The other way around: data reuse

Workshop FAIR data and data reuse for ESG researchers – Module 6

October 18, 2022, by Cindy Quik and Luc Steinbuch

WAGENINGEN
UNIVERSITY & RESEARCH

How can we reuse existing data?

F   Findable   🔍   Where do we find it?

A   Accessible   🚪   How do we obtain access to it?

I   Interoperable   🔌   Can we read the file types?

R   Reusable   ∞   1) Are we allowed to reuse it?
2) Do we understand what the data means?

More background:

Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018–160019. https://doi.org/10.1038/sdata.2016.18
http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/

Findable + Accessible

- Where to find existing data?

  - Data tables in manuscripts/reports

  - Data availability statements of manuscripts/reports

- Where to find **more** existing data?

Findable + Accessible

Where to find **more** existing data?

→ Repositories

- Often open access metadata   *Example [4TU.ResearchData](#)*

→ Databases of laboratories

- Open/restricted access   *Example [NCL](#)*

→ Contact the corresponding author

- You can always give it a try...   *Gabelica et al., 2022*

Findable + Accessible

- Contact the corresponding author... *Gabelica et al., 2022*:

  - Investigation of 3556 open access articles

  - Many researchers not compliant with their published data sharing statement

  - Most frequent (42%): data sets available on reasonable request But...

  - Among 1792 manuscripts, 1669 (93%) authors either did not respond or declined to share their data

Gabelica M, Bojčić R, Puljak L (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. Journal of Clinical Epidemiology 150:33-41. https://doi.org/10.1016/j.jclinepi.2022.05.019

Findable + Accessible

Where to find **more** existing data?

→ Repositories

- Often open access metadata   *Example* *4TU.ResearchData*

→ Databases of laboratories

- Open/restricted access   *Example* *NCL*

→ Contact the corresponding author

- You can always give it a try...   *Gabelica et al., 2022*

Findable + Accessible

Where to find **more** existing data?

→ Repositories

- Often open access metadata

→ Databases of laboratories

- Open/restricted access

→ Contact the corresponding author

- You can always give it a try…

Most transparent
Higher chance of success

Least transparent
Lower chance of success

Interoperable + Reusable

- First step: What am I allowed to do with this data?
  - → Licenses

- Second step: Can I read the file types?
  - → Evaluate cost-benefit (when conversions are needed, e.g. hardcopy data)

- Third step: Trying to understand what the data exactly is about
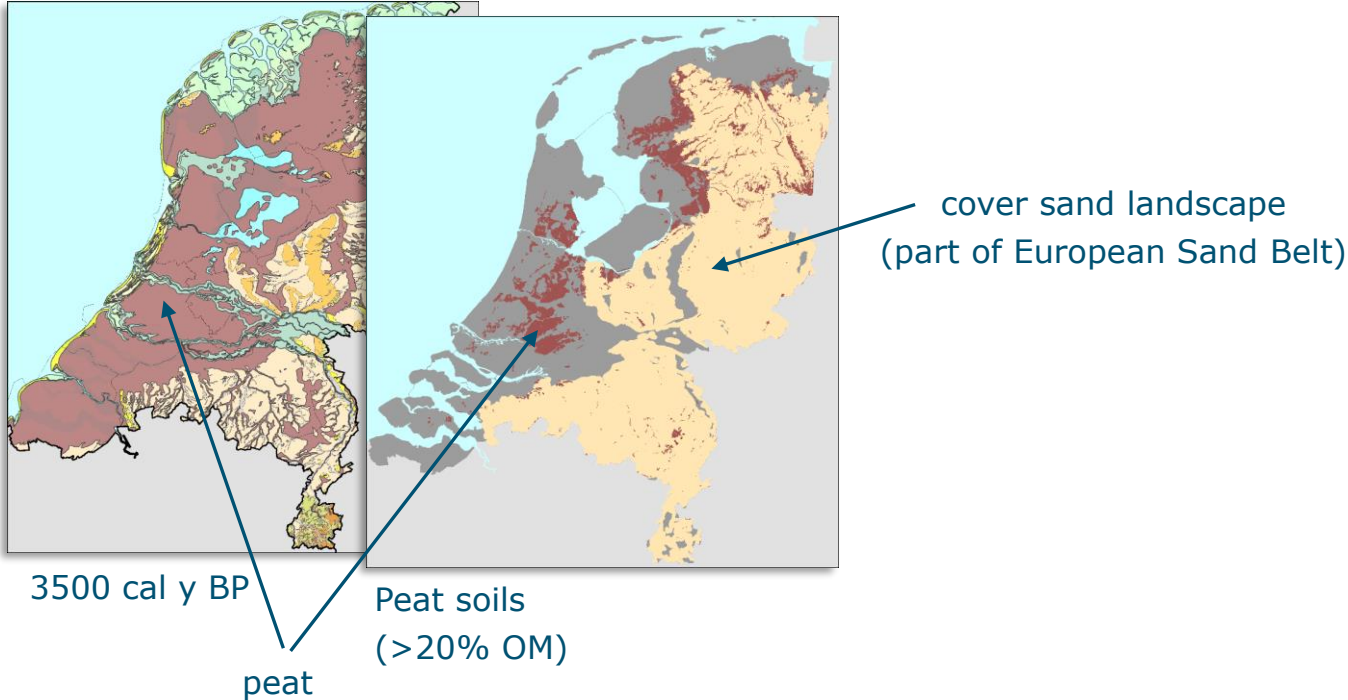 - → Documentation, metadata

Meta-analysis: reuse in the leading role

- Meta-analyses are increasingly relevant

- Overarching trends in space and time

- Analyses on large spatial scales
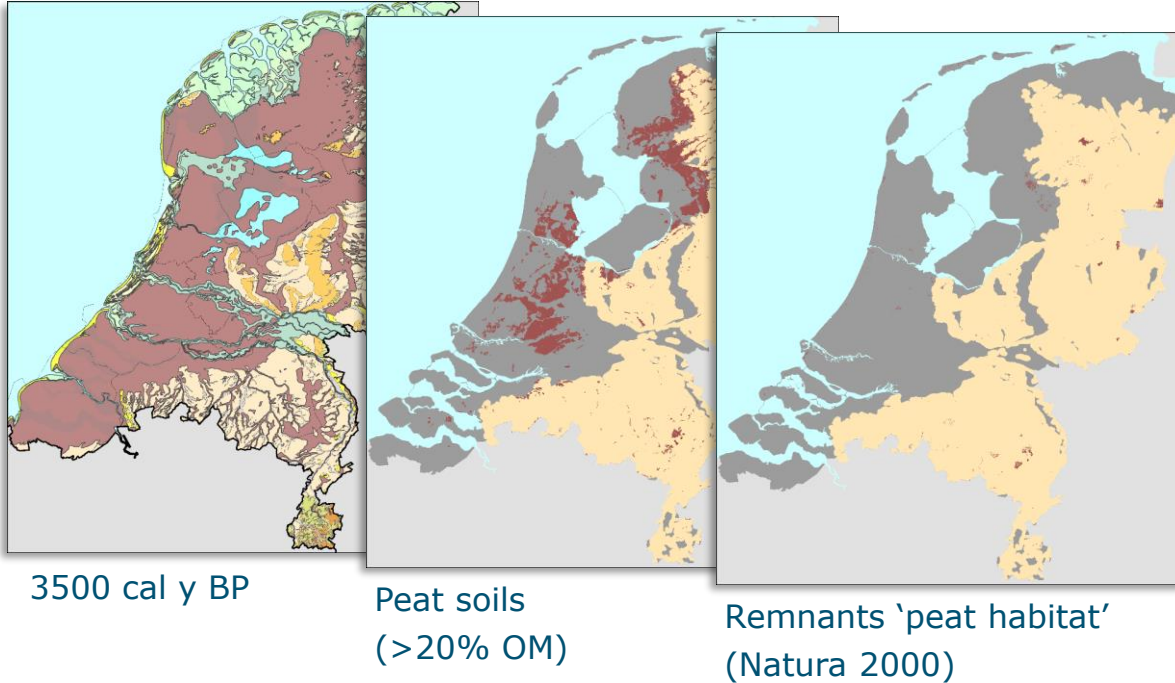
WAGENINGEN
UNIVERSITY & RESEARCH

Case study – introduction

- Patterns in peat growth in the cover sand landscape of the Netherlands:

  - When did these swamps begin to form?

  - Ages are determined with radiocarbon dating

- Why are existing data especially important for this topic?

Case study – why existing data?

cover sand landscape
(part of European Sand Belt)

3500 cal y BP

Peat soils
(>20% OM)

peat

WAGENINGEN
UNIVERSITY & RESEARCH

Case study – why existing data?

3500 cal y BP

Peat soils
(>20% OM)

Remnants 'peat habitat'
(Natura 2000)

Field studies are
limited due to
large-scale
disappearance of
peatlands

WAGENINGEN
UNIVERSITY & RESEARCH

Case study – data search

- Where did we find existing data?

  - Data tables in manuscripts/reports   open/restricted access

  - Databases of laboratories   restricted access
    digital + physical location

  - Contact the corresponding author   to obtain use rights

Case study – data search

■ Study area in northern NL

■ 313 dates of peat layers

■ Dates of diverse origin

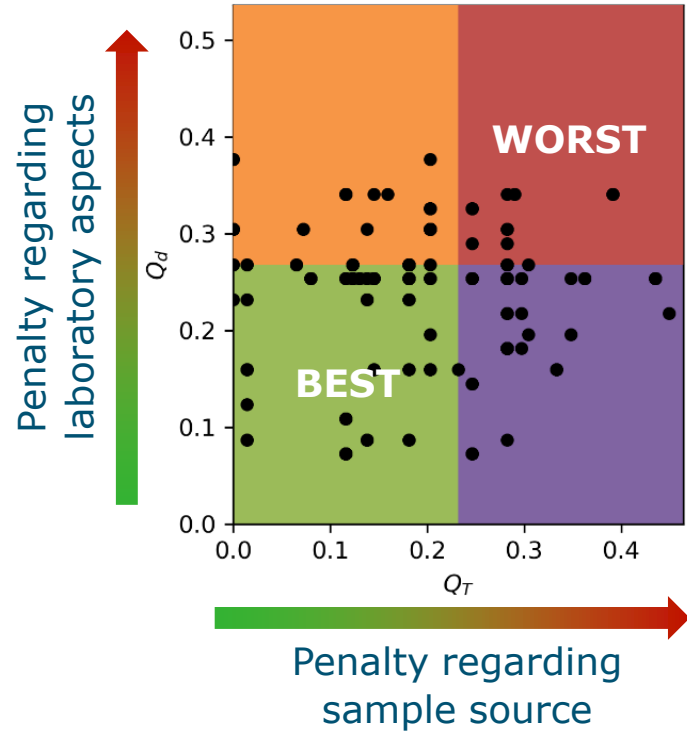→ Quality assessment

Case study – quality assessment

- Details about the sample source (information on sampling context)

  E.g. location, sampling depth

- Details about laboratory processing of the sample
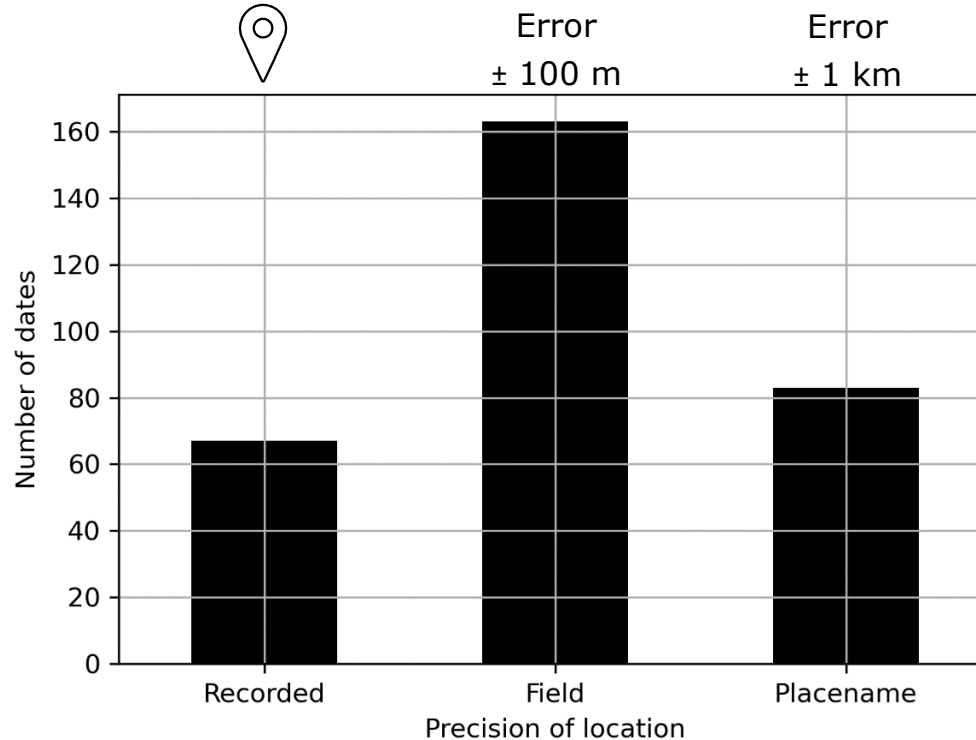
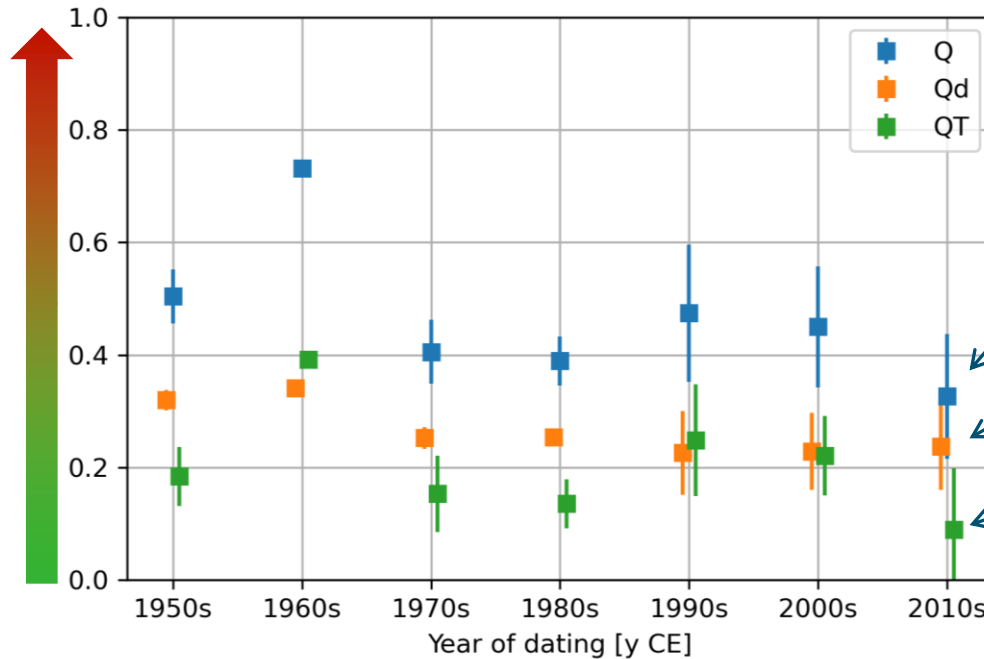 E.g. small sample, large sample

Case study – quality assessment

Four quality quadrants

Penalty regarding sample source

Case study – quality assessment

Case study – quality assessment

old ≠ worse

new ≠ better

Total penalty

Penalty regarding laboratory aspects

Penalty regarding sample source

Case study – some key points

- A quality assessment provides a clear format to distinguish certainty levels of reused data

- Certainty levels may provide insight in potential sources of error

→ Data reuse requires a tailor-made approach for data-analysis

- Studies that reuse legacy data may yield new insights that require a bird's-eye view to be discovered

Case study – FAIR dataset of reused data

- Dates performed at the radiocarbon lab of University of Groningen, either longer than 10 years ago or published with CC-BY license

- *Dates performed at the radiocarbon lab of University of Groningen, either less than 10 years ago or not published with CC-BY license*

- Dates performed at other labs, published with CC-BY license

- *Dates performed at other labs, published without CC-BY license*

WAGENINGEN
UNIVERSITY & RESEARCH

Case study – FAIR dataset of reused data

- Contacted authors

- Permission to:

  - reuse the data

  - publish the data in a CC-BY dataset of the reused dates

Quik C, Van der Velde Y, Harkema T, Van der Plicht H, Quik J, Van Beek R, Wallinga J (2021) Data from: Using legacy data to reconstruct the past? Rescue, rigor and reuse in peatland geochronology. 4TU.ResearchData https://doi.org/10.4121/14406347

Case study – FAIR dataset of reused data

Approach to property rights of the used legacy data:

We have taken utmost care in our reuse of existing data. Our approach to property rights of the used legacy data is explained below.

The legacy dataset contains 313 radiocarbon dates, of which:

1. 269 dates were performed by the Centre for Isotope Research (CIO) at the University of Groningen (The Netherlands) longer than 10 years ago (anno 2021), or less than 10 years ago but published under a CC-BY license.
2. 5 dates were performed by the CIO less than 10 years ago and <u>not</u> published under a CC-BY license.
3. 15 dates were performed by other labs and published under a CC-BY license.
4. 24 dates were performed by other labs and <u>not</u> published under a CC-BY license.

Dates in group 1 or 3 that were published under a CC-BY license may freely be shared and/or adapted.
Source: https://creativecommons.org/licenses/by/4.0/, accessed April 2021.

For dates in group 1 that were performed longer than 10 years ago (anno 2021), the Terms and Conditions of the CIO state that:
"Sample(s) submitted to the CIO for analysis by the Submitter in relation to the analytical services will be archived for at least 10 years, unless marked for return upon submission. Once this 10-year period has lapsed, the CIO, potentially in collaboration with third parties, will be at liberty to utilise the samples for further academic study and to publish any data arising from such research without seeking the permission of the original submitters."
Source: https://www.rug.nl/research/centre-for-isotope-research/customers/general-conditions, accessed April 2021.

For dates from group 2 and 4, we contacted the authors of the publications in which these dates were found. We thank for their kind permission to reuse the dates, and to publish the legacy dataset under a CC-BY license.

Tips & tricks for data reuse

→ Handout *Tips & Tricks for Data Reuse*

Tips & tricks for data reuse

■ Data collection

- ● Design a database format for standardized registration of reused data

- ● It can be useful to keep a list of data that was found but not included in the database

Tips & tricks for data reuse

- Methodology

  - How were the data obtained? E.g., which search terms were used to find these data?

  - On what terms were data in- or excluded? E.g., excluding hardcopy data

Tips & tricks for data reuse

■ Methodology (continued)

- How was the suitability of the data for the present meta-analysis determined?

- How was the quality of the data evaluated?

Tips & tricks for data reuse

■ Results and Discussion

- Is divergent quality of the reused data affecting conclusions?

- Keep in mind that it is a balancing act to reach robustness of conclusions without being too strict and discarding the majority of reused data.

Tips & tricks for data reuse

■ Results and Discussion (continued)

● Keep an eye out that a quality assessment does not lead to

over-representation of reused data from only a few studies

(data from the same study often have comparable quality).

Tips & tricks for data reuse

- Closing the loop: FAIR dataset of reused data

  - Include a section in the README file where you explain which licenses apply to the reused data, and whether additional permissions were obtained.

-> Discipline & Motivation talk!