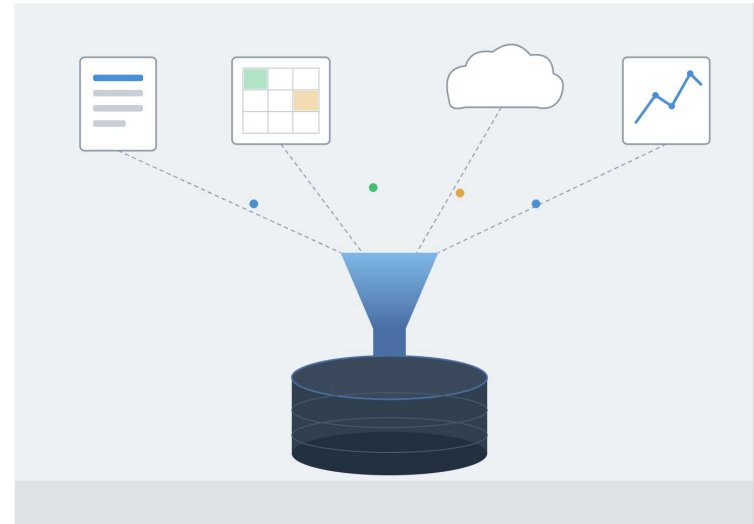
**Universiteit
Leiden**

Data Aggregation

- Data aggregation increasingly common in research
- Enabled by interconnected data infrastructures and research commons
- Provides valuable insights not possible with individual data

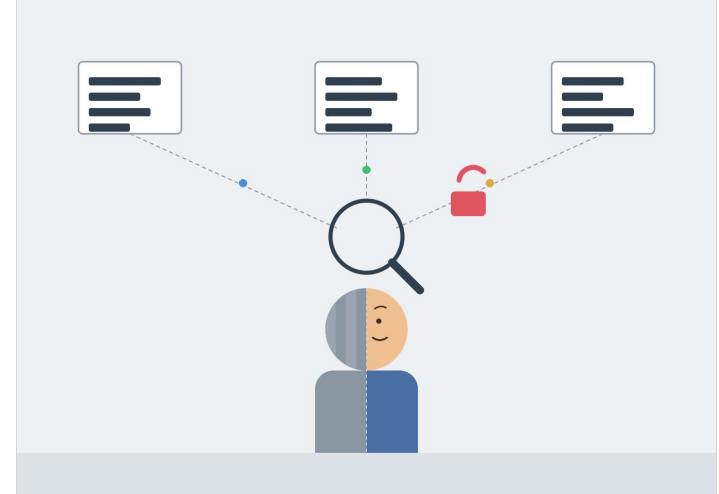
BUT

- Raises ethical issues beyond privacy
- Can create new forms of knowledge and power that not present in the individual datasets
- Makes responsibility and provenance harder to trace



Privacy and Re-identification

- Aggregation does not necessarily anonymise data
- Combining de-identified datasets can make individuals identifiable
- Innocuous variables can become identifying when combined with other information

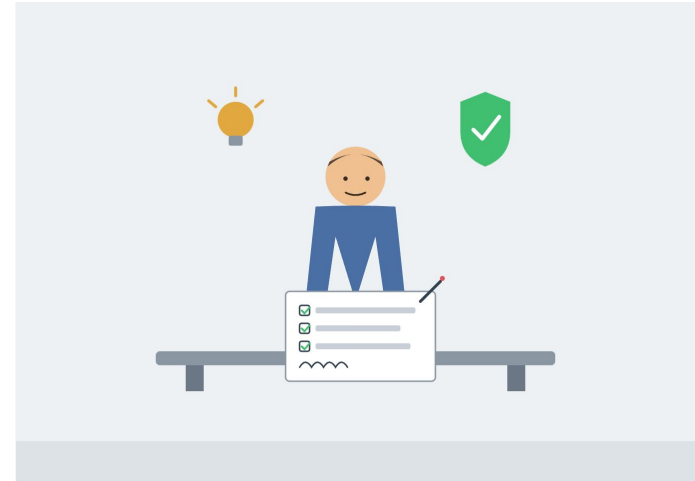


From: *Is this dataset anonymous?*

To: *What can become knowable when this dataset is combined with other datasets over time?*

Function Creep and Secondary Use

- Aggregation enables uses that were not anticipated in original data collection
- Reuse can undermine the expectations under which data were originally provided
 - particularly when stakeholders were not informed about subsequent uses
- Tension between data openness and contextual integrity
 - public value vs consent



From: Is the research valuable for society?

To: Is the research in keeping with the original intentions of the data collected?

Loss of Individual and Community Agency

- Aggregation can undermine meaningful control over use
- Aggregations makes it impossible to determine:
 - who contributed the information
 - who has subsequently used it
 - what inferences have been made
 - whether the original contributor would have agreed to the new use
 - how to withdraw or correct the information
- Particularly significant for community or Indigenous data

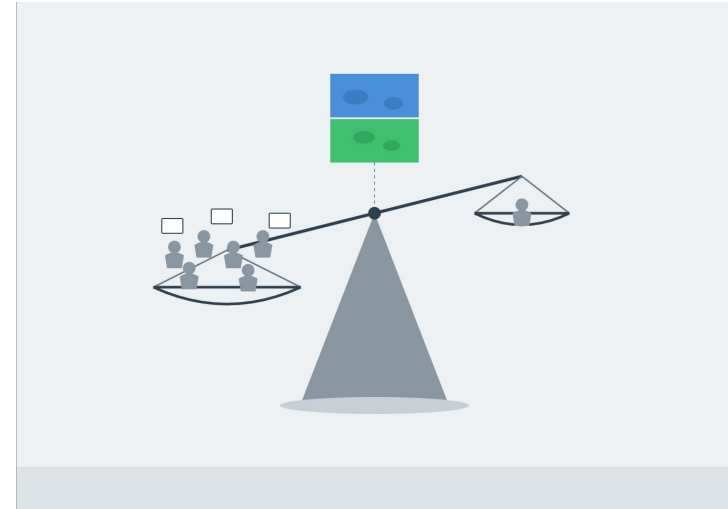


From: Is this dataset useful?

To: Who needs to consent to its reuse?

Bias and Representational Inequality

- Aggregation can make existing biases less visible rather than less significant
- Not only *bias in the data* but bias introduced by decisions about the level at which data are aggregated
- Fairness, human dignity, autonomy, self-determination, and discrimination as ethical considerations in data governance

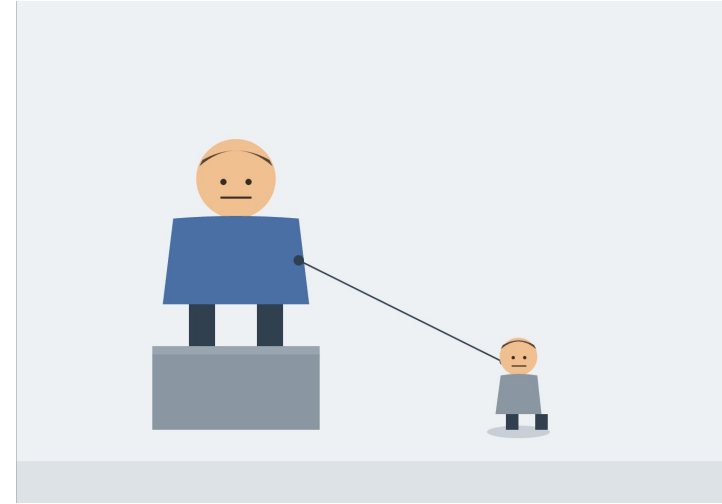


From: *What do we assume from the dataset?*

To: *Who gets represented by the aggregate, and who disappears into it?*

Power Dynamics

- Decisions about data are not value-neutral
- Reflect worldview of those in power
 - Who makes decisions about what categories can legitimately be combined?
 - Bigger is better” problem: interoperability does not automatically produce epistemic comparability
 - Who is subject to the knowledge produced by aggregation?



From: *What categories are useful to analyse the dataset?*

To: *How do these categories constructs a particular representation of the world?*

Systemic Risk

- Aggregation can make data more useful while simultaneously making the infrastructure holding them more vulnerable
- Data sovereignty
- Accountability and provenance, metadata, documentation, and governance are ethical as well as technical concerns



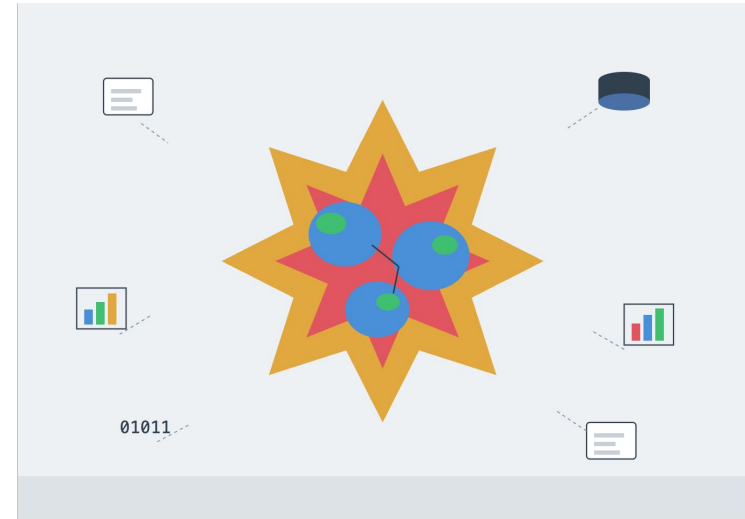
Data Rescue Project

From: Are the data available?

To: How are the data vulnerable to geopolitics and malicious attack?

Sustainability

- Data-infrastructure sustainability - What data should be aggregated, at what resolution, for what purpose, and for how long?
- Considerable environmental impact

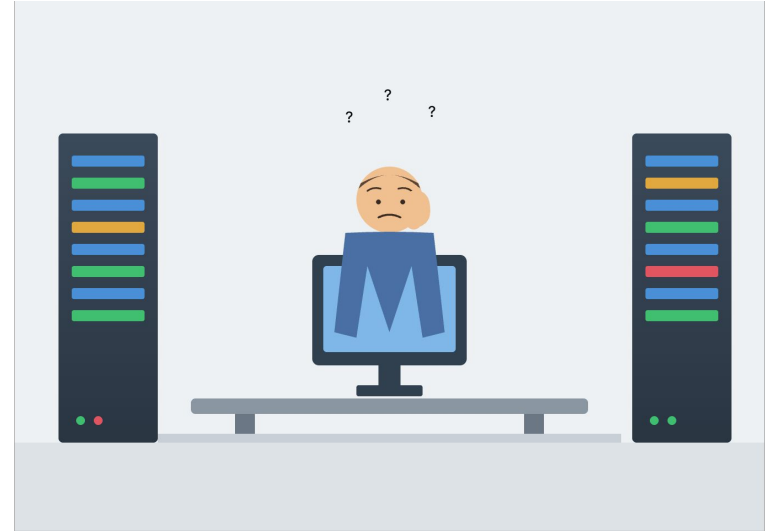


From: All data have potential reuse value and should be saved?

To: How can strategic data management reduce environmental impact?

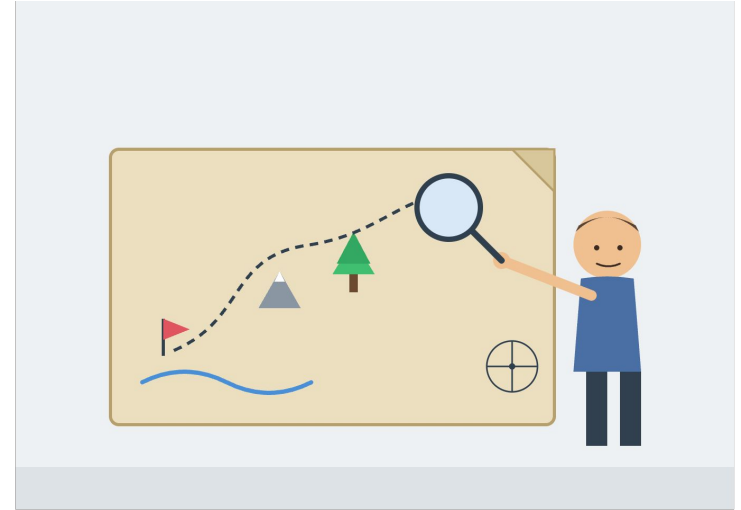
Ethics and Data Aggregation

- Work in progress
 - Cultural lag
 - Changing policy landscape
- Disconnect between RECs and data stewards
- Speculative element difficult to manage
- Can't put responsibility on solely researcher



Starting to be Ethical About Aggregation

- Know that datasets from different disciplines have different requirements
- Understanding personal worldviews and assumptions is vital
- Being critical of data infrastructures is important
- Link your REC to your data stewards
- Make use of tools that can enhance data provenance and trace reuse



Thank You

Thank you for your attention!

Please feel free to contact me at l.m.bezuidenhout@cwts.leidenuniv.nl

You can cite this presentation using the DOI [10.5281/zenodo.22275455](https://doi.org/10.5281/zenodo.22275455).



Universiteit
Leiden

Presentation images
created using AI

