gesis Leibniz Institute for the Social Sciences

DDP BILDUNG Domain Data Protocols for Educational Research

Measuring the Costs of Research Data Management

Dr. Anja Perry and Dr. Sebastian Netscher

ESRA Conference | Data management and processing| 23.07.2021

Mitglied der Leibniz-Gemeinschaft

This work is licensed under Creative Commons Attribution 4.0 International License.

Research Data Management means 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-**

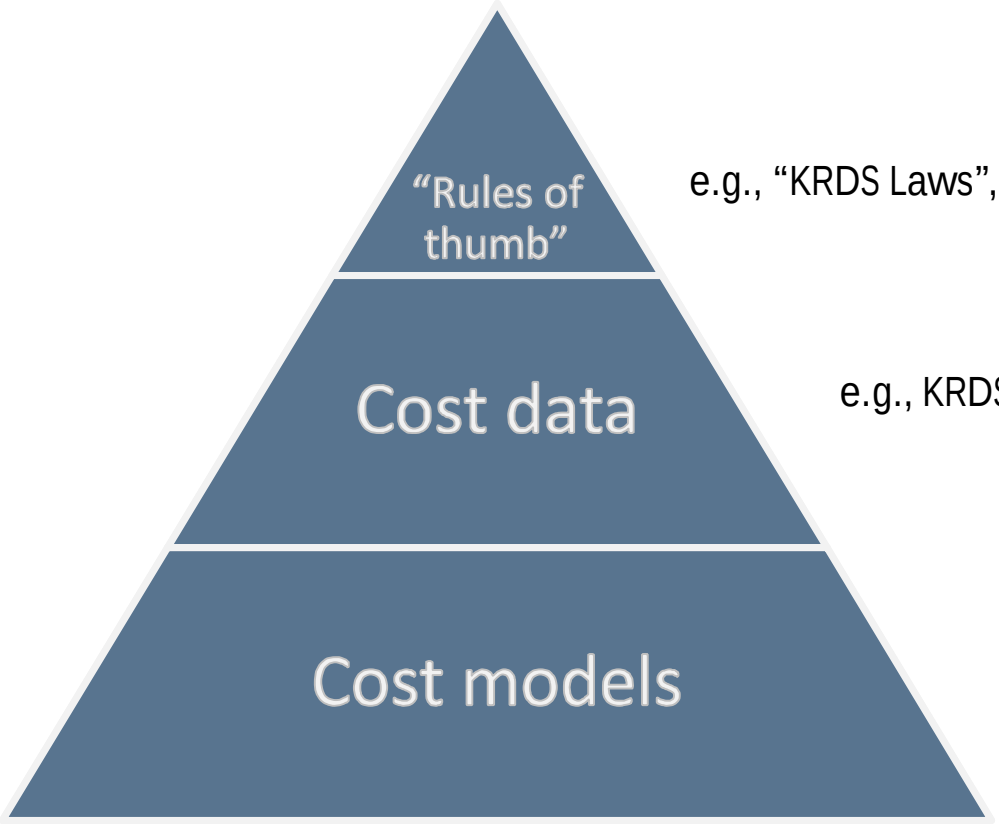
Data stewardship offers excellent returns on investment.

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

Previous work 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

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

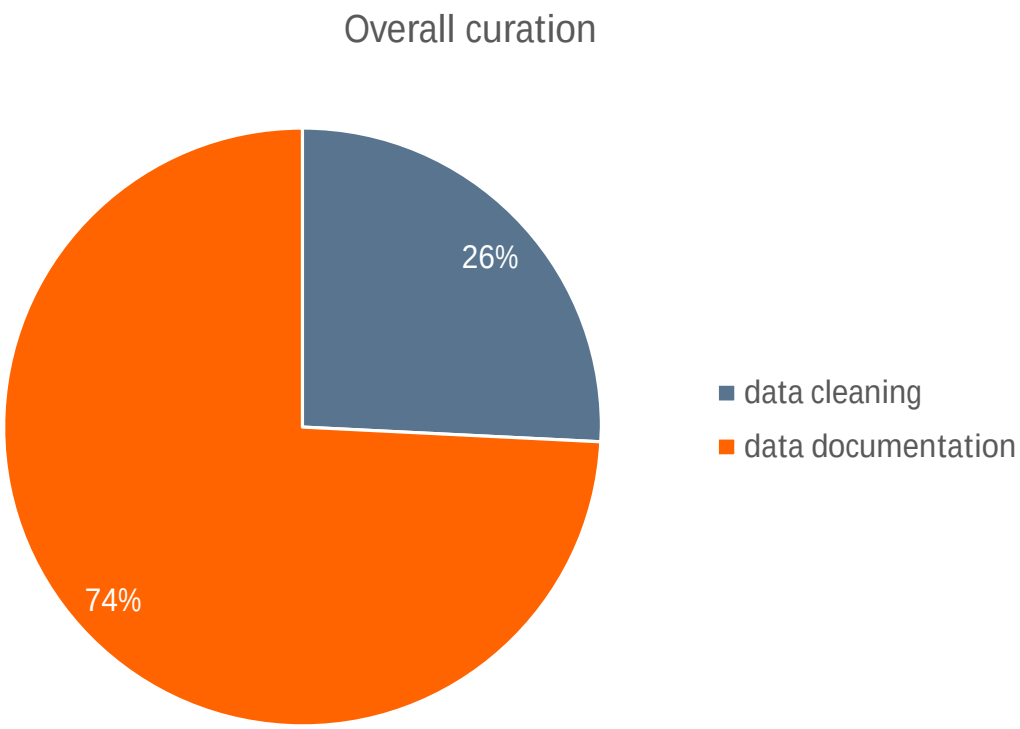
Our approach

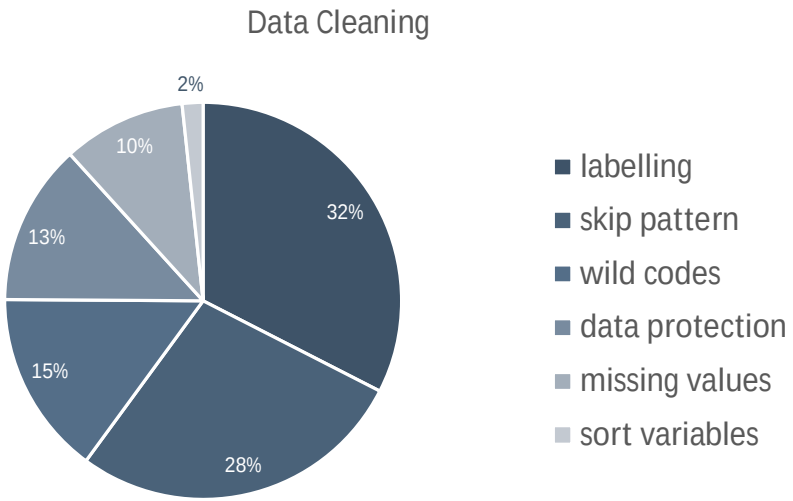
- Times were recorded for curation tasks (Dec. 2016 – Sept. 2017)
  - 3 multi wave studies
  - 10 datasets
  - Documentation times not recorded for 1 study (2 datasets)

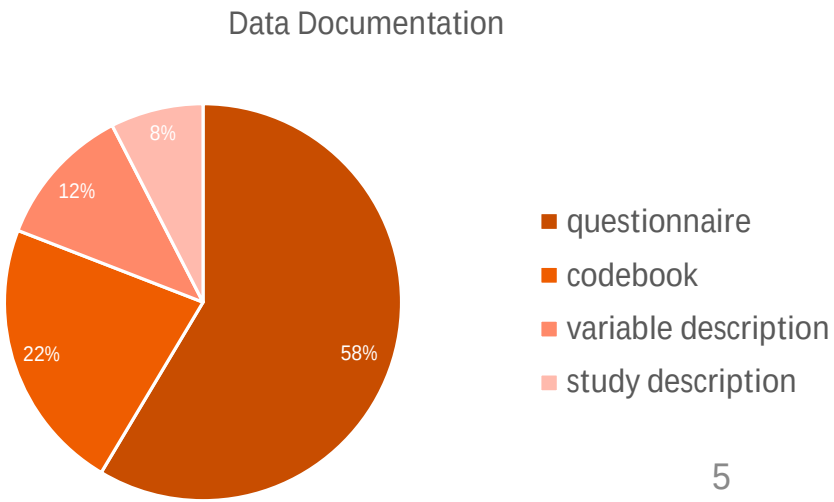
- Focus group interview (Feb. 2021)

- We look at four main drivers (Committee on Forecasting Costs… (2020)):
  - # of variables
  - # of questions
  - # of open answer questions
  - # of questions affected by filters

Time spent on different activities







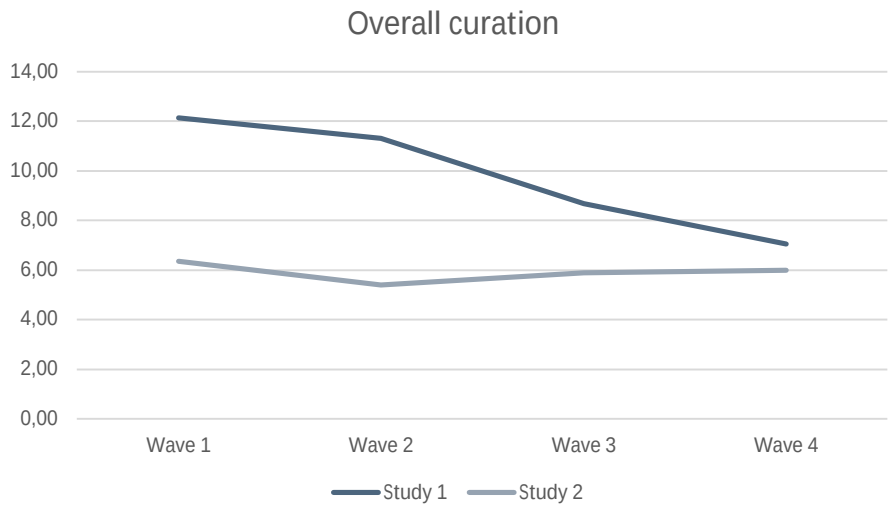
Average time spent on one study: 63h (≈ 8 days)

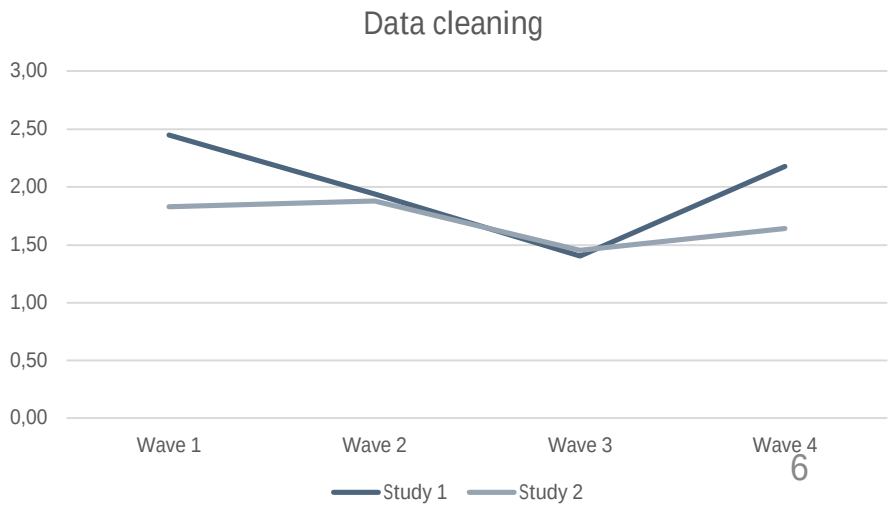
gesis Leibniz Institute for the Social Sciences

DDP BILDUNG Domain Data Protocols for Educational Research

Is there a learning effect?

- Curators confirm that they need less time for follow-up studies
- Reuse of code and routines
- We find it only for Study 1 overall
- Last wave of Study 1 needed much more time for cleaning due to open answer questions – number of cases matter
- For Study 2, wave 3 and 4 differed greatly – no reuse of former code





gesis Leibniz Institute for the Social Sciences

Correlations

- Main drivers:
  - # variables
  - # open answer questions
- All four characteristics correlated with cleaning, esp.
  - skip patterns
  - labels
- Interdependencies:
  - initial checks and immediate corrections
  - labeling and wild codes
  - questionnaire documentation partly contains the remaining steps

	# variables	# questions	# open answer questions	# variables affected by filter
overall curation time	0.8774		0.8071	
	(0.004)		(0.015)	
cleaning	0.9156	0.6261	0.7858	0.6813
	(0.001)	(0.097)	(0.021)	(0.063)
missing values	0.7584	0.8400	0.8857	
	(0.029)	(0.009)	(0.003)	
wild codes				
skip pattern structure	0.7673	0.9531	0.8464	0.8671
	(0.026)	(0.000)	(0.008)	(0.005)
data protection				
variable and value labels	0.8331	0.8326	0.7877	0.8943
	(0.010)	(0.010)	(0.020)	(0.003)
sorting variables				
documentation	0.7593		0.7351	
	(0.030)		(0.038)	
questionnaire documentation	0.7084		0.7988	
	(0.049)		(0.017)	
variables documentation				
study description				
codebook (print)				

Note. Displayed are correlation coefficient and p-value below; insignificant correlations not displayed; correlations with $r < 0.75$ in grey; number of observation n = 8.

Further findings

- Not number of filters are decisive but their complexity and their documentation
- Low data quality increases time spent on cleaning
  - Inconsistent missing scheme
  - Incorrect filter criteria
- Routines can make curators blind for unexpected errors
- DDI standard and tools are most helpful for documentation
  - Provides a predefined structure
  - Metadata can be reused
- Documentation and cleaning split between curators which increases need for documentation
- When mistakes are found during documentation, data have to undergo cleaning again, corrections are then, again, documented

Lessons learned

- Documentation is the cost driver in our case (75% of overall curation time)
  → the use of documentation standards saves time
- Investments in planning and field work have pay-offs in data curation
  → how complex does it need to be?
  → make use of field instrument documentation
  → get “clean” data from the field
- Re-use of earlier work (code etc.) saves time, BUT it can “hide” unexpected problems in the data

Outlook

- Work package 5 in the DDP project on RDM costs
- We will make use of the tasks identified in the DDP project
- Develop rules of thumb for research projects on how to budget these tasks
  - This analysis is the first step
  - We will also speak with experts on RDM costs
  - And we will use the outcomes and expertise of DDP

gesis Leibniz-Institut für Sozialwissenschaften

DDP BILDUNG Domain Data Protocols for Educational Research

Thank you!

www.ddp-bildung.org

How to cite:
Perry, Anja & Netscher, Sebastian (2021): Measuring the Costs of Research Data Management. Presented at the European Survey Research Association Biennial Conference 2021 (ESRA21), virtual: Zenodo. http://doi.org/10.5281/zenodo.5126773

Contact

Dr. Anja Perry

anja.perry@gesis.org

Tel: +49 221 47694-464

References

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

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

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

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

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

Study characteristics

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

Curation tasks

Data cleaning

1. Checking and correcting
   - missing value schemes (identify unlabeled missing values and deviant codes, harmonize)
   - skip patterns (identify filters and correct)
   - wild codes (identify and correct)
2. Data protection (check for identifiable persons and delete information)
3. Labeling of variables and values (harmonize, rename, shorten, assign labels)
4. Sorting variables (sort according to documentation available)

Data documentation

1. Questionnaire documentation (documentation of questions by variables)
2. Variable description (documentation of all information related to variables (surveyed and derived), incl. corrections made on variable level)
3. Study description (documentation of all information related to the study , incl. corrections made on study level)
4. Codebook (create final pdf variable report)