Measuring the time spent on data curation

Anja Perry and Sebastian Netscher
GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

Received 27 August 2021
Revised 2 December 2021
Accepted 12 December 2021

Abstract

Purpose – Budgeting data curation tasks in research projects is difficult. In this paper, we investigate the time spent on data curation, more specifically on cleaning and documenting quantitative data for data sharing. We develop recommendations on cost factors in research data management.

Design/methodology/approach – We make use of a pilot study conducted at the GESIS Data Archive for the Social Sciences in Germany between December 2016 and September 2017. During this period, data curators at GESIS - Leibniz Institute for the Social Sciences documented their working hours while cleaning and documenting data from ten quantitative survey studies. We analyse recorded times and discuss with the data curators involved in this work to identify and examine important cost factors in data curation, that is aspects that increase hours spent and factors that lead to a reduction of their work.

Findings – We identify two major drivers of time spent on data curation: The size of the data and personal information contained in the data. Learning effects can occur when data are similar, that is when they contain same variables. Important interdependencies exist between individual tasks in data curation and in connection with certain data characteristics.

Originality/value – The different tasks of data curation, time spent on them and interdependencies between individual steps in curation have so far not been analysed.

Keywords Data curation, Digital curation, Curation tasks, Research data management, Data sharing

Paper type Case study

Introduction

Barend Mons states in his 2020 Nature article that around 5% of the overall research budget should go towards data stewardship, that is research data management (RDM) tasks (Mons, 2020). Indeed, FAIR data (Wilkinson *et al.*, 2016), data sharing and, thus, RDM are increasingly important. Research funders, like the European Research Council, now demand data sharing and carefully planned RDM in terms of specified data management plans (European Commission, 2019; European Research Council, 2019). But even if data sharing appears to be free, for example if data is shared via a data repositorium free of charge, there are costs involved. These are, at least, costs for data cleaning and documentation to prepare data for reuse. Such costs are often not anticipated when planning the research project (National Academies of Sciences, Engineering, and Medicine, 2020).

However, such lump sums as suggested by Mons ignore differences across research projects. For instance, an estimate for biological databases suggests that only 0.088% of the overall research project costs go towards data curation (Karp, 2016). In contrast, whenever the requested RDM budgets are higher than 5%, because the projects in planning are more

The authors thank the three curators for their participation in the focus group discussion and their very valuable input. This work was funded by the German Federal Ministry of Education and Research as part of the "DDP Bildung – Domain Data Protocols for Education Research" project (www.ddp-bildung.org). Grant number: 16QK01A.

Journal of Documentation
Vol. 78 No. 7, 2022
pp. 282-304
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-08-2021-0167

complex, this may cause suspicion on the funder's side and the necessary funds may not be granted. Therefore, a more careful evaluation of such costs is needed.

Research on RDM costs is still in its infancy and precise statements about which RDM tasks are eligible for funding are rare (Donaldson and Ensberg, 2018). For example, the guidelines by UK Research and Innovation state that "any element of research data management may be included as a directly incurred cost" (UKRI, 2015). Clear and more precise recommendations on how to budget RDM tasks would help to accommodate for these tasks when applying for research funding and to devote appropriate time and effort towards RDM tasks during the project.

With this paper, we investigate the time spent on specific RDM tasks. These are data curation tasks, more specifically data cleaning and documentation of quantitative social science studies. We make use of a pilot study conducted between December 2016 and September 2017 at the *GESIS* - Leibniz Institute *for the Social Sciences* (from here on "GESIS"). For this pilot, our data curators documented their working times and tasks while cleaning and documenting data from ten quantitative datasets belonging to three different multi-wave survey studies. We analyse the data curators' records of their working times, make use of their working reports and interview them to better understand their workflows.

We contribute to existing research by analysing limited but very rare data on data curation tasks. Thereby, we identify factors that increase hours spent on data curation and factors that can save time. In addition, we obtain a list of individual curation tasks and feedback from curators on important interdependencies between these tasks and with important data characteristics. Various work on curation tasks already exists, for example, by Lee and Svilia (2017). We go beyond this work and provide more detailed information about data cleaning and data documentation tasks and how they depend upon one another. We provide insights for researchers and data managers initiating a survey project to better plan data curation tasks and organize data collection in a way that saves them time during data curation. Likewise, data infrastructures may profit from our findings when providing data services for researchers (German Research Foundation, 2021).

The paper is structured as follows: In section 2 we review existing literature and projects devoted to investigating and analysing RDM costs. We then describe the pilot project on determining curation times at GESIS and the ten datasets curated as well as our approach to examine curators' efforts in section 3. In section 4 we present the results of the analysis and in section 5 the outcome of a group discussion with our curators. We conclude with section 6.

Approaches to measuring RDM costs

Budgeting RDM tasks is difficult, as it covers a wide range of processes, strategies and measurements to manage data within a project as well as beyond. Not all tasks purely serve the purpose of data sharing. Even when researchers decide not to share their data, they need at least some basic documentation of the data to ensure its understandability and interpretability for their own purposes. In this article we solely focus on data curation aimed at preparing data for data sharing. In this theory section, after highlighting the importance of data curation, we give an overview of the literature on RDM and curation costs to identify the cost factors researched in this literature.

The role of data curation in data sharing

When analysing data sharing, we must distinguish between tasks relevant to the research itself and ensuring good scientific practice from those tasks devoted to sharing data beyond the research project. In this context, Klar and Enke (2013) as well as Treloar and co-authors (Treloar and Harboe-Ree, 2017; Treloar and Klump, 2019) suggest a domain model, grouping

data management tasks into four domains: private, group, persistent and public. The first two domains concern tasks by one researcher or by a group of researchers in a project, that is efforts that need to be undertaken to reach the original research aim. The latter two are tasks devoted to making data available beyond the project: persistent and public. In the following, we look at tasks to prepare data for sharing. Hence, we focus on the migration beyond the research project to provide insights in how to measure costs to clean and document data for the purpose of reuse by other researchers, that is the domains persistent and public.

Essential for this migration are data curation tasks that ensure information flows between the four different domains (Lafferty-Hess *et al.*, 2020). Digital curation has developed into a mature and stand-alone discipline over the last years. It encompasses digital preservation, data curation and information management and becomes more and more important in data driven sciences (Higgins, 2018; Koltay, 2015; Lee and Stvilia, 2017; Poole, 2016). It includes, among others, tasks such as data ingest, data checks and corrections, documenting data and the changes made and secure storage. The labour costs of digital curation, however, are often underestimated (National Research Council, 2015). For this paper, we narrow in on data cleaning and data documentation and go, for these two curation tasks, beyond the work by Lafferty-Hess *et al.* (2020) by analysing the individual steps involved.

Estimating RDM and curation costs in research

There have been attempts to investigate RDM and curation costs, most of them focusing on the costs of infrastructure to share data. Projects like Keeping Research Data Safe (KRDS, Beagrie *et al.*, 2008) [1], the 4C Project [2] (L'Hours *et al.*, 2014) or the DANS Cost Model for Digital Archiving (CMDA) (Palaiologk *et al.*, 2012) have built important groundwork in this field. They focus on requirements of data archiving and model costs that occur at data repositories. This important work raised awareness for archiving costs and created a business case for relevant stakeholders. In addition, some projects examine the costs of actual curation practices at data archives and repositories. For example, within the Keeping Research Data Safe 2 (KRDS2) project (Beagrie *et al.*, 2010), the costs for digital data ingest, that is obtaining and importing data into an archive, were calculated as 21.5% of the overall activity costs for UK Data Archive. A similar approach is the Curation Costs Exchange (CCEx) that compares archiving costs between institutions (Thirifays *et al.*, 2014). Finally, there are rules of thumb or observations of costs for digital archiving, that are often applied by small data archives with fewer resources (Beagrie, 2017).

However, these approaches focus on the costs of data archiving and preservation, but little on the costs of preparing data for archiving. And while it is of no less importance to implement good RDM in research projects and to allocate project funds for RDM tasks, there is little work so far that derives recommendations for allocating costs for RDM tasks in research projects.

At the time of writing this article, we are aware of two approaches to budget RDM costs in research projects. One approach is RDM checklists. The most prominent tool here for researchers is the Costing Tool developed by UK Data Service (2015). It provides a checklist containing 18 RDM tasks that researchers can use to identify tasks applying to their project and recommendations on how to estimate and plan their costs. The final step of estimating actual costs for each activity is left to the researchers. While it cannot provide actual costs for RDM, the costing tool is a great resource to consider necessary RDM tasks and raise awareness for possible working times and costs connected with them. Similarly, the German EWIG project developed workflows for long-term preservation specifically of Earth sciences data. They list 23 RDM cost factors along the research data lifecycle. For each cost factor further explanations are provided that help researchers to understand particular cost factors (Bertelmann *et al.*, 2014). Utrecht University in the Netherlands goes one step further and lists

actual budget proposals, in addition to only naming cost factors. For example, for transcribing and simultaneously anonymizing data, they suggest calculating an hour for each five-minute fragment of an interview. This is, however, a very general estimate and does not take specific aspects of the data into account, which affect the efforts spent on this task (Utrecht University, n.d.).

A second approach is to calculate specific budget recommendations for RDM tasks. These tools are rare and are tied to specific institutions. They always consider the resources available at this specific institution. The Service Team RDM at Hannover University (2018) in Germany base their recommendation on project size, the types of data (e.g. texts, scans, spreadsheets, automatically generated data), costs for hard- and software and costs that occur when publishing data. They provide estimates for resources needed in form of personnel working hours and costs for infrastructure, such as server costs. Similar recommendations are provided by 4TU.Research Data at TU Delft (2020) in the Netherlands. Their recommendations are based on file size, nature of the data (personal data, commercially sensitive data) and project size. Recommendations for small vs large projects range between 0.1 and 1.0 of a data manager's full-time equivalent (FTE) (4TU.ResearchData, TU Delft, 2020) and 0.25 and 0.5 FTE (Service-Team Forschungsdaten der Uni Hannover und der TIB, 2018). These recommendations, however, cannot be generalized for other projects outside the respective institution. But they reflect that certain data characteristics are in fact cost drivers in RDM and, more specifically, in data curation.

Identifying cost drivers in data sharing

While recommendations about RDM costs are often not very data specific or specific only for certain institutions, the literature and tools available identify certain cost drivers. In this section we describe these cost drivers and how we expect them to affect the curators' work analysed in this paper.

Doing so, we can focus on two key publications that identify certain data characteristics as cost drivers. According to the National Academies of Sciences, Engineering, and Medicine, (2020) the size of data, its complexity and existence of different data types, the amount of documentation needed to understand the data, data quality and access control affect the costs of managing research data for reuse. Likewise, Bingert *et al.* (2019) recommend considering data heterogeneity and complexity as well as data quality and personal information included in the data as cost drivers, each increasing the efforts to clean and document data.

Data heterogeneity means that data can be of different types, such as data matrixes, text files or transcripts, video or audio records etc. Data can also be collected from different sources, for example online, face-to-face, by recording, observing and so on. Accordingly, different types of data require different strategies of cleaning and documenting. However, data from one data type and using just one mode can be heterogenous across surveys, too. Different types of variables and missing value schemes require curators to carefully check for such differences and adjust their strategies and work routines. We therefore expect a learning curve when cleaning and documenting different datasets from the same survey. That is, higher initial workload when curating data from the first wave of one survey and a reduction of workload with each new wave of the same survey.

The *size of data* covers the number of variables, the number of questions in the questionnaire as well as the number of cases or observations included in the data. Many variables in a data set increases the amount of time to check for errors in the data as well as to document the variables. Likewise, the combination of information about respondents, revealed by different variables, may increase the chances of respondents' re-identification and thus requires additional efforts in data cleaning, that is pseudonymization, and data documentation (Corti *et al.*, 2014). Many observations increase the time worked on the data

when, for example open answer questions are involved that can possibly include sensitive information. Consequently, we expect that efforts to clean and document larger datasets are higher than for smaller datasets.

Complexity of data involves filters used in the questionnaire and the resulting skip pattern structure in the dataset. Filter questions and skip patterns are error prone. Checking them is therefore a standard procedure when preparing datasets for reuse to ensure that filters are applied correctly and that the distinction of the different subsets of cases in the data is correct. Consequently, we expect an increase of time spent on data cleaning when more elaborate filters and skip patterns are involved.

Personal information, such as names or addresses, *or even sensitive information* on personal health, religious beliefs, political and sexual orientation and union membership (European Parliament and Council of the European Union, 2018) in the data require additional checks to carefully control for information that allow re-identifying participants. Detailed questions on respondents' background and open-answer questions containing personal or sensitive information increase the likelihood of (indirect) re-identification by combining information included in different variables (see above and Corti *et al.*, 2014). Such data requires pseudonymization, which is very labour intensive (Roertgen *et al.*, 2019). We expect that an increasing number of variables and an increasing number of open answer questions in the data increase the efforts to clean and document data.

Research methodology

Data collection

From December 2016 to September 2017 the *Data Archive for the Social Sciences*, a department at the *GESIS – Leibniz Institute for the Social Sciences*, conducted an internal project to assess its efforts in data curation. The major goal of this project was to gain knowledge about tasks performed for each dataset and the times spent on them. This allowed better planning of the tasks and started a decision-making process regarding how much curation is devoted to which type of study. As a result of the project, we split the archiving process into different tracks (BASIC, PLUS and PREMIUM) for different data, ranging from longtail data with low reuse potential to data with very high reuse potential. That way we achieved a second major goal of the project, to implement a modularized service structure for data publication. These services are aimed at researchers who want to outsource RDM tasks to prepare their data for sharing [3]. Hence, the tasks analysed in this paper were completed by professional curators at GESIS, but these are tasks which are typically done in a research project.

To better understand the tasks, workflows and efforts of data curation, the project conducted one pilot test with a long-term data provider. The data provider chose four, mainly multi-wave study programs with a total of 11 datasets they wanted to publish. They were willing to follow a different acquisition process than usual for these studies to accommodate for this pilot, meaning additional communication with GESIS to agree on a set of services to be performed and closer cooperation with the acquisition team when problems with the data occurred during the curation process.

We chose three out of the four studies provided in this pilot test for our analyses. The curation process for the fourth study, consisting of a single dataset, was not documented well and their task descriptions cannot be compared to the other three studies. Hence, the final sample consists of three multi-wave studies with a total of 10 datasets.

During the pilot test, data curators at GESIS (working in teams of two) documented the tasks and working hours needed to curate the datasets. During an initial check, they noted down all necessary steps for data documentation and the estimated time they need for these steps. While then working on the data, they carefully documented all completed steps, all

problems that occurred and have not been foreseen during the initial check, how they resolved these problems and the hours actually spent on each task.

Curation tasks can be split into data cleaning and data documentation and were conducted according to the data provider's request. Each task consists of several individual activities (see Table 1). Data cleaning ensures data quality by checking and correcting for errors. Data documentation is the creation of metadata to make the data findable and understandable to users. The curators created metadata that comply with the Data Documentation Initiative (DDI) standard (DDI Alliance, 2021a) and to be used for registering the data to receive a digital object identifier (DOI; Koch *et al.*, 2017).

Conducting these tasks, curators followed predefined instructions described in a GESIS-internal handbook for data curation. Tasks are, thus, standardized, and efforts spent on each

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
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

Table 1.
Data curation tasks

of the ten datasets are, at least to some degree, comparable with each other [4]. In addition to time recording, the curators also documented specifics in the data that affected their working times.

Description of datasets included in the analysis

Data at hand include ten datasets from three different multi-wave studies. For all three studies, repeated representative samples of the German adult population were interviewed by telephone or online interview. The questions included basic demographics (such as age, gender, education), knowledge about certain topics and their attitude towards them. Within each study, a core set of questions is repeated in consecutive waves. In addition, each dataset includes a set of unique, non-repeated, questions that typically focus on the specific topics of that particular wave. The data sets' download numbers range in the medium field compared to all data available at GESIS.

As shown in Table 2, study 1 comprises four datasets, conducted between 2010 and 2016. On average, these four datasets include about 381 variables and replies from 4,001 respondents. About 87 variables are affected by filters. The questionnaires contain on average 89 questions with 19 open answer questions.

The four waves of study 2 were conducted between 2007 and 2013. On average, these datasets have 704 variables and 10.376 observations, and 235 variables affected by filters. The questionnaires include on average 266 questions with an average of 61 open-answer questions [5].

Study 3 consists of two datasets conducted in 2014 and 2016. These two datasets include on average 528 variables and 4,752 respondents. On average 61 variables are affected by filters. The questionnaires have on average 259 questions, including 18 open answer questions. For study 3, we do not have information on time spent on data documentation as it has not been recorded. However, we decided to keep the two datasets from study 3 in our analysis for additional information on data cleaning.

Data analyses and focus group discussion with data curators

We make use of the time records to examine patterns in the time spent on data cleaning and data documentation tasks. We focus on four data characteristics: the number of variables, the number of questions, the number of open answer questions and the number of variables

Table 2.
Study characteristics

Dataset	Year	Number of variables	Number of cases	Number of questions	Open-answer questions	Variables affected by filters
Study 1 – wave 1	2010	288	4,001	71	12	37
Study 1 – wave 2	2012	240	4,000	79	11	71
Study 1 – wave 3	2014	301	4,002	93	12	90
Study 1 – wave 4	2016	696	4,002	114	41	150
Study 2 – wave 1	2007	639	10,001	217	44	237
Study 2 – wave 2	2009	574	10,000	241	43	244
Study 2 – wave 3	2011	737	10,002	283	68	244
Study 2 – wave 4	2013	867	11,501	324	87	215
Study 3 – wave 1	2014	428	4,491	229	14	44
Study 3 – wave 2	2016	627	5,012	270	22	77
Average		539.70	6701.20	192.10	35.40	140.90
Std. dev.		213.89	3207.65	93.97	26.28	86.80

affected by filters. In terms of curation tasks, we concentrate on different data cleaning and data documentation tasks (compare Table 1).

Due to various aspects the following analyses are rather a case study than an empirical analysis: First, we only have a small number of eight to ten waves, that is datasets, which limits statistical analyses and the interpretation of correlation coefficients and their p-values. Second, most of the datasets were in a good shape and pre-documented when delivered, reducing data curation efforts. Third, tasks were carried out by highly experienced curators who are well-trained in cleaning and documenting data. This implies time-saving effects compared to researchers in a research project. Fourth, working hours are usually not recorded at GESIS and the curators may have had problems to correctly record their hours, especially when tasks were conducted in parallel. We therefore expect that time recordings may sometimes be inaccurate.

Consequently, we use correlations, trends and patterns in the quantitative data only as frameworks for further investigations. When interpreting these frameworks, we first used written comments made by the curators while working on the data. In addition, we conducted a focus group discussion with the curators to understand their workflows and work routines and to strengthen our interpretation of patterns found in the data. For the focus group discussion we invited all three curators who worked on studies 1 and 2. Due to the Covid-19 pandemic the discussion was held online on February 17th, 2021. We prepared our questions for the group based on our quantitative analyses and structured the discussion along the curators' workflows and our findings for each step. Hence, we first asked questions about the preparation of their work, then about data cleaning, followed by data documentation. Finally, we asked for general feedback on the process and their work in general. We asked for the curators' consent to use their recorded working times and their feedback during the discussion for this paper. However, due to the sensitivity of the data, we did not record the group discussion.

In sum, the quantitative results presented in this paper can only be interpreted with great care and should not be generalized right away. Rather, they can be only be interpreted together with further findings and explanation based on written comments and on the focus group discussion with our curators.

Analysing curators' time spent to curate data

We start by looking at the overall time spent on curating and then focus on the time for data cleaning and documentation, separately. For overall curation and data documentation we examine descriptive statistics for eight datasets from study 1 and 2. For data cleaning we can make use of time recordings for two more study 3 datasets. Because our data are very limited, we use this descriptive analysis as a framework for the following focus group discussion (section 5).

Overall curation time

Curators spent on average 3801.38 min (std. dev. 981.05 min), or about 63 h, for curating one dataset. About 75% of this time, that is 2820.75 min (std. dev. 636.08 min), or 47 h, were spent on data documentation, compared to an average of 980.63 min (std. dev. 412.66 min), or 16 h, for data cleaning (Table 3).

It took almost twice as long to clean and document a single variable in the first two waves of study 1 (11.7 min on average, std. dev. 0.6 min) than in the remaining six datasets of study 1 and 2 (6.6 min on average, std. dev. 1.2 min). The time spent to curate data from the study 2 waves is generally lower than for those of study 1.

Time spent on each dataset within study 1 decreases. It drops from 12.14 min to 7.05 min per variable in study 1. We, however, do not find this decrease for study 2. Here, the time spent on curation stays relatively constant between 6.25 min and 5.38 min per variable across all four waves (Figure 1).

Correlation between the number of variables in a dataset and the overall time spent on data curation is quite strong ($r = 0.877$; $p = 0.004$; Table 4). Figure 2 illustrates this pattern by plotting the time spent and the number of variables for each of the eight datasets examined. While study 2 waves follow the trend of longer curation times with an increasing number of variables, study 1 waves cluster in the bottom left corner, except for wave 4 on which we will focus latter on.

Table 3.
Time spent on overall data curation, cleaning and documentation, in minutes

	Overall		Cleaning		Documentation	
Dataset	In total	Per variable	In total	Per variable	In total	Per variable
Study 1 – wave 1	3,495	12.14	705	2.45	2,790	9.69
Study 1 – wave 2	2,715	11.31	465	1.94	2,250	9.38
Study 1 – wave 3	2,610	8.67	420	1.40	2,190	7.28
Study 1 – wave 4	4,905	7.05	1,515	2.18	3,390	4.87
Study 2 – wave 1	4,056	6.35	1,170	1.83	2,886	4.52
Study 2 – wave 2	3,090	5.38	1,080	1.88	2010	3.50
Study 2 – wave 3	4,335	5.88	1,065	1.45	3,270	4.44
Study 2 – wave 4	5,205	6.00	1,425	1.64	380	4.36
Study 3 – wave 1			1,680	3.93		
Study 3 – wave 2			2,535	4.04		
Average ($n = 8$)	3801.38	7.85	980.63	1.84	2820.75	6.00
Std. dev.	981.05	2.60	412.66	0.36	636.08	2.43
Proportion	100.00%		25.00%		75.00%	
Average ($n = 10$)			1206.00	2.27		
Std. dev.			631.51	0.96		

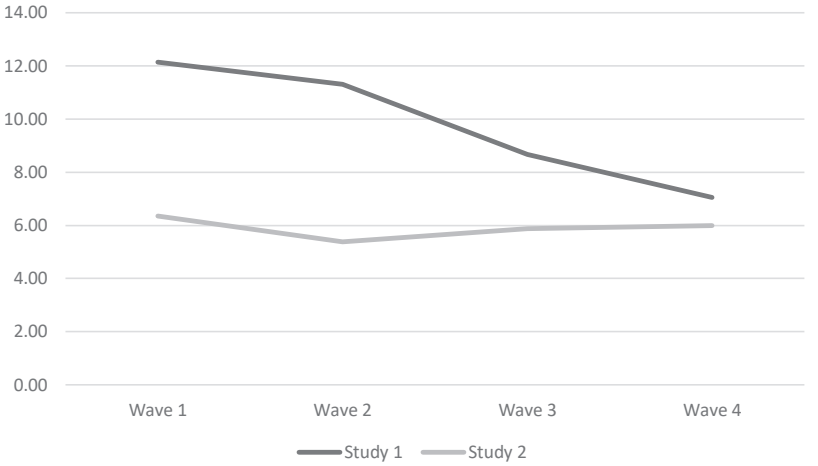

Figure 1.
Time spent on overall curation per variable for each wave of study 1 and study 2

	# Variables	# Questions	# Open answer questions	# Variables affected by filter
Overall curation time	0.8774** (0.004)	0.5765 (0.136)	0.8071* (0.015)	0.4918 (0.216)
Cleaning	0.9156** (0.001)	0.6261 (0.097)	0.7858* (0.021)	0.6813 (0.063)
Missing values	0.7584* (0.029)	0.8400** (0.009)	0.8857** (0.003)	0.5878 (0.125)
Wild codes	−0.3788 (0.355)	−0.5082 (0.199)	−0.3364 (0.415)	−0.7635* (0.028)
Skip pattern structure	0.7673* (0.026)	0.9531*** (0.000)	0.8464** (0.008)	0.8671** (0.005)
Data protection	0.2776 (0.506)	−0.2405 (0.566)	0.0329 (0.938)	−0.0334 (0.937)
Variable and value labels	0.8331* (0.010)	0.8326* (0.010)	0.7877* (0.020)	0.8943** (0.003)
Sorting variables	−0.4352 (0.281)	−0.4301 (0.288)	−0.4027 (0.323)	−0.5864 (0.127)
Documentation	0.7593* (0.029)	0.4830 (0.225)	0.7351* (0.039)	0.3167 (0.445)
Questionnaire documentation	0.7084* (0.049)	0.6053 (0.112)	0.7988* (0.017)	0.3167 (0.445)
Variable documentation	0.4031 (0.322)	0.5523 (0.156)	0.4078 (0.316)	0.5796 (0.132)
Study description	−0.2028 (0.630)	−0.2704 (0.517)	−0.3087 (0.457)	−0.1358 (0.749)
Codebook (print)	0.1580 (0.709)	−0.3696 (0.368)	−0.0927 (0.827)	−0.1640 (0.698)

Note(s): Displayed are correlation coefficient and p-value in brackets; number of observation $n = 8$; ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table 4.
Correlations between data set characteristics and data curation activities for datasets of studies 1 and 2

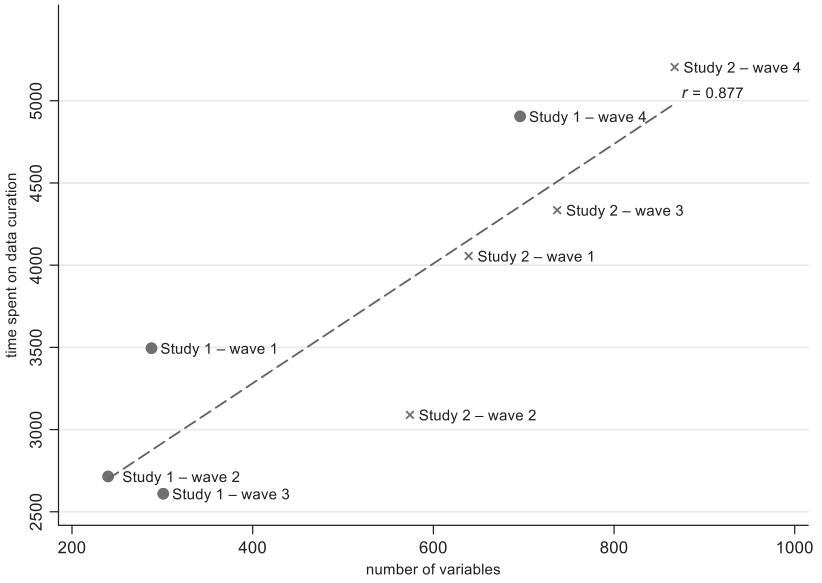

Figure 2.
Data curation by numbers of variables

In addition to the number of variables in the data, the number of open-answer questions in the questionnaires increases the overall time spent to curate data ($r = 0.807$; $p = 0.015$). The more variables contain free-text fields, the more time curators spent to clean and document the datasets. In contrast, neither the total number of questions in the questionnaires ($r = 0.577$; $p = 0.136$) nor the number of variables affected by filters ($r = 0.492$; $p = 0.216$) seem to impact the time spent on data curation (Table 4).

In summary, the main drivers of time spent on overall curation are the number of variables and the number of open answer questions. For study 1 we also observe decreasing curation times with each follow-up wave.

Data cleaning

When cleaning the data, curators spent most of the time on labelling (32%, $n = 8$) and checking skip pattern structures (26%, $n = 8$). Taking the two additional datasets from study 3 into account, this finding remains stable. But with study 3 the pattern tends to shift slightly towards relatively more time spent on missing values 14% ($n = 10$) vs 11% ($n = 8$). This indicates additional efforts to control and correct the missing value schemata in the two datasets of study 3. The smallest proportion and fastest task per variable and relatively constant for each dataset is variable sorting (about 2%, Table 5).

When looking at time spent on data cleaning across waves of studies 1 and 2, we find that times decrease for study 1 between waves 1 and 3 but increases again for wave 4. This is true for the total time spent on data cleaning as well as for almost all individual cleaning tasks and the respective times spent per variable. Overall time spent on cleaning drops from 2.45 min per variable in wave 1 to 1.40 min per variable in wave 3 and then goes up to 2.18 min per variable in wave 4. In contrast, we do not see decreasing times for study 2 where time spent per variable varies between 1.88 min and 1.45 min across the four waves (Figure 3).

As seen above, in comparison to the remaining three datasets of study 1, wave 4 stands out. It is much larger, containing more than twice as many variables (696, Table 2). In addition, wave 4 requires a long time to check for data protection issues (855 min or 14.25 h, Table 5). The dataset includes 41 open-answer questions, around 4 times more than in other waves (Table 2), with some of them potentially containing sensitive information. Accordingly, for all 4,002 observations in wave 4 each of these open-answer questions needed to be checked manually and its information needed to be either deleted or at least pseudonymized. Curators therefore needed more than 30-times longer for data protection for this dataset than for any other dataset in the analysis (0.42 h on average, Table 5) [6].

Study 1 also stands out regarding checks for wild codes. Even though study 1 waves have fewer variables than those of studies 2 and 3. Per variable, curators spent 0.58 min on checking wild codes for study 1, more than the overall average of 0.31 min (Table 5).

For study 3, curators needed longer time for data cleaning. It took 3.99 min for a single variable, which is more than twice as long as for the remaining eight datasets (1.84 min, on average, std. dev. = 0.36 min). Checking for missing values required more time for study 3 than for all other studies (1.06 min for study 3 compared to 0.19 min for studies 1 and 2). Moreover, wave 2 of study 3 had a more complex skip pattern structure, including a section for specific subgroups, and filters did not always contain correct filter criteria. The skip pattern structure thus demands additional checks, and its correction was more difficult. This characteristic is also apparent in Table 5 where time spent on the skip patterns (1,260 min) is more than three times higher than for the average time spent on all datasets (372 min, std. dev. = 355.55 min).

As shown in Table 4, the number of variables ($r = 0.916$, $p = 0.001$) and the number of open answer questions ($r = 0.7858$, $p = 0.021$) correlate with the time spent on data cleaning for studies 1 and 2. When looking at the single tasks of data cleaning, we find positive correlations of all data characteristics examined here with checking the skip pattern, and with labelling variables. However, taking all ten datasets into account (i.e. including study 3), none of our dataset characteristics correlate with the time spent on overall data cleaning. We find a positive correlation between the number of questions and the skip pattern structure ($r = 0.689$, $p = 0.027$) and between the number of questions and labelling ($r = 0.74$, $p = 0.015$; Table 6).

Dataset	Cleaning		Missing values		Wild codes		Skip pattern		Data protection		Labels		Sort variables	
	In total	Per variable	In total	Per variable	In total	Per variable	In total	Per variable	In total	Per variable	In total	Per variable	In total	Per variable
Study 1 – wave 1	705	2.45	60	0.21	270	0.94	120	0.42	15	0.05	210	0.73	30	0.10
Study 1 – wave 2	465	1.94	60	0.25	165	0.69	90	0.38	15	0.06	120	0.50	15	0.06
Study 1 – wave 3	420	1.40	60	0.20	135	0.45	60	0.20	30	0.10	120	0.40	15	0.05
Study 1 – wave 4	1,515	2.18	60	0.09	165	0.24	120	0.17	855	1.23	300	0.43	15	0.02
Study 2 – wave 1	1,170	1.83	120	0.19	90	0.14	375	0.59	30	0.05	540	0.85	15	0.02
Study 2 – wave 2	1,080	1.88	60	0.10	90	0.16	480	0.84	30	0.05	405	0.71	15	0.03
Study 2 – wave 3	1,065	1.45	150	0.20	90	0.12	375	0.51	30	0.04	405	0.55	15	0.02
Study 2 – wave 4	1,425	1.64	210	0.24	180	0.21	540	0.62	30	0.03	450	0.52	15	0.02
Study 3 – wave 1	1,680	3.93	540	1.26	60	0.14	300	0.70	15	0.04	750	1.75	15	0.04
Study 3 – wave 2	2,535	4.04	540	0.86	30	0.05	1,260	2.01	30	0.05	660	1.05	15	0.02
Average ($n = 8$)	980.63	1.84	97.50	0.19	148.13	0.37	270.00	0.46	129.38	0.20	318.75	0.58	16.88	0.04
Std. dev.	412.66	0.36	57.26	0.06	61.81	0.30	192.93	0.22	293.28	0.42	156.86	0.16	5.30	0.03
Proportion	100.00		10.54		19.24		26.02		9.90		32.15		2.15	
Average ($n = 10$)	1206.00	2.27	186.00	0.36	127.50	0.31	372.00	0.64	108.00	0.17	396.00	0.75	16.50	0.04
Std. dev.	631.51	0.96	193.29	0.38	70.09	0.29	355.51	0.52	262.56	0.37	214.73	0.41	4.74	0.03
Proportion	100.00		13.77		15.87		27.57		8.13		32.79		1.87	

Table 5.
Descriptive overview of activities on data cleaning in minutes

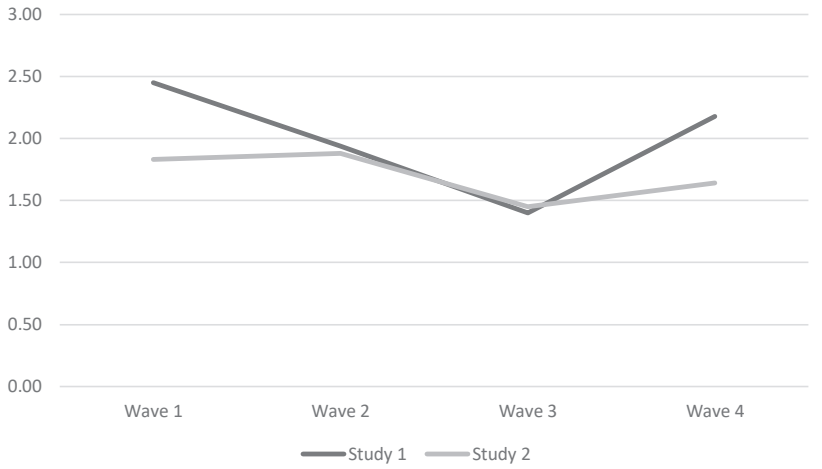

Figure 3.
Time spent on data cleaning per variable for each wave of study 1 and study 2

Hence, for data cleaning we find a strong impact of the number of variables and the number of open answer questions. All four data characteristics correlate with checking and correcting the skip pattern structure and with variable and value labelling. We also observe decreasing times across waves 1 to 3 of study 1 and increased efforts on data protection when open answer questions are involved.

Data documentation

More than half of the documentation time is spent on documenting the questionnaire (57% on average). Various steps within this process include harmonizing variable names and questions in the questionnaire, documenting the linkage between variables and questions as well as between variable values and answer categories, examining the skip pattern structure in the questionnaire and transferring it to the variable documentation and identifying question batteries and their coding in the data. Questionnaire documentation thus overlaps with other documentation tasks, such as variable documentation and adding further information to the documentation, for example corrections made during data cleaning.

	# Variables	# Questions	# Open answer questions	# Variables affected by filter
Cleaning	0.5619 (0.091)	0.6146 (0.059)	0.1835 (0.612)	0.0021 (0.995)
Missing values	0.1642 (0.650)	0.5173 (0.126)	−0.1206 (0.740)	−0.3379 (0.334)
Wild codes	−0.2908 (0.415)	−0.5821 (0.078)	−0.0353 (0.923)	−0.2218 (0.538)
Skip pattern structure	0.4795 (0.161)	0.6895* (0.027)	0.2131 (0.554)	0.1220 (0.737)
Data protection	0.2748 (0.442)	−0.2768 (0.439)	0.0911 (0.802)	0.0564 (0.877)
Variable and value labels	0.4789 (0.161)	0.7388* (0.015)	0.2024 (0.575)	0.1211 (0.739)
Sorting variables	−0.4135 (0.235)	−0.4528 (0.189)	−0.3128 (0.379)	−0.4206 (0.226)

Note(s): Displayed are correlation coefficient and p-value in brackets; number of observation $n = 10$

Table 6.
Correlations between data set characteristics and data cleaning activities for datasets of studies 1, 2 and 3

The smallest proportion of time spent on data documentation is creating the study description (8%).

Just like for data cleaning, time spent for documentation is highest for the first dataset in study 1 and decreases for follow-up waves (Table 7). When looking at time spent per variable, wave 4 is not an outlier here as it follows the pattern of decreasing time spent. Again, we do not find this pattern for study 2. More than 3 min per variable are spent on questionnaire documentation for waves 3 and 4 of study 2, while times spent per variable are well below 3 min for the first two waves of this study.

There is little variation in the total time spent for study description and for processing the codebook, except of wave 4 of study 1. In comparison to all other datasets, it took more than twice as long to process a codebook for this wave.

Correlation between the time spent on data documentation and the number of variables is once again quite strong ($r = 0.759$, $p = 0.029$, Table 4). And again, the time spent on documentation increases with the number of open answer questions ($r = 0.735$; $p = 0.039$). This relationship is driven by the significantly positive relationship of both variables with the time spent on questionnaire documentation ($r = 0.708$, $p = 0.049$ for number of variables and $r = 0.799$, $p = 0.017$ for number of open questions). However, neither the number of questions in the questionnaire nor the number of variables affected by filters correlate with time spent on data documentation or its individual tasks.

To sum up, we find a strong correlation of questionnaire documentation with the number of variables and with the number of open questions while all other tasks are not affected by dataset characteristics. The times spent on documenting study 1 decrease with each wave and, contrary to data cleaning, this pattern also holds for wave 4.

Discussing the findings with data curators

To improve our understanding of the patterns found, we discussed our findings with three data curators, who processed the datasets from studies 1 and 2 [7]. Thereby, we focused on the workflows behind the cleaning and documentation tasks. In this section we recap the important findings and back these up by the curators' feedback. We start the discussion with the overall time spent on curation and proceed with discussing the learning effect, major drivers of curation times, interdependencies between work tasks and a general discussion about data quality.

Overall curation time

Our curators clarify that their first step is to get familiar with the data and to do some initial data checks. This takes about 1–2 h for each dataset. They start with creating frequency tables of all variables and cross checking them with the questionnaire to get a general understanding of the data as well as to ensure that the data corresponds to the questionnaire and that data are complete. The amount of time spent to get familiar with the data varies depends on how large and complex a dataset is. In general, the data looked at here are not very complex [8]. Curators do not aim to familiarize themselves with the project itself, only with the data.

During this initial step they already correct for minor errors such as typos in variable and value labels. Hence, the time recordings do not entirely reflect each individual task. The different steps of checking and correcting the data can therefore not be easily separated.

Our curators confirm that most of the time is spent on data documentation. They also state that data always need to be documented. Data cleaning becomes an effort only if data need to be harmonized, for example for trend-files, or if problems in the data or deviations from the questionnaire occur. Data quality in the datasets considered here is generally high and data

Table 7.
Descriptive overview of activities on data documentation in minutes

Dataset	Documentation		Questionnaire		Variable description		Study description		Codebook	
	In total	Per variable	In total	Per variable	In total	Per variable	In total	Per variable	In total	Per variable
Study 1 – wave 1	2,790	9.69	1,515	5.26	360	1.25	300	1.04	615	2.14
Study 1 – wave 2	2,250	9.38	1,290	5.38	210	0.88	195	0.81	555	2.31
Study 1 – wave 3	2,190	7.28	1,245	4.14	210	0.70	180	0.60	555	1.84
Study 1 – wave 4	3,390	4.87	1770	2.54	210	0.30	195	0.28	1,215	1.75
Study 2 – wave 1	2,886	4.52	1,386	2.17	540	0.85	390	0.61	570	0.89
Study 2 – wave 2	2010	3.50	990	1.72	360	0.63	150	0.26	510	0.89
Study 2 – wave 3	3,270	4.44	2,250	3.05	360	0.49	150	0.20	510	0.69
Study 2 – wave 4	380	4.36	2,760	3.18	360	0.42	150	0.17	510	0.59
Average	2820.75	6.00	1650.75	3.43	326.25	0.69	213.75	0.50	630.00	1.39
Std. dev.	636.08	2.43	588.42	1.31	113.88	0.29	86.63	0.32	239.20	0.69
Proportion	100.00		57.47		11.90		7.87		22.76	

cleaning took much less time (only 1/4 of the total curation time) than documenting the data. The curators further explain that data cleaning needs more preparation than documentation. Moreover, data cleaning affects data documentation and vice versa. Changes made during data cleaning must be documented. Data documentation then serves as a first check of the final data. If errors are found while documenting the data, the data are passed back to data cleaning to re-run the process of correcting errors, documenting these changes and recreating the documentation files.

In comparison to their usual working routines with other datasets coming to GESIS, our curators agreed that datasets from studies 1 and 2 were easier to handle. The studies were, in comparison, less complex, for example regarding the skip pattern structures. In addition, the datasets were delivered in a good shape. Variables and values were well-labelled, and the missing schemes were consistent. This was different, however, for study 3, for which incorrect and inconsistent missing labels and redundant variables were documented in the curators' notes. The filters used in the study 3 questionnaires were more complex. As we can see above, including study 3 in our analyses changes the correlation coefficients which reflects the additional work caused by higher complexity and, at the same time, lower quality of the data.

Learning effect

Curators confirm our findings that learning effects occur for both, data cleaning and data documentation. It takes considerably longer to familiarize with a new study than to curate follow-up waves. For follow-ups, curators rather only look for deviations from previous data. However, the learning effect depends on by whom and how long ago the previous study was curated, and on how similar the waves within a study are, for example in terms of the number of repeated questions. Further, to increase efficiency, curators reuse scripts from earlier datasets whenever possible. For example, they adapted syntax files for study 1 and 2, where core sets of variables were repeatedly employed. Likewise, checking routines on the missing schemata and the skip pattern structures were reused due to a high consistency across datasets within each study.

Also, learning a software or tool once can lead to learning effects that even affects the work on studies beyond this pilot project. For example, during the pilot project one curator developed a Stata ado that can now also be used for other data.

When documenting a dataset, curators reuse existing metadata, that is variable and value labels, whenever possible. In this context, our curators state that employing the DDI standard (DDI Alliance, 2021a) is very time saving. It provides a clear structure and controlled vocabulary, which would otherwise need to be developed and implemented.

However, if deviances between the actual and earlier data are large, reuse of earlier work is restricted or even impossible. This was the case with study 1. There, wave 4 contained additional questions that were also mostly open answer questions. Likewise, we do not find a learning effect for study 2 as the last two waves of this study differed greatly and pre-existing syntax files could not be reused.

Finally, reusing pre-existing syntaxes can be risky. The curators discussed the fact that scripts and routines from previous datasets can be misleading and foster ignorance towards new and unexpected problems that may occur.

Main drivers of curation times

In our descriptive analysis, we identify the number of variables and the number of open answer questions as the main drivers for time spent on curation tasks. For data cleaning we find positive relations also between time spent and the number of questions and the number of variables affected by filters.

Generally, our curators confirm that the *number of variables* affect the time spent on cleaning and documenting data and it is the most crucial characteristic for them. When cleaning data, curators first check variable and value labels and correct typos and irregularities. At the same time, they check for the completeness of data, that is complete sets of variables, the missing schema, as well as for unlabelled or extraordinary values, potentially indicating wild codes. These tasks are directly related to the number of variables in a dataset.

They also confirm that the *number of open answer questions* can affect the time spent when requiring to check open answers for each observation. It is almost the only circumstance when the number of observations affect data cleaning times. Most often, the number of observations in the datasets have no effect on the time spent for data curation [9].

The *number of questions affected by filters* is in positive relation to data cleaning, more specifically to the skip pattern structure (a direct result of the filters used in the questionnaire) and to labelling. The discussion with the curators, however, revealed that filters in the datasets examined here are relatively simple. Only complex filters would increase the time spent on cleaning in a meaningful way. However, if filters are not well-documented, it becomes difficult to check for correct skip patterns. Filters are detected by word searches in the questionnaire and the documented criteria are used for checking the data.

The *number of questions* is different from the number of variables in the dataset. Often, the number of variables in the resulting dataset exceeds the number of questions in the questionnaire, due to variables which are derived from original questions after fieldwork and due to additional technical variables, such as weights and disposition codes. Hence, the number of variables is often a more direct indicator for time spent on data cleaning while the number of questions, that is the length of a questionnaire, relates to the time spent on documentation. The curators also state that the type of questions is decisive. Item batteries, for example are questions that are very similar to one another (often just one word or one group of words differ). These can be cleaned and documented in bulk, which saves a lot of time compared to working on each question individually.

The main drivers for curation times, number of variables and number of open answer questions, are not correlated with the study description and with creating a codebook. These two documentation tasks generally do not depend on the size of the data but vary with further specifics on study level.

Interdependencies between curation tasks

There are important interdependencies between curation tasks that became apparent when discussing our findings with the curators. First, many steps within cleaning and documentation are done in parallel. Hence, it is difficult to separate between individual tasks. The correlations we present in the analysis section may therefore either be underestimated or positive relations are subsumed under one task while additional tasks were done at the same time. This is, for example the case for wild codes for which we don't find positive correlations with dataset characteristics. Wild codes are either found while checking all frequencies (as part of the initial checks right at the beginning) or by examining unlabelled values. Hence, this work is subsumed almost entirely under the task "labelling". The task "labelling" includes checks and corrections of completeness, typos and length reduction to make it suitable for standard statistics software. Once value labels are checked and corrected, checking and correcting wild codes are done in a very fast and efficient way. At this point, one curator states that running checks in parallel is restricted by the data's complexity and quality. Once data are too complex or too messy, data cleaning must be done step by step to ensure high-quality data. However, this was not the case for the datasets examined here.

Very similar interdependencies occur in data documentation, even more so when using a DDI documentation tool. The first step to document the questionnaire is to compare and link

variables and values to the underlying questions and answer options in the questionnaire. This already includes tasks related to variable documentation. The time reported for the questionnaire documentation, thus, already includes time that would have been spent on variable documentation at a later stage.

Because of these workflows, our curators are not surprised that we cannot find stronger correlations between dataset characteristics and individual tasks. They even state that the distinction between cleaning and documenting is probably the only meaningful distinction that can be reasonably made when analysing their time recordings.

It is also important to note that data documentation and data cleaning were split between different curators, regarding their expertise. This is because the data provider requested data documentation according to DDI codebook standard (DDI Alliance, 2021b). To ensure this, a GESIS-internal DDI tool, called *Dataset Documentation Manager* (DSDM) (Zenk-Möltgen, 2006), was used, which requires specific skills and expertise. To increase efficiency, a curator highly capable using this tool was assigned to these tasks while data cleaning was done by others. However, this split is not necessarily recommended for research projects as it demands a lot of documentation of cleaning and documentation tasks, a process that is naturally done at a data archive.

Data quality and data curation

Generally, datasets of study 1 and 2 were in very good shape when delivered. This also results in a relatively low share of time spent on data cleaning (25%) of the total time spent on curation. The curators confirm that there was little to be done in terms of cleaning. At the same time, their standard work routines could be applied, and they did not need to deviate often. This may be the reason why times spent on data cleaning also varies less with the number of variables than time spent on documentation [10]. Study 3 was not of such a good quality and the two datasets of study 3 differ greatly. This is one reason why curators spent more time cleaning the data and why we cannot observe a learning effect from wave 1 to 2. The limited data we have available for study 3 must therefore be interpreted with even greater care.

At the end of our discussion, we asked our curators what makes a messy dataset, that is what are factors that increase their time spent on data cleaning and documentation tasks. We summarized their comments in Table 8.

Also, unexpected changes in the data collection process that are not documented make data curation difficult. They therefore highlight the relevance of having a well-documented planning of the studies and well-documented field work. The better the documentation of survey instruments and the collection of data – also of unexpected changes that occurred during the data collection process – the easier it is to understand the data and, thus, to document and clean it. It is helpful for curators to have access to the coding of the questionnaire to easily spot possible coding errors.

Factors increasing time spent on data curation
Unlabelled data or missing labels
Variable names that are not meaningful for curators (and users) outside the research project
Arbitrary missing schema
Coding errors
A differing order of variables compared to the questionnaire
A general lack of structure in the data and its documentation

Table 8.
Data quality and data curation

Conclusions

Until now, information on curation tasks and cost drivers of data curation are very limited and rare. In this paper, we make use of time records at GESIS during a pilot project. Our data is based on the cleaning of ten datasets and the documentation of eight datasets. Due to this very limited database, we cannot infer general characteristics of data curation efforts. We therefore relate our descriptive results to actual curation processes by discussing them with the curators involved in this work.

We find two very strong patterns: The *size of the data* and *personal information* potentially contained in the data are very important drivers of time spent on curation overall as well as its two components, data cleaning and data documentation.

Heterogeneity in data plays a role and we see strikingly in how learning effects, by reusing previous work routines and codes, are apparent for waves 1 to 3 within study 1. We cannot find these for study 2, due to large differences between waves.

The studies analysed for this paper did not vary much regarding *complexity*. However, our curators confirm that complexity plays a role, for example when complex filters are used. For comparison the curators also named more complex studies where, due to their complexity, many decisions are made during data collection and data curation that need to be carefully documented and communicated to others involved.

The discussion with the curators revealed important interdependencies between curation tasks. For example, certain tasks were done in parallel and certain tasks needed to be completed first to complete others in an efficient way. This makes it more difficult to interpret the recorded time spent on each task. However, it also highlights the importance of data quality and data documentation throughout the project, for example the importance of having a fully labelled dataset. The curators also emphasize that the use of the DDI standard is very helpful as it provides a general structure used for all data coming to GESIS instead of defining this structure again and again for every new dataset.

We contribute to existing literature on RDM and the efforts connected with it in several ways. First, we analyse the recorded time spent on data curation tasks in the social sciences which is despite its limitations, to our knowledge, very unique data. We deepen our understanding gained by these analyses by discussing them with the curators who worked with the data and who recorded the times spent on this work. While the cost drivers examined here are identified in previous literature, we can back these findings up with data and with further insights into curation processes. Knowing these cost drivers is crucial for data collection projects in survey research in the social sciences and related fields. Second, an outcome of the original project for which times were recorded are a detailed list of curation tasks that can serve as a workflow template for planning curation tasks in research projects. Third, we also identify interdependencies between tasks and how the time spent on each task is related to data quality and the complexity of the data examined. Both help research projects to accommodate for them early in the project and save time and effort in the long run. Fourth, we aim to raise awareness to carefully plan data curation tasks in a data collection project by providing rough estimates on how much time needs to be spent on data curation. Very often this work is unnoticed, and its costs are not considered when writing grant proposals for new projects. And fifth, we can state that the work and the funding of research data infrastructure is highly important in guiding and helping researchers to make data reusable.

Our paper has several limitations. First, we have a very small sample size for our descriptive analyses and curators at GESIS are not used to documenting their working times, so those times may therefore not be entirely accurate. The data are, thus, to be interpreted with great care and only in relation with the focus group discussion that followed. We urge readers to not overinterpret our quantitative results, but rather, as we did, use it as a framework for possible further investigations.

Second, we also need to keep in mind that our curators are experts in cleaning and documenting data. Hence, this work was done much faster and scale effects could be realized, for example by using the appropriate tools and documentation standards. It is very likely that individual researchers will spend more time on these tasks (National Academies of Sciences, Engineering, and Medicine, 2020).

Third, the datasets in these analyses were already well prepared for publication and, compared to other data coming to GESIS, of very good quality. Not much time had to be spent on data cleaning. The amount of time spent and the ratio between data cleaning and data documentation will likely differ when data are of lower quality.

Further research will be needed on the topic of RDM efforts. We can only provide a first insight into one pilot project with one data provider. We need a better understanding of the identified cost drivers in various settings. This is especially true for data of lower quality or more complex data than the datasets considered here. Also, levels of curation and their associated effort should be considered (see e.g. ICPSR, 2020) to shed light on differences in time spent on RDM between data expected to be highly used and data of limited reuse potential.

Notes

1. https://www.beagrie.com/krds.php
2. https://www.4cproject.eu/
3. https://www.gesis.org/datenservices/home
4. GESIS archived the data and made them available to scientific users for research purposes, providing data in SPSS and Stata format along with the questionnaire and the variable report. Additional data cleaning and documentation tasks were done which specifically served the purpose of archiving data at GESIS and are not relevant for data sharing in general. We therefore did not include them in our analyses.
5. A preceding study 2 wave had been processed at GESIS before the pilot project started. Time spent on cleaning and documenting this dataset had not been recorded.
6. When excluding data protection checks from the analysis, cleaning wave 4 took 660 min (1,515 min–855 min = 660 min, or 11 h) and, thus, less than the average time spent on other datasets (1,172 min or about 19:30 h).
7. For additional data on data cleaning for study 3 we relied on comments made during documentation by the respective curators of this study.
8. An example for more complex data is the German Longitudinal Election Study (https://www.gesis.org/en/elections-home/gles). There, a lot of decisions have to be made during data collection and data cleaning that need to be carefully documented.
9. The only other ways that the number of observations may have an effect is that there can be more deviations in the dataset when more cases are involved, and that software needs longer to process when this number is very large.
10. We also find this in the data: Bartlett's equal-variances test: $\chi^2 = 13.7873$; $p = 0.000$.

References

4TU.ResearchData, TU Delft (2020), "Data management costing tool", available at: https://zingtree.com/host.php?style=buttons&tree_id=511095771&persist_names=Restart&persist_node_ids=1&start_node=1&start_tree=511095771 (accessed 27 August 2021).

Beagrie, C. (2017), "CESSDA SaW costs factsheet", doi: 10.18448/16.0003.

Beagrie, N., Chruszcz, J. and Lavoie, B. (2008), "Keeping research data safe - a cost model and guidance for UK universities", Final report, available at: https://www.webarchive.org.uk/wayback/archive/20140615221657/http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf (accessed 27 August 2021).

Beagrie, N., Lavoie, B. and Woollard, M. (2010), "Keeping research data safe 2", Final report, available at: https://www.webarchive.org.uk/wayback/archive/20140615221405/http:/www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf (accessed 27 August 2021).

Bertelmann, R., Gebauer, P., Hasler, T., Kirchner, I., Peters-Kottig, W., Razum, M., Recker, A., Ulbricht, D. and van Gasselt, S. (2014), "Einstieg ins Forschungsdatenmanagement in den Geowissenschaften", Potsdam, available at: https://gfzpublic.gfz-potsdam.de/rest/items/item_749901_8/component/file_749904/content (accessed 27 August 2021).

Bingert, S., Engelhardt, C. and Kusch, H. (2019), "Handlungsempfehlungen zu Forschungsdatenmanagement und -infrastruktur an Hochschulstandorten", Göttingen Research Online/Data, doi: 10.25625/PAYCKB.

Corti, L., Van den Eynden, V., Bishop, L. and Woollard, M. (2014), *Managing and Sharing Research Data: A Guide to Good Practice*, Sage, London.

DDI Alliance (2021a), "Document, discover and interoperate - the website of the DDI alliance", available at: https://ddialliance.org/ (accessed 6 August 2021).

DDI Alliance (2021b), "DDI codebook 2.5", DDI-Codebook 2.5, available at: https://ddialliance.org/Specification/DDI-Codebook/2.5/ (accessed 28 April 2021).

Donaldson, M. and Ensberg, V. (2018), "How to ensure that the costs of data management activities are budgeted in grant proposals?", Open Working, Blog, available at: https://openworking.wordpress.com/2018/03/09/how-to-ensure-that-the-costs-of-data-management-activities-are-budgeted-in-grant-proposals/ (accessed 8 January 2021).

European Commission (2019), "H2020 programme. AGA - annotated model grant agreement", 26 June, available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf (accessed 27 August 2021).

European Parliament and Council of the European Union (2018), "General data protection regulation 2016/678", available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed 27 August 2021).

European Research Council (2019), "Open research data and data management plans - information for ERC grantees", European Commission, available at: https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf (accessed 27 August 2021).

German Research Foundation (2021), "Handling of research data", Handling of Research Data - Information on the Resources Available, available at: https://www.dfg.de/en/research_funding/principles_dfg_funding/research_data/resources_available/index.html (accessed 5 August 2021).

Higgins, S. (2018), "Digital curation: the development of a discipline within information science", *Journal of Documentation*, Vol. 74 No. 6, pp. 1318-1338.

ICPSR (2020), "ICPSR curation levels", available at: https://www.icpsr.umich.edu/files/datamanagement/icpsr-curation-levels.pdf (accessed 12 November 2021).

Karp, P.D. (2016), "How much does curation cost?", *Database*, Vol. 2016, doi: 10.1093/database/baw110.

Klar, J. and Enke, H. (2013), "Organisation und Struktur, DFG-Projekt RADIESCHEN - Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur", Bericht, doi: 10.2312/RADIESCHEN_005.

Koch, U., Akdeniz, E., Meichsner, J., Hausstein, B. and Harzenetter, K. (2017), *da|ra Metadata Schema - Documentation for the Publication and Citation of Social and Economic Data*, GESIS Papers, 2017-25. Version 4.0, GESIS Leibniz Institute for the Social Sciences, doi: 10.4232/10.mdsdoc.4.0.

Koltay, T. (2015), "Data literacy: in search of a name and identity", *Journal of Documentation*, Vol. 71 No. 2, pp. 401-415, doi: 10.1108/JD-02-2014-0026.

Lafferty-Hess, S., Rudder, J., Downey, M., Ivey, S., Darragh, J. and Kati, R. (2020), "Conceptualizing data curation activities within two academic libraries", *Journal of Librarianship and Scholarly Communication*, Vol. 8 No. 1, p. 2347, doi: 10.7710/2162-3309.2347.

Lee, D.J. and Stvilia, B. (2017), "Practices of research data curation in institutional repositories: a qualitative view from repository staff", *PLoS ONE*, Vol. 12 No. 3, pp. 1-44, doi: 10.1371/journal.pone.0173987.

L'Hours, H., Kejser, U.B., Johansen, K.H.E., Thirifays, A., Wang, D., Strodl, S., Ashley, K., Davidson, J., McCann, P., Krupp, J. and Grindley, N. (2014), "D3.2 cost concept model and gateway specification", Final report, Colchester, available at: https://www.4cproject.eu/documents/D3.2%20Cost%20Concept%20Model%20and%20Gateway%20Specification.pdf (accessed 27 August 2021).

Mons, B. (2020), "Invest 5% of research funds in ensuring data are reusable", *Nature*, Vol. 578 No. 7796, pp. 491-491, doi: 10.1038/d41586-020-00505-7.

National Academies of Sciences, Engineering and Medicine (2020), *Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*, National Academies Press, Washington, DC.

National Research Council (2015), *Preparing the Workforce for Digital Curation*, National Academies Press, Washington, DC, doi: 10.17226/18590.

Palaiologk, A.S., Economides, A.A., Tjalsma, H.D. and 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*, Vol. 12 No. 4, pp. 195-214, doi: 10.1007/s00799-012-0092-1.

Poole, A.H. (2016), "The conceptual landscape of digital curation", *Journal of Documentation*, Vol. 72 No. 5, pp. 961-986, doi: 10.1108/JD-10-2015-0123.

Roertgen, S., Kusch, H., Engelhardt, C., Bingert, S., Savin, V., Wang, Y., Rice, R., Elsenga, C., Stokes, P., Donaldson, M. and Ebert, B. (2019), "2019_05_28_Wang.pdf", *Workshop on Data Management Costs and Efforts*, 28 May 2019, Göttingen, doi: 10.25625/NTRUKA/KD3XHY.

Service-Team Forschungsdaten der Uni Hannover und der TIB (2018), "Wie lassen sich die Kosten für das Forschungsdatenmanagement abschätzen?", December 2018, available at: https://www.fdm.uni-hannover.de/fileadmin/fdm/Dokumente/200727_KalkulationFDMKosten.pdf (accessed 27 August 2021).

Thirifays, A., Sisu, D., Davidson, J., Haage, K., Faria, L., Grootveld, M., Stokes, P. and Middleton, S. (2014), "D3.3 curation costs Exchange framework, collaboration to clarify the costs of curation", Final report, available at: https://www.4cproject.eu/documents/4C%20-%20D3%203%20-%20Curation%20Costs%20Exchange%20Framework%20-%2031%20Oct%202014%20-V1.0.pdf (accessed 27 August 2021).

Treloar, A. and Harboe-Ree, C. (2017), "Data management and the curation continuum: how the Monash experience is informing repository relationships", *Presented at the VALA2008 Conference*, Monash University, Melbourne, pp. 1-12.

Treloar, A. and Klump, J. (2019), "Updating the data curation continuum", *International Journal of Digital Curation*, Vol. 14 No. 1, pp. 87-101, doi: 10.2218/ijdc.v14i1.643.

UK Data Service (2015), "UK data service - data management costing tool and checklist", UK Data Archive and University of Essex, available at: https://ukdataservice.ac.uk/media/622368/costingtool.pdf (accessed 27 August 2021).

UK Research and Innovation (2015), "Guidance on best practice in the management of research data", available at: https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-GuidanceBestPracticeManagementResearchData.pdf (accessed 12 November 2021).

Utrecht University (n.d.), "Costs of data management - research data management support", available at: https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management (accessed 27 August 2021).

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, M. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3 No. 1, doi: 10.1038/sdata.2016.18.

Zenk-Möltgen, W. (2006), "Dokumentation von Umfragedaten in Länder vergleichender Perspektive mithilfe des ZA Dataset Documentation Managers (DSDM)", *ZA-Information/Zentralarchiv Für Empirische Sozialforschung*, Vol. 59, pp. 159-170.

Corresponding author
Anja Perry can be contacted at: anja.perry@gesis.org