gesis Leibniz Institute for the Social Sciences

Data management plans and data management costs

Dr. Anja Perry

International Workshop 2023 at the Social Science Japan Data Archive | online | 22.03.2023

Leibniz Association

gesis Leibniz Institute for the Social Sciences

Agenda

Benefits and Costs of Research Data Management and Data Sharing

Data Management Plans

- Purpose and structure
- Information and tools
- Other Initiatives

Data Sharing Costs

- Research / researchers' needs so far
- Analysis at GESIS
- Cost drivers and cost reduction

Research Data Management and Data Sharing

Picture by fauxels (Pexel-Lizenz; https://www.pexels.com/de-de/foto/kollegen-die-sich-das-umfrageblatt-ansehen-3183153/)

Benefits for your team

- Better workflows and knowledge exchange
- Saves resources
- Helps to achieve correct results

Picture by Tima Miroshnichenko (Pexel-Lizenz; https://www.pexels.com/de-de/foto/laptop-buro-internet-geschaft-6693655/)

gesis Leibniz Institute for the Social Sciences

Benefits for your own research

- Transparent and valid research
- Allows efficient re-use of your own data
- Additional data publication
- Better cited research
- More possibilities for cooperation and networking

Picture by Olena Bohovyk (Pexel-Lizenz; https://www.pexels.com/de-de/foto/braunes-holzregal-mit-buchern-3646172/)

Benefits for science

- Better access to data
- Efficient re-use of data by others
- Data can be used in different contexts
- Use in teaching
- …

Picture by SahraRichterArt (Pixabay license; https://pixabay.com/de/photos/jongleur-jongleure-juggler-toss-up-2329843/)

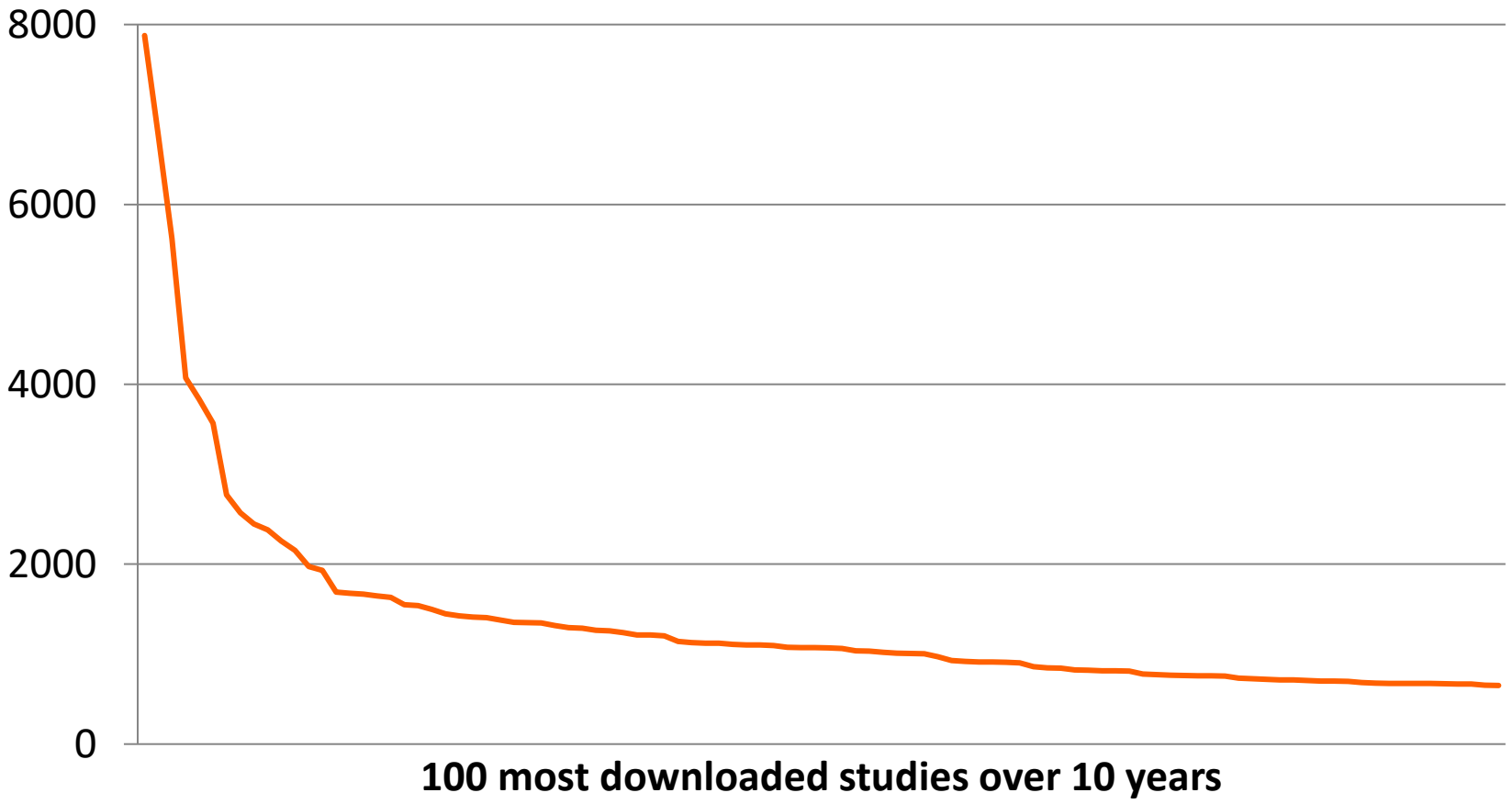
Data sharing and RDM may be required by

- your **institute**, e.g., in your contract or project agreements
- **Funding organizations**, e.g., to ensure subsequent use
- **Journals** that request your data before publishing your article
- Your **supervisor**...

Check for these conditions!

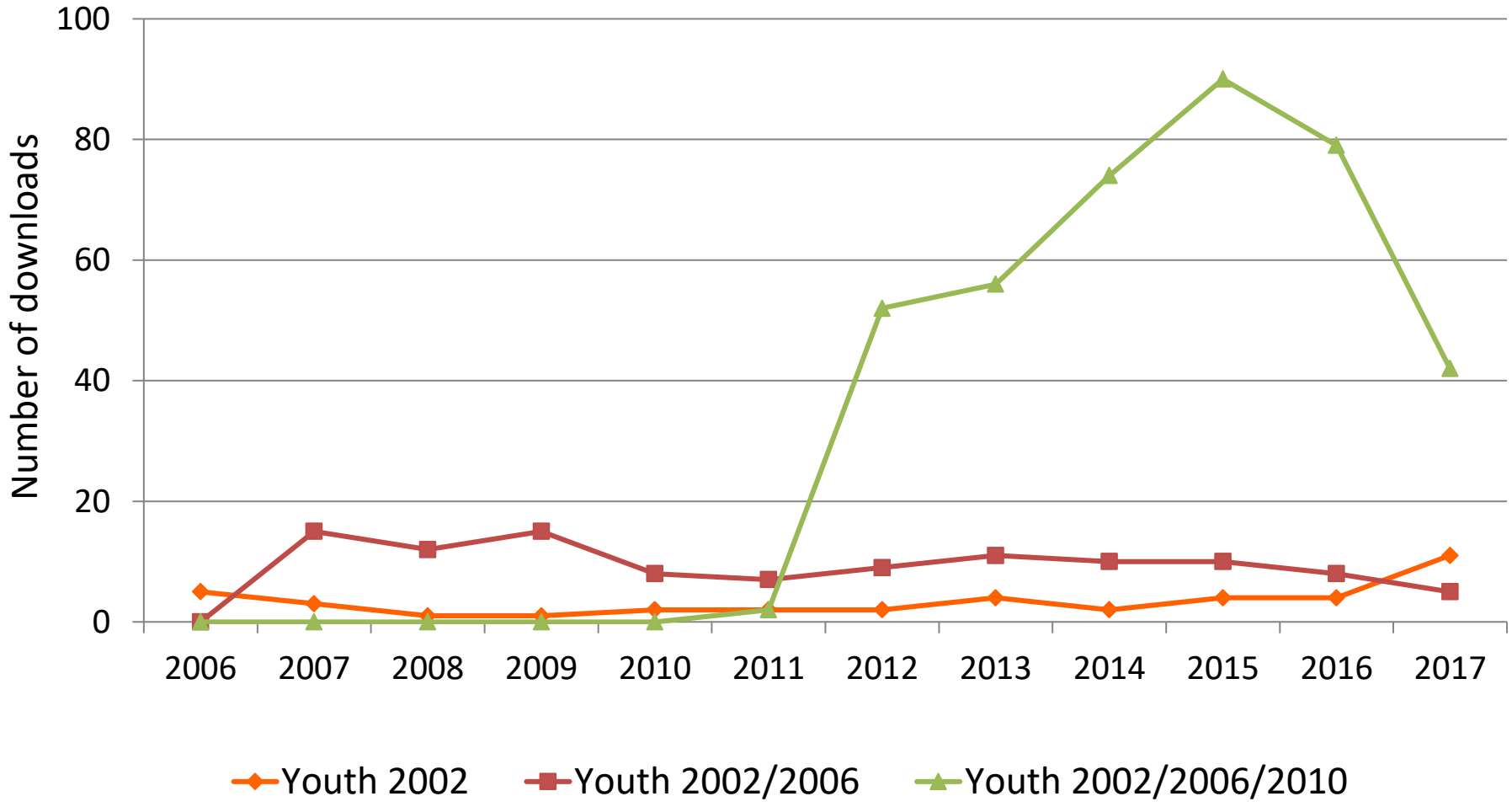
Picture by A. Herrema & H. Bouwteam (CC BY; https://www.fosteropenscience.eu/content/cartoonpublication-and-data)

Download distribution



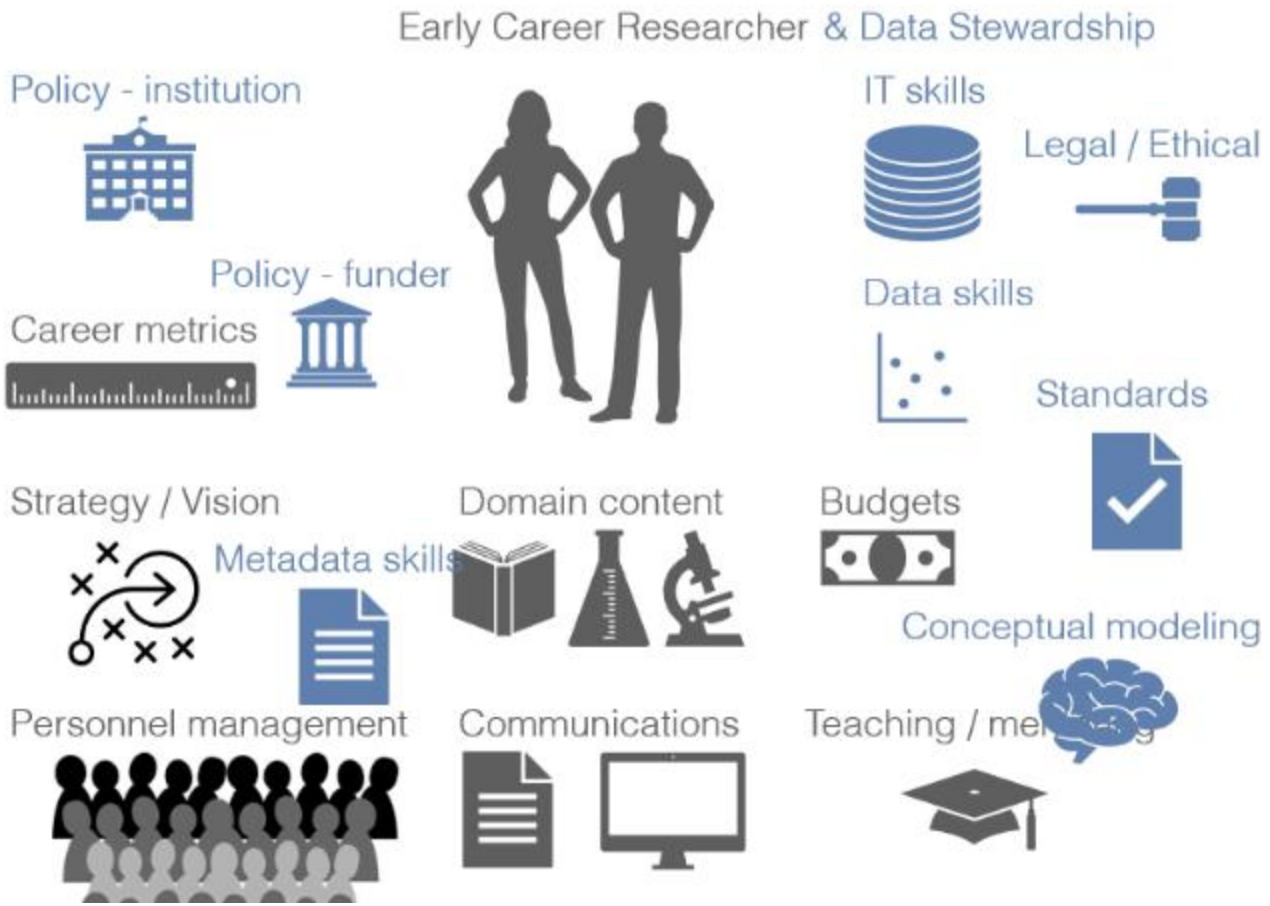
GESIS - Data Archive for the Social Sciences (2018): Downloadstatistik GESIS Datenarchiv. GESIS Datenarchiv, Köln. ZA6899 Datenfile Version 2.0.0, doi:10.4232/1.12979

Shell Youth Study 2002, 2006 and 2010



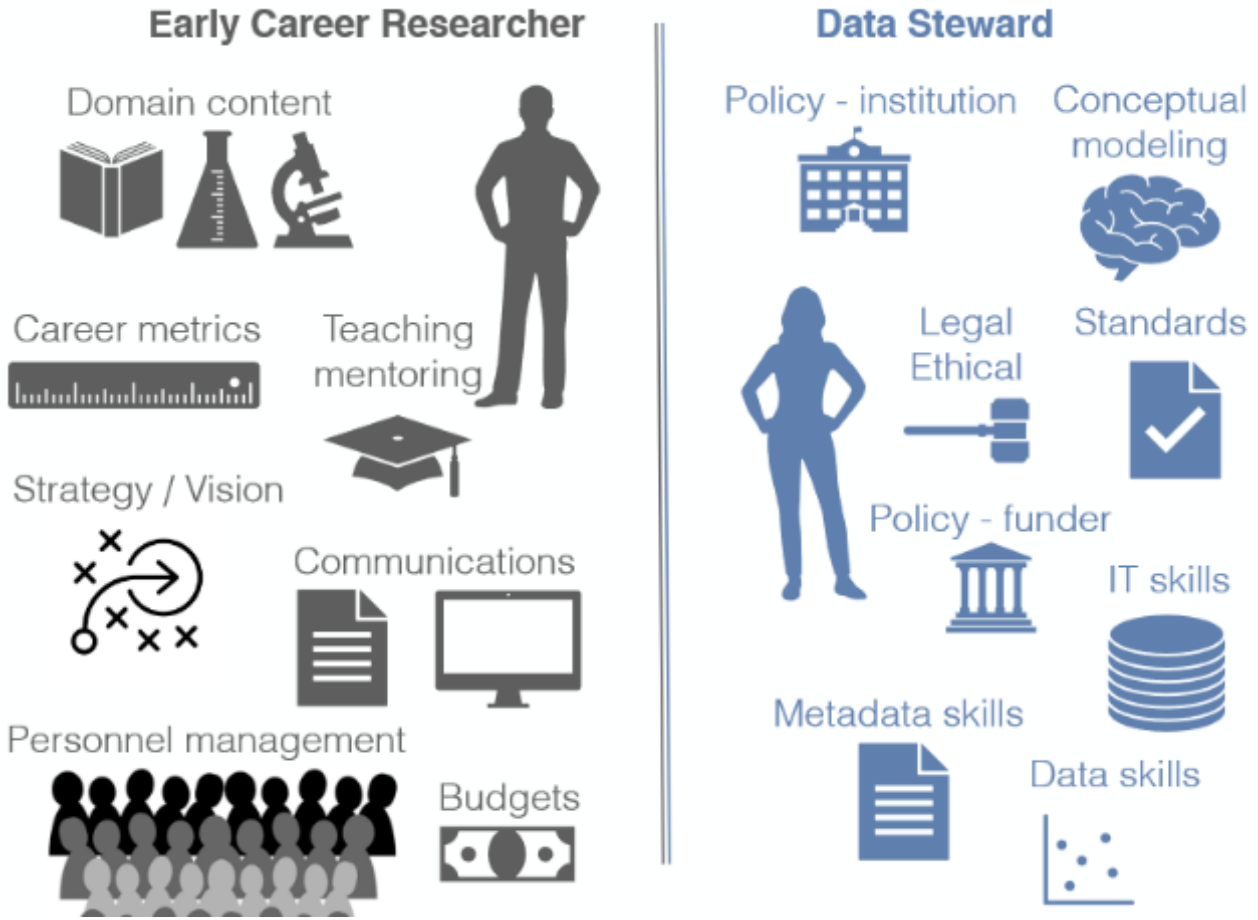
GESIS - Data Archive for the Social Sciences (2018): Downloadstatistik GESIS Datenarchiv. GESIS Datenarchiv, Köln. ZA6899 Datenfile Version 2.0.0, doi:10.4232/1.12979

Increased responsibilities



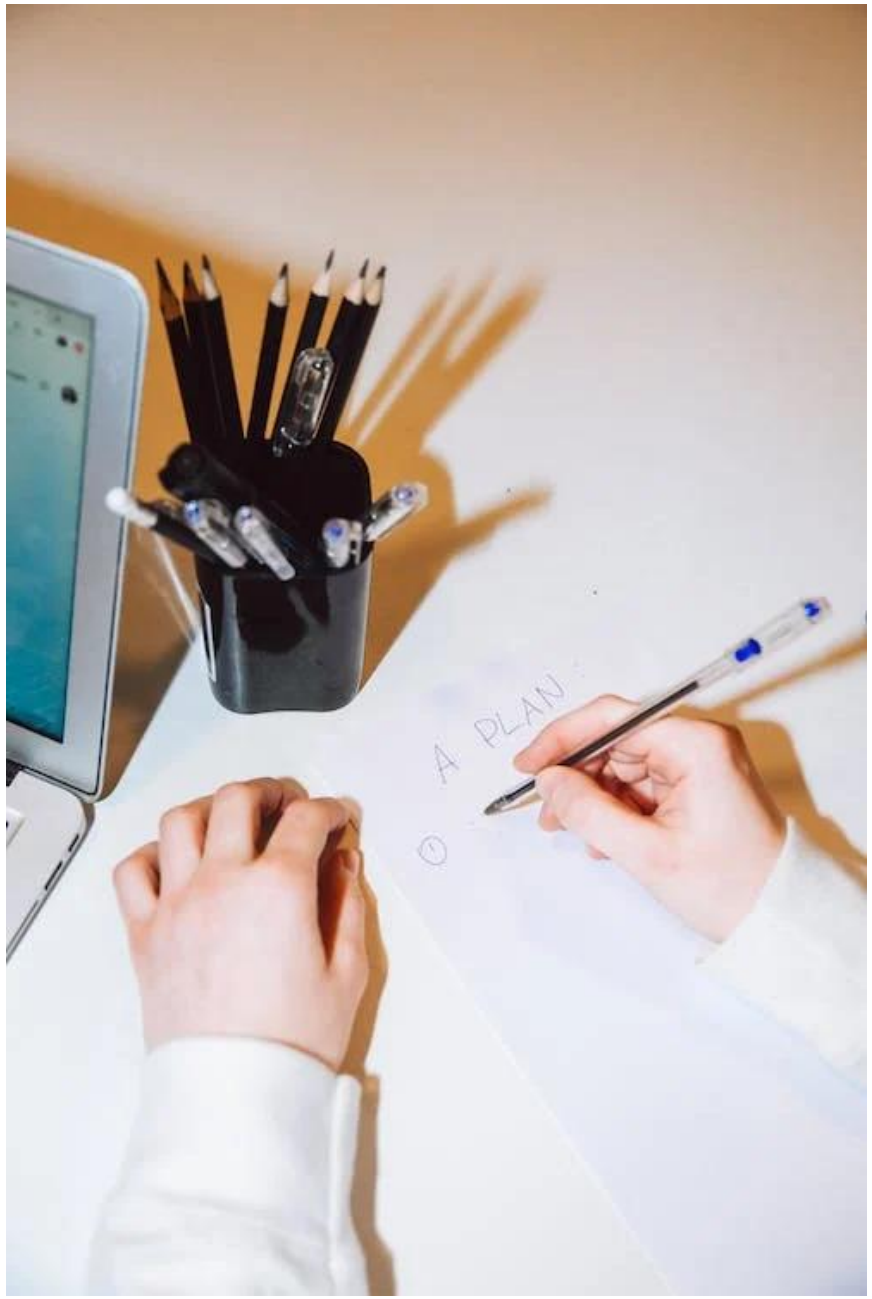
(Erik Schultes, CC BY, 2020)

Increased responsibilities



(Erik Schultes, CC BY, 2020)

Data Management Plan



Picture by Polina Zimmermann (Pexel-Lizenz; https://www.pexels.com/de-de/foto/person-die-auf-weissbuch-schreibt-3746963/)

What is a Data Management Plan?

- Describes treatment of data during and after the project
- Covers all steps of the data lifecycle
- Addresses important aspects and responsibilities **before** data is collected, such as
  - Data protection
  - Data storage
  - Data ownership
  - Data sharing
- “Living document”
- Requirement:
  - For EU funded projects since 2017
  - Same for more and more national funders

Data lifecycle: CESSDA Training Team (2017 - 2022). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. Retrieved from https://dmeg.cessda.eu/ (CC BY-SA)

Key components

- Overview
- Organizing and documenting your data
- Processing your data
- Storing your data and metadata
- Protecting your data
- Archiving and publishing your data
- Discovering data

- See also: CESSDA Checklist

Example on Protecting your data: Informed consent

Protecting your data

Ethical review (if applicable)

- Does your project require approval by a local ethics committee?
- How will possible ethical issues be taken into account, and codes of conduct followed?

Informed consent (if applicable)

- Do you require informed consent for your project?
- If so, how will permission be obtained?
- How are consent files organised and stored?

(sensitive) Personal data /confidential information (if applicable)

- How will access to (sensitive) personal data during the project be controlled?
- How will collaborators be granted access to the data in a secure way?
- If the research project is going to have data that includes confidential information or information that requires informed consent, is there a requirement to notify a privacy officer?
- Is there any confidential information within the material that requires special treatment and/or limits the access to it during/after the project?
- How will the material be protected during/after the project?
- How will permissions and restrictions be enforced?

Source: CESSDA Checklist

gesis Leibniz Institute for the Social Sciences

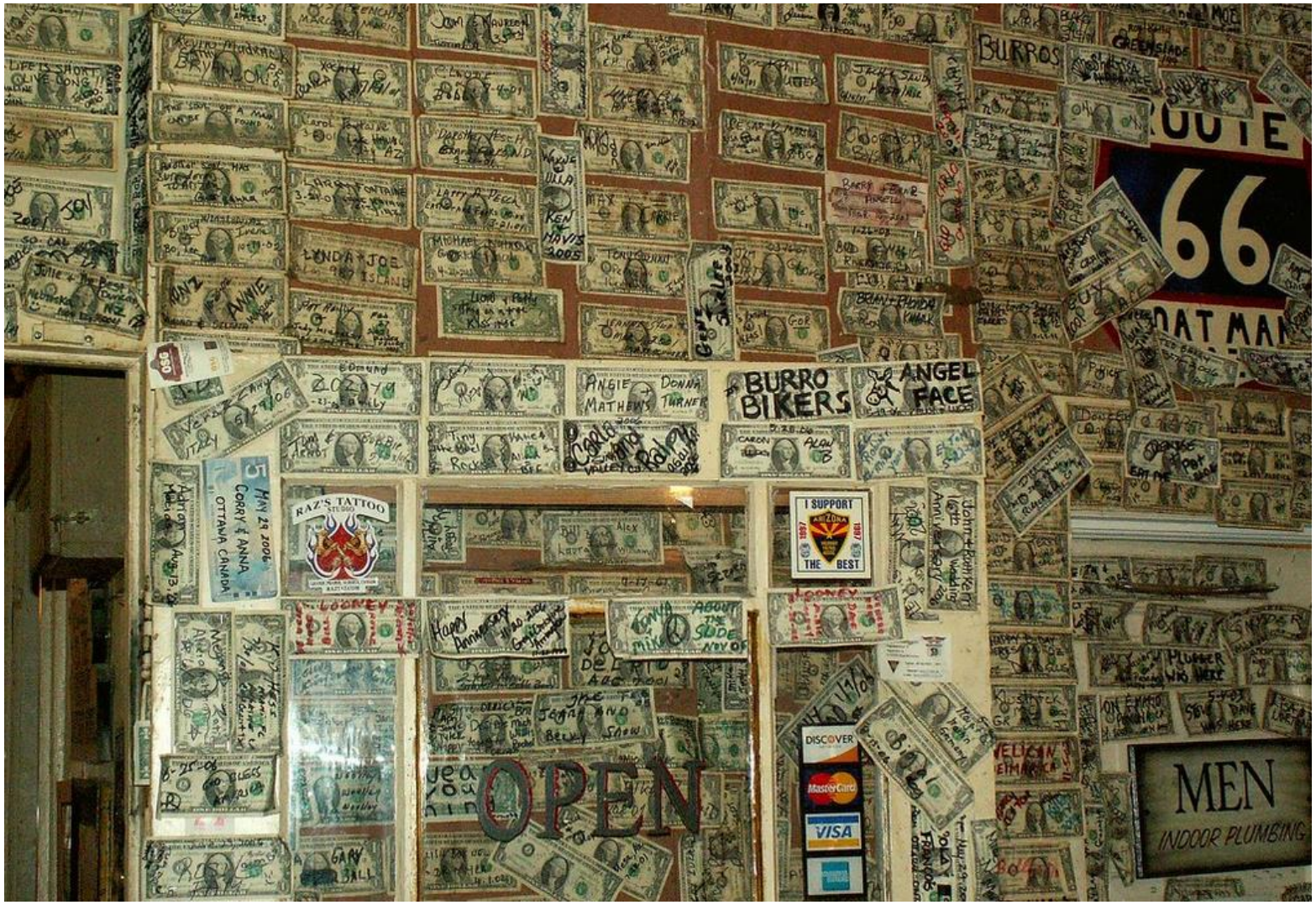
DMP Tools

Tool	Website	What it does	
DMPonline	https://dmponline.dcc.ac.uk/	- Data Management Online (DMPonline) - Online tool with templates from different funders - Guidance when answering DMP questions	Funder driven, combined into DMPRoad map
DMPTool	https://dmptool.org/	- Data Management Tool (DMPTool) - Online tool with templates from different funders - Guidance when answering DMP questions	
RDMO	https://rdmorganiser.github.io/en/	- Research Data Management Organizer (RDMO) - Online tool to organize research process - Users publish templates to be re-used, f.ex. for in a specific discipline - User community, mostly German-speaking, but also in France and Italy	Community driven
Stamp	https://www.forschungsdaten-bildung.de/stamp-nutzen	- Standardized Data Management Plan for Education Research (Stamp) - Focus on Germany (esp. regarding legal aspects), available in German - Predefined DMP that guides through RDM (checklists) for different types of data - Implemented in RDMO	

CESSDA Data Management Expert Guide

https://dmeg.cessda.eu/

Data Sharing Costs



Picture by cobalt123 (CC BY-NC-SA; https://www.flickr.com/photos/cobalt/42523 1363/in/photostream)

Research data management is work!

A personal take on science and society

World view

By Barend Mons

Invest 5% of research funds in ensuring data are reusable

It is irresponsible to support research but not data stewardship, says Barend Mons.

Many of the world's hardest problems can be tackled only with data-intensive, computer-assisted research. And I'd speculate that the vast majority of research data are never published. Huge sums of taxpayer funds go to waste because such data cannot be reused. Policies for data reuse are falling into place, but fixing the situation will require more resources than the scientific

> **Funders hold the stick: they should disburse no further funding without a data-**

Data stewardship offers excellent returns on investment.

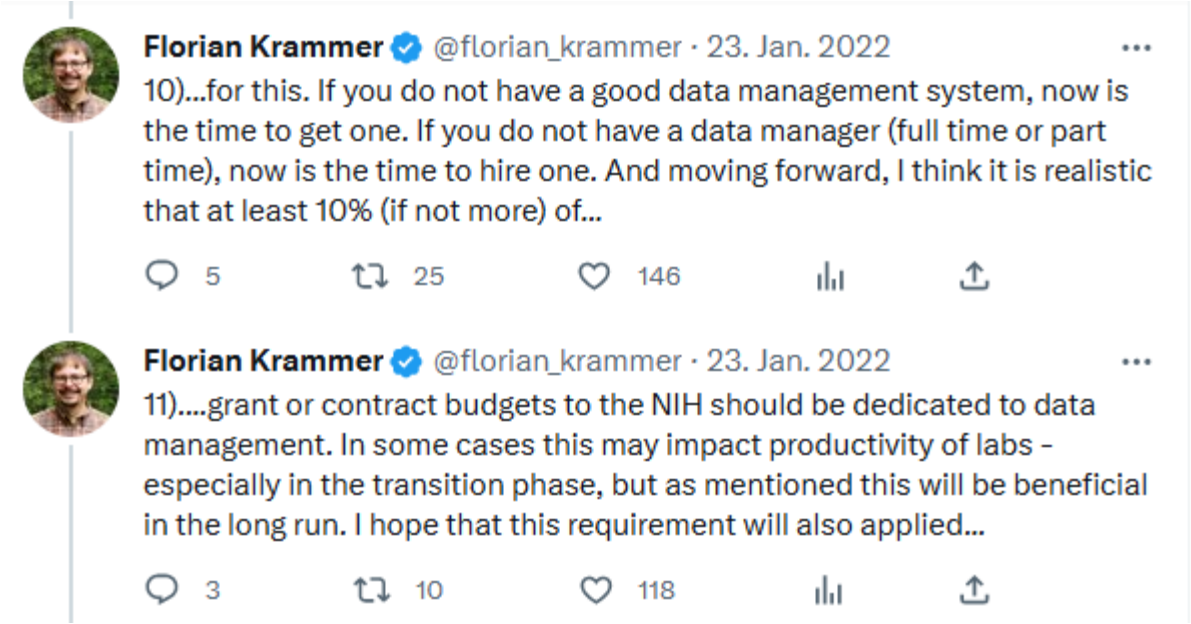
A 2018 European Commission report estimates that problems with the reuse of data cost the EU at least €10 billion each year in the academic sector alone, and €16 billion in lost innovation opportunities. I translate that as roughly €100 billion lost annually at the global level. That's not even counting related reproducibility problems.

The FAIR guiding principles are now cited three times per day, but citations do not equate to practice. My colleagues and I, along with European Open Science Cloud, an initiative aimed at promoting open-science practices, scoped requirements for the continent-wide data-shar-

Source: Nature | Vol 578 | 27 February 2020
https://doi.org/10.1038/d41586-020-00505-7

Research data management is work!

Or “at least 10%”?



Florian Krammer @florian_krammer · 23. Jan. 2022

10)...for this. If you do not have a good data management system, now is the time to get one. If you do not have a data manager (full time or part time), now is the time to hire one. And moving forward, I think it is realistic that at least 10% (if not more) of...

5 25 146

Florian Krammer @florian_krammer · 23. Jan. 2022

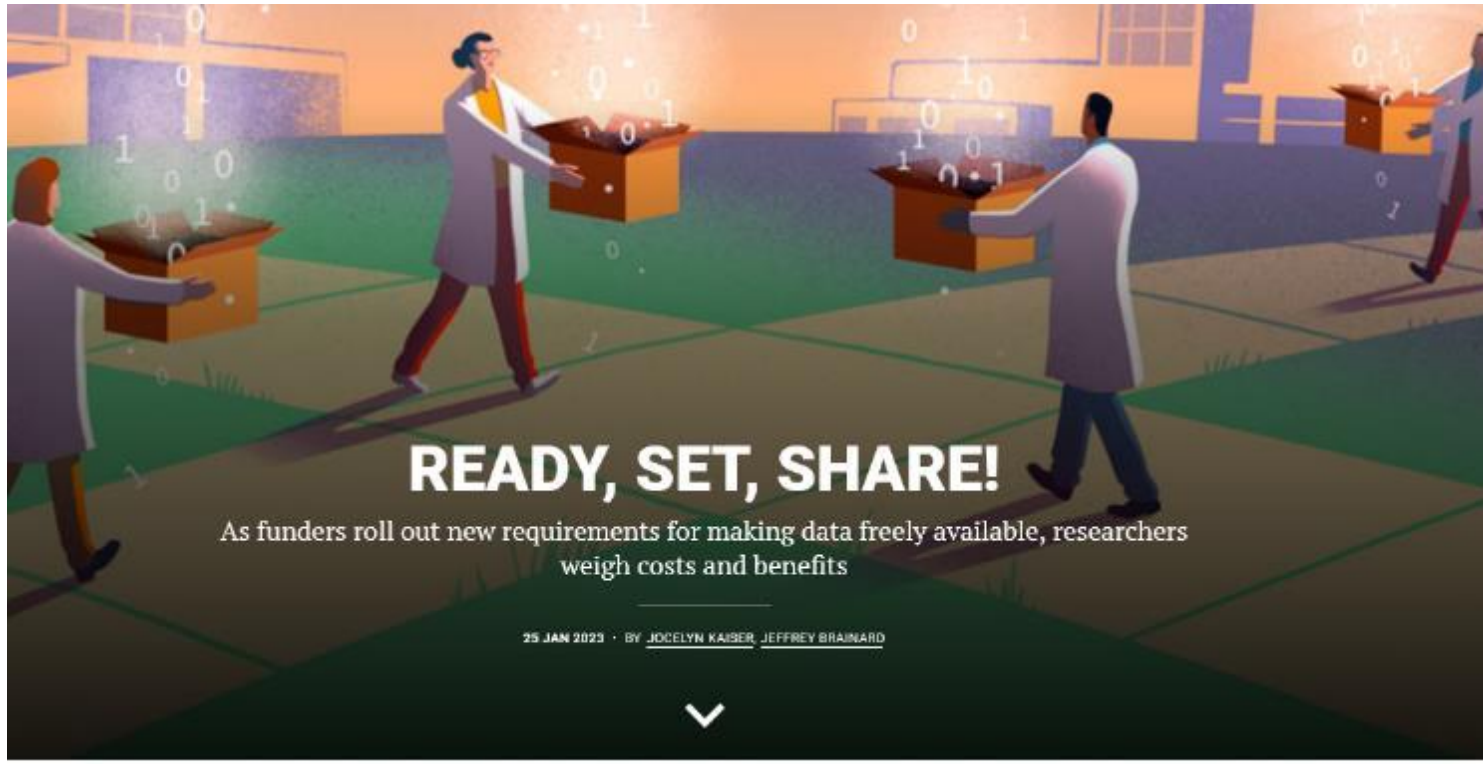
11)....grant or contract budgets to the NIH should be dedicated to data management. In some cases this may impact productivity of labs - especially in the transition phase, but as mentioned this will be beneficial in the long run. I hope that this requirement will also applied...

3 10 118

Krammer (2022) https://twitter.com/florian_krammer/status/1485271552325300230

Funders' role and support

- Researchers can often add costs for data managers, staff time to prepare data, and repository fees
- May cut into the funds available for research
- Universities sometimes have campuswide services
- Challenge: even repositories often don't have sustainable business models



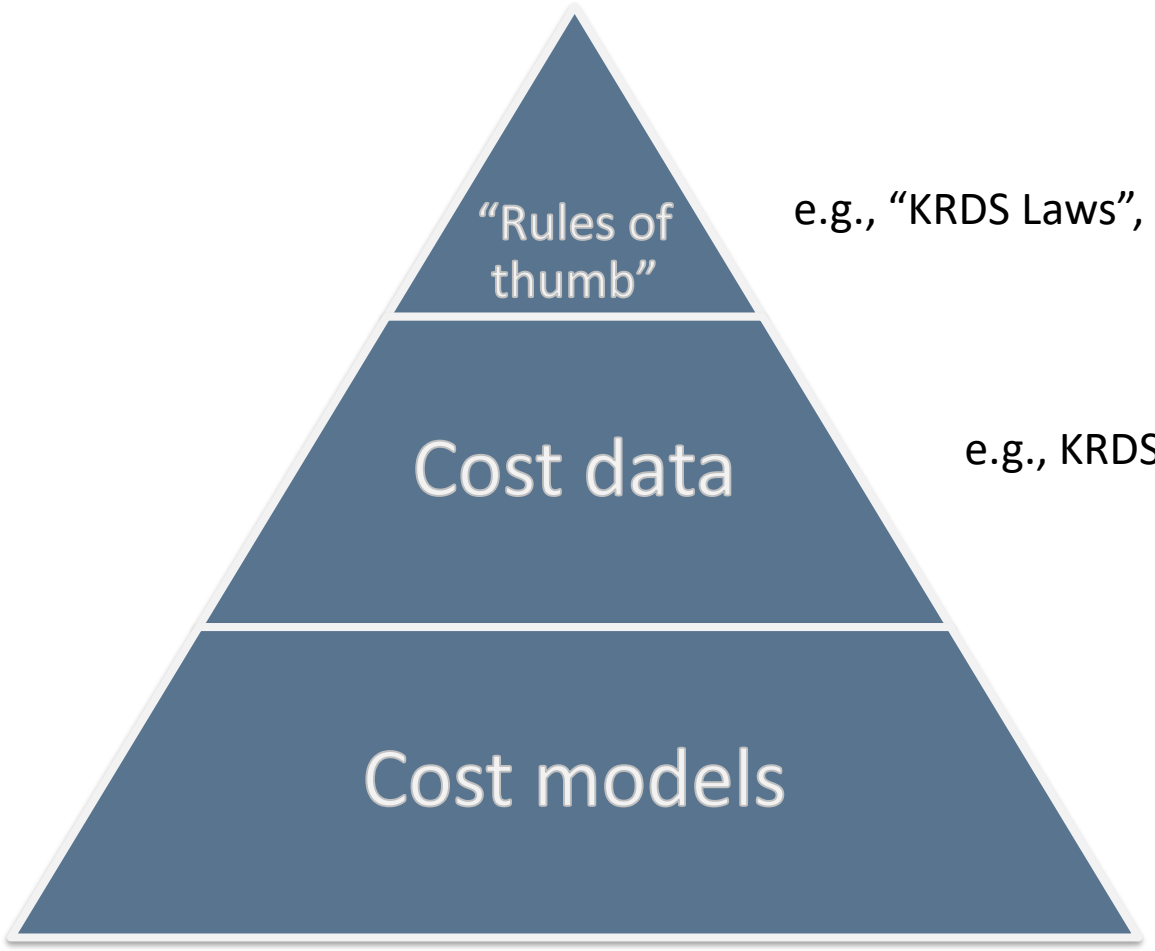
SHARE:

A version of this story appeared in Science, Vol 379, Issue 6630.

Physiologist Alejandro Caicedo of the University of Miami Miller School of Medicine is preparing a grant proposal to the U.S. National Institutes of Health (NIH). He is feeling

https://www.science.org/content/article/ready-set-share-researchers-brace-new-data-sharing-rules

Research on RDM costs is scarce



e.g., "KRDS Laws", Kryder's Law

e.g., KRDS Cost Survey, 4C Costs of Curation Calculator

e.g., Keeping research data safe (KRDS), 4C Project, DANS

More on costs of digital preservation here:
http://wiki.opf-labs.org/display/CDP/Home

Source: CESSDA (2017). Cost-Benefit Advocacy User Guide. https://dx.doi.org/10.5281/zenodo.3662438

Our research questions

Participants in our RDM workshops ask us:

“How much funding can we apply for to account for RDM tasks?”

Perry, A. and Netscher, S. (2022), "Measuring the time spent on data curation", *Journal of Documentation*, Vol. 78 No. 7, pp. 282-304. https://doi.org/10.1108/JD-08-2021-0167

Tools/guides for researchers

- UKDS costing tool (checklist): https://dam.ukdataservice.ac.uk/media/622368/costingtool.pdf
- Tools that recommend RDM budget based on services available at a specific university

Our approach

- **Working hours were tracked (Dec. 2016 – Sep. 2017)**
  - 3 studies with multiple waves
  - 10 datasets
  - Times for documenting one of the 3 studies were not tracked (2 datasets)

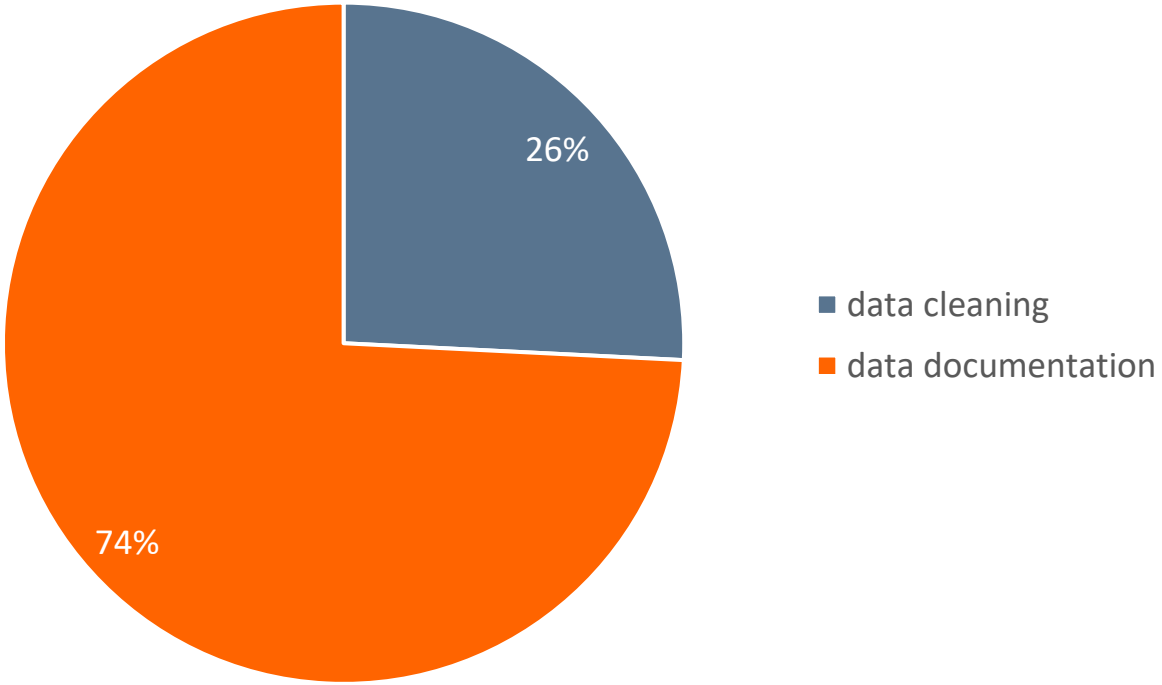
- **We look at 4 factors (Committee on Forecasting Costs… (2020)):**
  - # variables
  - # questions
  - # open answer questions
  - # questions affected by filters

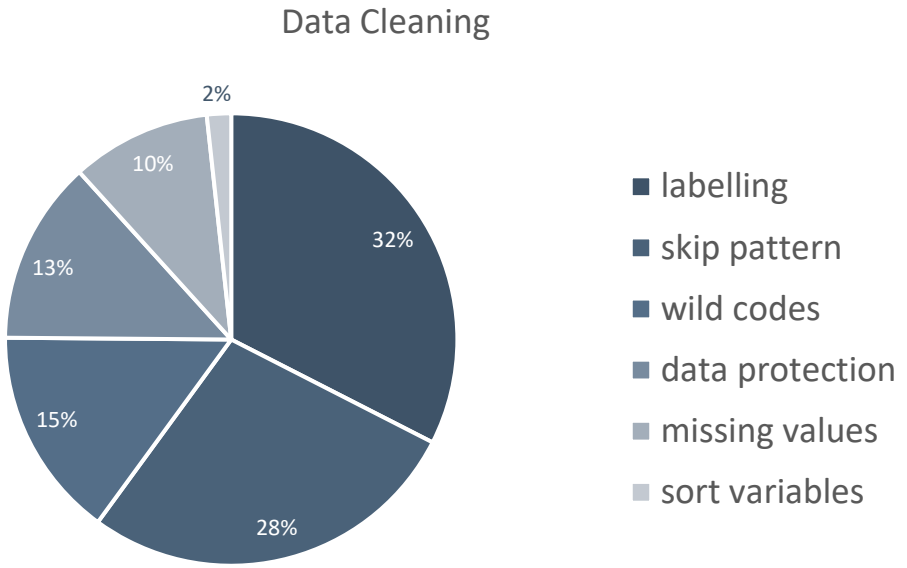
  } size of data

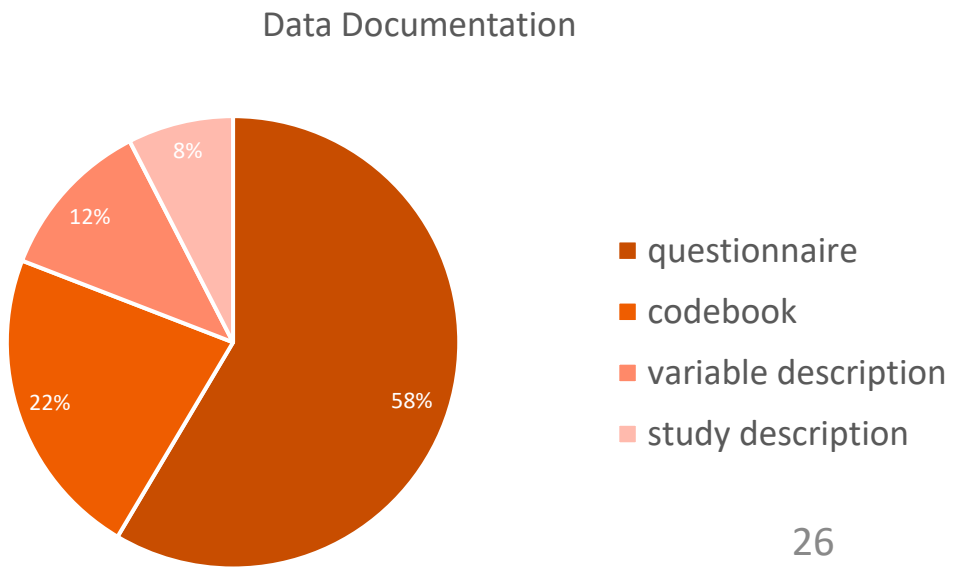
  } complexity of data

- **Focus group interview (Feb. 2021)**

Working times for different tasks







Average time for each dataset: 63h (≈ 8 days)
Average time for each variable: 7.85 min

gesis Leibniz Institute for the Social Sciences

Correlations

	# variables	# questions	# open answer questions	# variables affected by filter
overall curation time	0.8774**		0.8071*	
cleaning	0.9156**		0.7858*	
missing values	0.7584*	0.8400**	0.8857**	
wild codes				
skip pattern	0.7673*	0.9531***	0.8464**	0.8671**
structure				
data protection				
variable and value labels	0.8331*	0.8326*	0.787*	0.8943**
sorting variables				
documentation	0.7593*		0.7351*	
questionnaire documentation	0.7084*		0.7988*	
variables documentation				
study description				
codebook (print)				

Note. Displayed are correlation coefficients; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; insignificant correlations not displayed; correlations with $r < 0.75$ in grey; n = 8.

Correlations

- Main factors:
  - # variables
  - # open answer questions

	# variables	# questions	# open answer questions	# variables affected by filter
overall curation time	0.8774**		0.8071*	
cleaning	0.9156**		0.7858*	
missing values	0.7584*	0.8400**	0.8857**	
wild codes				
skip pattern	0.7673*	0.9531***	0.8464**	0.8671**
structure				
data protection				
variable and value labels	0.8331*	0.8326*	0.787*	0.8943**
sorting variables				
documentation	0.7593*		0.7351*	
questionnaire documentation	0.7084*		0.7988*	
variables documentation				
study description				
codebook (print)				

Note. Displayed are correlation coefficients; *** p < 0.001, ** p < 0.01, * p < 0.05; insignificant correlations not displayed; correlations with $r < 0.75$ in grey; n = 8.

Correlations

- All four characteristics correlate with filters and labels

	# variables	# questions	# open answer questions	# variables affected by filter
overall curation time	0.8774**		0.8071*	
cleaning	0.9156**		0.7858*	
missing values	0.7584*	0.8400**	0.8857**	
wild codes				
skip pattern structure	0.7673*	0.9531***	0.8464**	0.8671**
data protection				
variable and value labels	0.8331*	0.8326*	0.787*	0.8943**
sorting variables				
documentation	0.7593*		0.7351*	
questionnaire documentation	0.7084*		0.7988*	
variables documentation				
study description				
codebook (print)				

Note. Displayed are correlation coefficients; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; insignificant correlations not displayed; correlations with $r < 0.75$ in grey; n = 8.

Correlations

- Only questionnaire documentation correlates with characteristics when looking at documentation

	# variables	# questions	# open answer questions	# variables affected by filter
overall curation time	0.8774**		0.8071*	
cleaning	0.9156**		0.7858*	
missing values	0.7584*	0.8400**	0.8857**	
wild codes				
skip pattern	0.7673*	0.9531***	0.8464**	0.8671**
structure				
data protection				
variable and value labels	0.8331*	0.8326*	0.787*	0.8943**
sorting variables				
documentation	0.7593*		0.7351*	
questionnaire documentation	0.7084*		0.7988*	
variables documentation				
study description				
codebook (print)				

Note. Displayed are correlation coefficients; *** p < 0.001, ** p < 0.01, * p < 0.05; insignificant correlations not displayed; correlations with $r < 0.75$ in grey; n = 8.

How are the steps connected?

- During initial checks (1-2h): Small mistakes are corrected right away
- Check for wild codes:
  - By checking labels
  - When no labels exist or when they are incomplete, checking for wild codes takes more time
- Questionnaire documentation involves many of the following steps, e.g., variable documentation

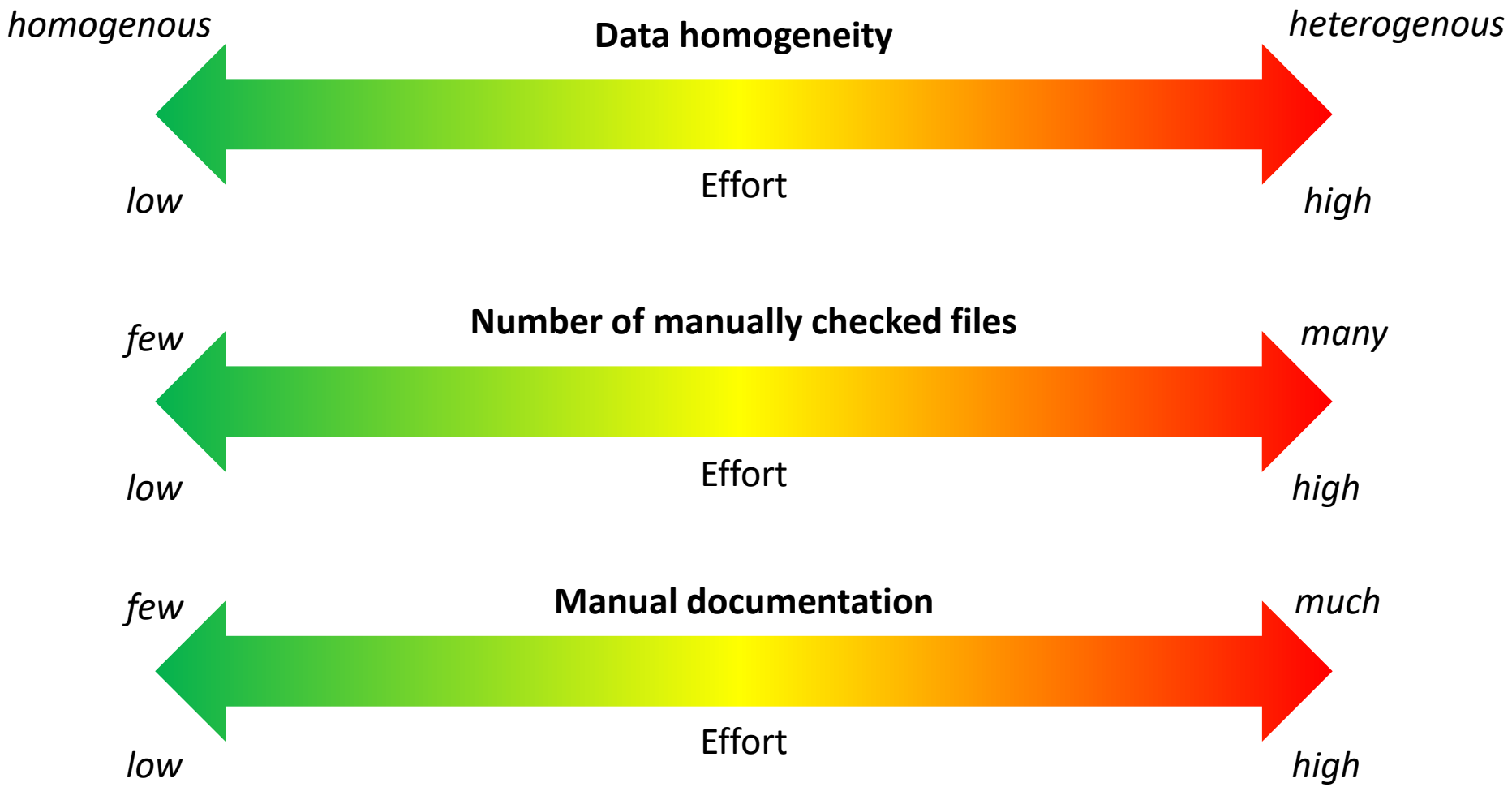
Picture by Ankush Rathi (Pexels license; https://www.pexels.com/photo/brown-concrete-door-925067/)

What else have we learned?

- Learning effect
- Not the number of filters is decisive, but their complexity and their documentation
- Low data quality increases the time spent on cleaning
- Open answer questions and data protection → number of cases play a role
- DDI standard and tools very helpful for documentation

Picture by Wokandapix (Pixabay license; https://pixabay.com/de/photos/lernen-wort-scrabble-briefe-1820039/)

General factors



gesis Leibniz Institute for the Social Sciences

Limitations

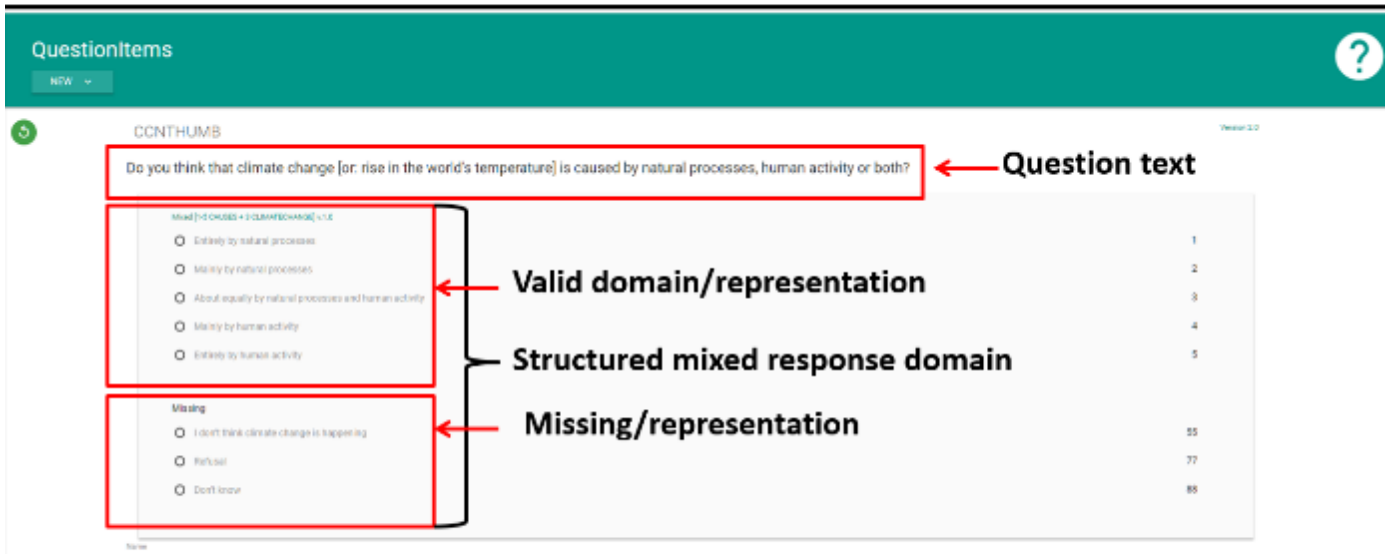
- Very small dataset!
- Data were already of high quality and not very complex
- Our curators are experienced and they use DDI tools
  - Established routines
  - Allows them to make multiple steps in parallel

How can DDI help?

Questionnaire Design and Documentation Tool

QDDT

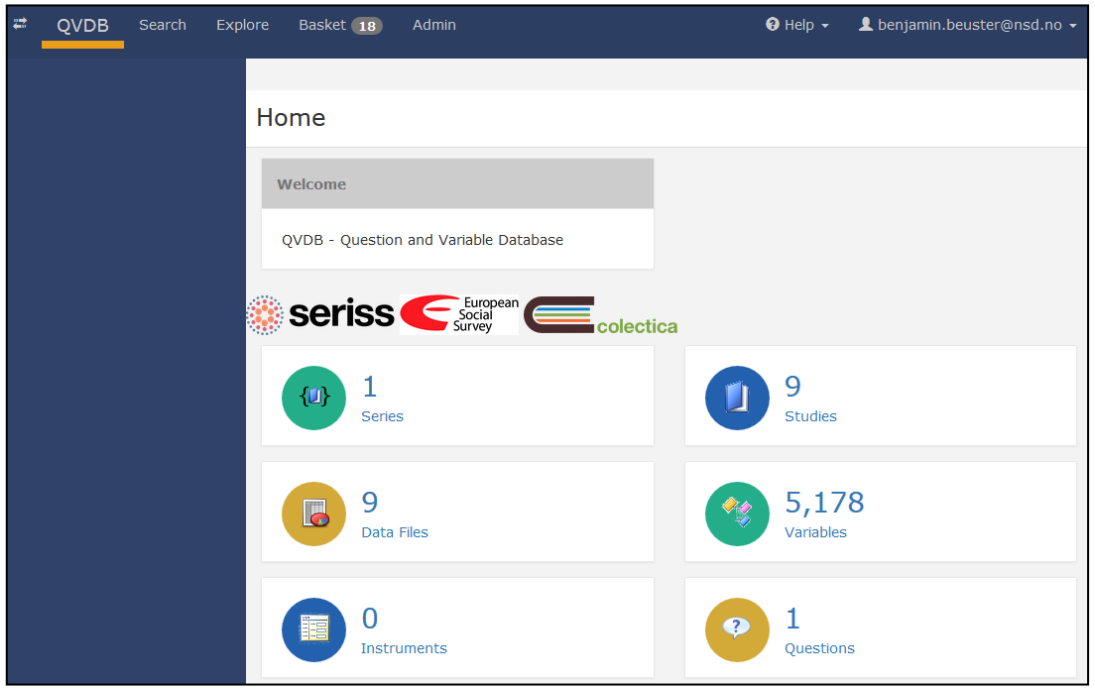
Structuring Question Items and Response Domains for reuse



- **Only core content is added at the question item level**
- **Responses are maintained separately and linked to the question by reference**
- **Valid responses are maintained and reused separately from the missings**

Orten, Norland, Butt (2018)

How can DDI help?



Similar database model (DDI3.2)

Beuster (2018)

How can DDI help?

https://ddialliance.org/learn

Outlook

- Analysis was part of the Stamp project
- Results went into Stamp section „Responsibilities and Ressources“

But…

- We need more data to make recommendations for RDM funding
  - Data on data protection matters in qualitative research
  - Researcher‘s RDM vs. professional data curators
  - This data is difficult to collect (time consuming, legal aspects)
- RDM is still an underdeveloped field, needs professionalization

gesis Leibniz-Institut für Sozialwissenschaften

ありがとうございました – Thank you!

Contact

Dr. Anja Perry

anja.perry@gesis.org

Tel: +49 221 47694-464

@Datendealerin@fediscience.org

Sources

4C Project (2013). Collaboration to Clarify the Costs of Curation. https://www.4cproject.eu/

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). Keeping Research Data Safe—A Cost Model And Guidance For UK Universities [Final Report]. https://www.webarchive.org.uk/wayback/archive/20140615221657/http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Beuster, B. (2018). The Question and Variable Database (QVDB) - A portal for the ESS. 10th Annual European DDI User Conference (EDDI18), Berlin, Germany. Zenodo. https://doi.org/10.5281/zenodo.2530051

CESSDA (2019). Adapt your Data Management Plan - A list of Data Management Questions based on the Expert Tour Guide on Data Management. https://dmeg.cessda.eu/content/download/4302/48656/file/TTT_DO_DMPExpertGuide_v1.3.pdf (CC-BY)

Charles Beagrie Ltd. (2017). CESSDA SaW Cost-Benefit Advocacy Toolkit User Guide. Zenodo. http://doi.org/10.5281/zenodo.3662438

Sources

Committee on Forecasting Costs for Preserving and Promoting Access to Biomedical Data, Board on Mathematical Sciences and Analytics, Committee on Applied and Theoretical Statistics, Computer Science and Telecommunications Board, Board on Life Sciences, Board on Research Data and Information, Division on Engineering and Physical Sciences, Division on Earth and Life Studies, Policy and Global Affairs, & National Academies of Sciences, Engineering, and Medicine. (2020). Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs (S. 25639). National Academies Press. https://doi.org/10.17226/25639

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. Nature, 578(7796), 491–491. https://doi.org/10.1038/d41586-020-00505-7

Orten, H., Norland, S., & Butt, S. (2018). The Questionnaire Design and Documentation Tool (QDDT) - a DDI based tool for assisting questionnaire design teams in their work. 10th Annual European DDI User Conference (EDDI18), Berlin, Germany. Zenodo. https://doi.org/10.5281/zenodo.2530046

Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: The case of DANS. International Journal on Digital Libraries, 12(4), 195–214. https://doi.org/10.1007/s00799-012-0092-1

Service-Team Forschungsdaten der Uni Hannover und der TIB (2018). Wie lassen sich die Kosten für das Forschungsdatenmanagement abschätzen? Presentation.

Schultes, E. (2020). Challenging areas for the Early Career Researcher in Data Stewardship. International FAIR Convergence Symposium 2020. https://osf.io/yhu85 (CC-BY)

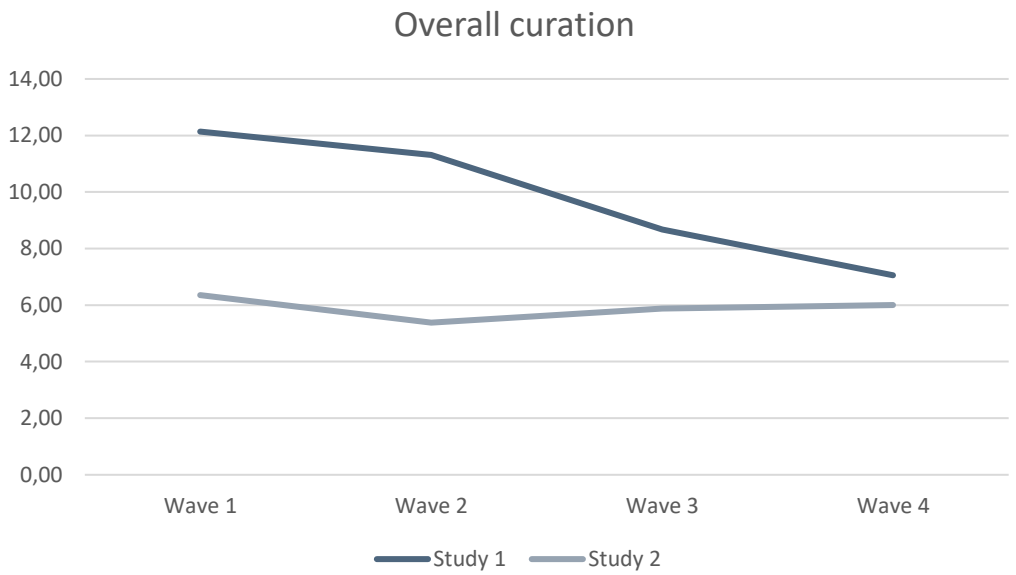
gesis Leibniz Institute for the Social Sciences

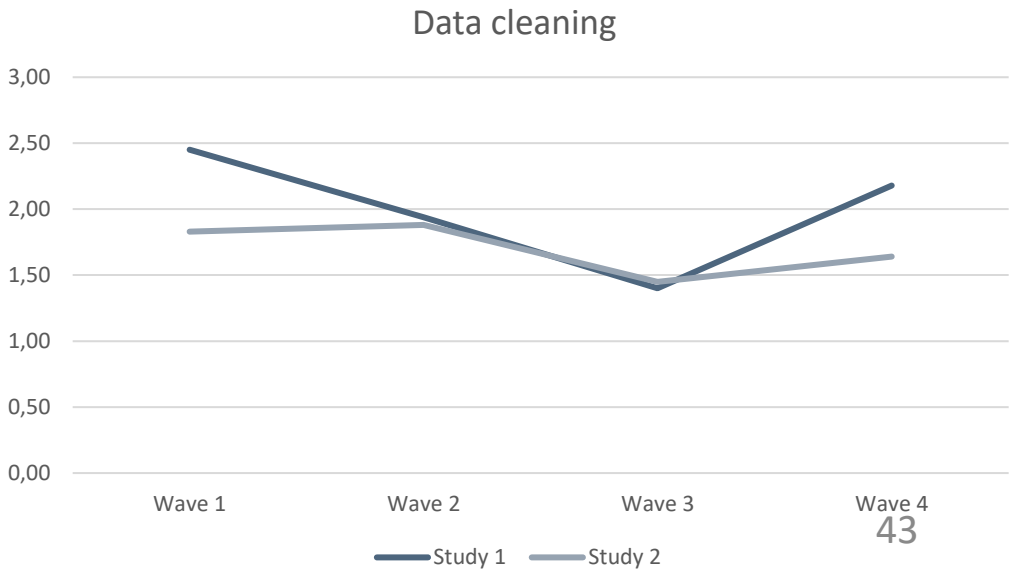
Sample description

dataset	year	number of variables	number of cases	number of questions	open-answer questions	variables affected by filters
Study1 - 2010	2010	288	4001	71	12	37
Study1 - 2012	2012	240	4000	79	11	71
Study1 - 2014	2014	301	4002	93	12	90
Study1 - 2016	2016	696	4002	114	41	150
Study2 - 2007	2007	639	10001	217	44	237
Study2 - 2009	2009	574	10000	241	43	244
Study2 - 2011	2011	737	10002	283	68	244
Study2 - 2013	2013	867	11501	324	87	215
Study3 - 2014	2014	428	4491	229	14	44
Study3 - 2016	2016	627	5012	270	22	77
average		539.7	6701.2	192.1	35.4	140.9
std. deviation		213.9	3207.7	93.9	26.3	86.8

Is there a learning effect?

- Curators can confirm this
- Codes and routines can be re-used
- We find a leanring effect only for study 1
- Except for data cleaning: wave 4 had many open answers
  - Number of cases play a role here
  - In total still time saving effect
- For study 2 the waves 3 and 4 differed greatly, codes could not be re-used





Curation tasks

Data cleaning tasks	Activities within task
Missing values	– Check for consistent use and labelling of missing values
	– Correct deviant use and inconsistent labelling of missing values
Wild codes	– Search for wild codes and outliers in the data
	– Correct wild codes and outliers in the data
	– Document changes made or wild codes and outliers themselves if not corrected
Skip pattern structure	– Search for filters in the questionnaire
	– Check for irregularities in the skip pattern structure
	– Correct irregularities
	– Document changes made
Data protection	– Search open-answer questions for information that allows re-identification
	– Pseudonymize or delete information
	– Document changes made
Variable and value labels	– Check for consistent use of variable names and labels
	– Check for typos
	– Harmonize names and labels
	– Shorten labels to accommodate for statistical programs' restrictions
Sorting variables	– Sort variables within one wave/dataset according to the questionnaire
	– Harmonize order of variables within one study with multiple waves/datasets

Perry, A. and Netscher, S. (2022), "Measuring the time spent on data curation", *Journal of Documentation*, Vol. 78 No. 7, pp. 282-304. https://doi.org/10.1108/JD-08-2021-0167

Curation tasks

Data documentation tasks	Activities within task
Questionnaire documentation	– Compare and link variables and values to the underlying questions and answer options in the questionnaire – Harmonize variable names and questions – Document linkage between variables and questions and between values and answer categories – Examine skip pattern in the questionnaire and transfer it to the variable documentation – Identify item batteries and their coding in the data
Variable documentation	– Document field notes on particular variables – Combine and finalize reports on the various checks, findings and corrections made during data cleaning
Study description	– Document study's metadata – Process cover-page for data documentation
Codebook (print)	– Combine documentation in a single variable report in PDF format – Run final checks on data and data documentation

Perry, A. and Netscher, S. (2022), "Measuring the time spent on data curation", *Journal of Documentation*, Vol. 78 No. 7, pp. 282-304. https://doi.org/10.1108/JD-08-2021-0167