DATA SCIENCE JOURNAL

Pronk, TE. 2019. The Time Efficiency Gain in Sharing and Reuse of Research Data. *Data Science Journal,* 18: 10, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2019-010

RESEARCH PAPER

The Time Efficiency Gain in Sharing and Reuse of Research Data

Tessa E. Pronk[1,2]

[1] Utrecht University Library, Heidelberglaan 3, Utrecht, NL

[2] KWR watercycle research institute, Groningenhaven 7, Nieuwegein, NL

Tessa.Pronk@kwrwater.nl

Among the frequently stated benefits of sharing research data are time efficiency or increased productivity. The assumption is that reuse or secondary use of research data saves researchers time in not having to produce data for a publication themselves. This can make science more efficient and productive. However, if there is no reuse, time costs in making data available for reuse will have been made with no return on this investment. In this paper a mathematical model is used to calculate the break-even point for time spent sharing in a scientific community, versus time gain by reuse. This is done for several scenarios; from simple to complex datasets to share and reuse, and at different sharing rates. The results indicate that sharing research data can indeed cause an efficiency revenue for the scientific community. However, this is not a given in all modeled scenarios. The scientific community with the lowest reuse needed to reach a break-even point is one that has few sharing researchers and low time investments for sharing and reuse. This suggests it would be beneficial to have a critical selection of datasets that are worth the effort to prepare for reuse in other scientific studies. In addition, stimulating reuse of datasets in itself would be beneficial to increase efficiency in scientific communities.

Keywords: sharing; reuse; research data; secondary use; model; open science

Introduction

Sharing research data is generally regarded as yielding benefits for the scientific community, society as a whole, and as having potential benefits for the researchers that share their data (e.g. Ascoli, 2007; Kim, 2013; Pitt & Tang, 2013). One of the specific benefits of shared data for the scientific community is more efficient or productive science (e.g. Koslow, 2002; Figueiredo, 2017, Kolb et al., 2013). When research data are shared, they become available for reuse. This may be time-efficient for a reuser because the often time consuming task of data collection or generation is omitted. Researchers who have the option to reuse data rather than having to produce it, have net more time to produce scientific products or to start other scientific inquiries, leading to increased productivity in science. Curty et al. (2017) indeed found that perceived efficacy and efficiency of data reuse are strong predictors of reuse behavior, and perceived importance of data reuse corresponds to greater reuse. In addition, for a whole community the reuse and combining of data can bring about great efficiency leaps in terms of understanding complex systems such as in brain research (Koslow, 2002) and biodiversity (Enke et al., 2012). The secondary use of the shared data in such individual cases can be very helpful for both scientific community and individual reusers.

Besides from individual advantages that come with reuse, the overall time efficiency of sharing and consequent reuse is generally believed to be real. However there have not been many investigations to ascertain this assumption. Fry et al. (2008) in a report calculated the theoretical return on investment of a single dataset, in terms of the money invested. This calculation however doesn't take into account the fact that there is an ecosystem involved in which researchers may share and reuse datasets, and not all datasets will be reused. Beagrie (2014) did a study on return on (financial) investment for three repositories, viewed from the following perspectives: the number of users and downloads (including the willingness to pay), the estimated user community time efficiency, and the additional use based on user products made with data from the repositories. He found that all three repositories had a positive return on investment.

Although basing reuse on user counts and downloads will probably overestimate the actual reuse, this is still a positive message.

The question remains *how much* reuse will compensate for the time spent in the sharing effort in a scientific community. This sharing effort, as well as the effort in reusing data, can be large. Sharing for reuse implies that the raw data itself should be fully understood in terms of origin and context (Jirotka et al., 2005; Faniel et al., 2013) to prevent misinterpretation and inappropriate use (Zimmerman, 2007), should be thoroughly described, and of good quality (Enke et al., 2012), must be relevant, and compatible to existing standards (e.g. Zimmerman, 2008; Enke et al., 2012; Bruland et al., 2016; Sprague et al., 2017), and should be clean to be useful for new research. Reuse of an available dataset with the intention to perform a new scientific analysis as well is actually a complicated process (Kim and Yoon, 2017; Zimmerman, 2007). Finding and downloading data are only the first of many steps in this process. Downloading is followed by the often necessary appraisal of a dataset on completeness, trustworthiness, and appropriateness, as these characteristics are key factors for re users to reuse the dataset in the end as a basis for their scientific work (Faniel et al., 2015; Yoon, 2016; Kim and Yoon, 2017). If the dataset seems trustworthy, and there is the intention to use it (Kim and Yoon, 2017) the data may still need filtering, cleaning and may need to be integrated with other information, before it is fit for reuse purposes. The fact that a time investment is needed to reuse a dataset, means that less time profit is made at reuse. At each of these activities the choice can be made to discard the dataset after all and not reuse it. Expectedly, actual reuse of datasets is a fraction of the downloads and views of datasets (e.g. see data from Fear, 2013).

For the current study a theoretical approach is taken to find the break-even point in time invested in sharing and time-gain by reuse. We focus on the question how much reuse is needed in a scientific community to compensate for the investment of making data available for sharing. For a calculation of the time efficiency in sharing and reuse, the model of Pronk et al. (2015) is used. It was previously used for a game theoretic analysis on research data sharing. In that study the main focus was not on the quantity of reuse as this was a fixed parameter. For the current study the same model is used, and the reuse necessary for at least a break-even point in terms of efficiency is further quantified. As values for time investment and gains related to data sharing and reuse can differ between disciplines, different scenarios are explored. They range from simple datasets that are shared and reused with limited time investments to complex datasets that are time consuming to prepare for reuse. The scenarios explored for different rates of sharing researchers. These scenarios together give an indication of needed reuse, and can be used in further studies to assess if current reuse is enough to give the desired and often assumed efficiency boost to the scientific community.

Methods

The game theoretic model on research data sharing that is used in the current study is described in detail in the paper of Pronk et al. (2015). The mathematical model assumes steady state. In the model of Pronk et al. (2015a) impact was the measure of success for a scientific community. For the current paper we focus on productivity (papers per year) as a measure of efficiency. The model and code to produce the resulting visuals for the current study are written in R statistical language and can be found in the current data package (Pronk, 2018). For anyone who wishes to run the model via a user interface and calculate the break-even point for a wide range of parameter settings, please visit https://tessapronk.shinyapps.io/ReuseResearchDataMinimum/.

In short, in the model, the output of all scientists consists of published scientific papers with research data (any information that conclusions are based on) at the basis. This is a simplifying assumption, as scientists may have other output as well. Scientific papers cost time to write, and it takes time to produce or assemble a dataset (See **Table 1** for values used). A percentage of the scientific community shares their research data as reusable datasets. This effort costs additional time, decreasing their output potential. For some of the research leading to publications, a suitable existing research dataset is found. In that case, time is saved as no new dataset has to be produced. The more datasets are deposited (with more sharing researchers), the better the chance to find an appropriate dataset. This chance levels off at a very high availability of datasets. There is also a time cost at reuse, to appraise and adapt the dataset. However, overall more time is left to (partly) write an extra scientific publication, resulting in increased output.

In the current study four scenarios are simulated (see **Table 1**). In scenario A short time (one day) is required or taken to prepare the dataset for sharing. The dataset is easy to interpret and reuse, requiring a short time (one day) to get to know the set and prepare it for reuse (e.g. a simple, straightforward, or well-maintained dataset).

In scenario B the dataset is difficult to reuse. In this scenario the sharer does not go through the effort of preparing the data adequately (one day), but leaves it to the reuser to do so. In a real-world situation this

Table 1: Model scenarios A B, C, and D.

Scenario	Dataset Description	Time for sharing	Time for reuse	Time for producing a dataset	Time to write a paper	Dataset decay rate
A	Easy share Easy to reuse	1 day	1 day	73 days	47 days	10% per year
B	Easy to share Hard to reuse	1 day	15 day	73 days	47 days	10% per year
C	Hard to share Easy to reuse	15 days	1 day	73 days	47 days	10% per year
D	Hard to share Hard to reuse	15 days	15 days	73 days	47days	10% per year

could occur, for instance, when there is an extra requirement on interoperability on the part of the reuser, or if the information is scattered in different files that have to be put together, or if the data is very messy and needs cleaning. This costs the reuser much time (fifteen days).

In scenario C the time for preparing the dataset for reuse by the sharing researcher is long (fifteen days). This could occur for instance when the dataset is complex and there are extra requirements on interoperability, providing the necessary context. Or, if a there is a backlog on annotation needed to properly interpret the data that has to be caught up on by lack of good data management. It is assumed that the effort to prepare for sharing was useful and the reuser requires a short time (one day) to get to know the set and further prepare it for reuse.

Scenario D is a case where the dataset requires a lot (fifteen days) of work to prepare. The reuser has to do preparations as well (fifteen days), for instance because the reuser has different requirements than the sharer.

We calculate the scenarios for three different sharing rates. These are 25%, 50%, and 75% researchers sharing. The following is fixed in the scenarios: Time spent to write a paper (47 days) and produce a dataset (73 days). These values are based on the empirical evidence that researchers on average author three papers per year (source: Scopus, analyzed in Pronk et al., 2015), so the time needed for these two tasks is on average one third of a year. This means the time won at reuse is 73 days (minus time spent for the reuse), as the researcher doesn't have to produce a new dataset. The decay rate of datasets in the model is set at 10%, which means datasets remain in the dataset pool for an average of 10 years (Fry et al., 2008). This is based on the notion that expectedly datasets go out of date or fashion because of continuously improved measurement techniques, bad archiving practices (i.e. outdated unusable file formats, data loss), lack of interest to store the data longer than a minimum of ten years, or policies that limit the archiving period to ten years.

For all scenarios and the three different sharing rates the break-even point is calculated. This is the point where reuse is enough to compensate for the time spent in preparing datasets for sharing. Two different metrics are calculated to represent minimum reuse. Firstly, reuse of datasets relative to the total amount of datasets available. Secondly, papers that reuse a dataset, relative to the total papers published in the scenario. Both are model simulation outcomes.

As an additional investigation we calculate the efficiency gain for the different scenarios in the case where the reuse rate is fixed to 5% of yearly published papers. The analysis should be seen as a proof-of-principle, in the absence of reliable numbers on actual current reuse rates. The analysis will give an indication if this 5% of reuse papers will in theory be enough to make up for sharing effort in all scenarios and at all sharing rates.

Results

Figure 1 shows results from modeling scenarios A, B, C, D (see **Table 1**) at 50% sharing researchers. This Figure shows the percentages of reuse (y-axis) and the average percent net time efficiency of the researchers in the community (x-axis). The break-even point in terms of time efficiency is in each subplot indicated by a black vertical line that intersects the x-axis at zero net added productivity. Large dots represent the reuse at which a break-even point is reached. The dotted horizontal lines intersect the y-axis at the needed percentage reuse for the break-even point.

Clearly, scenarios A and B where the sharing researchers spend one day to share data need much less reuse (in terms of either datasets or publications) to break even compared to scenarios C and D, where sharing researchers spend fifteen days to prepare data for sharing. The time that reusers spend to prepare the dataset for use (either 1 day in scenario A and C or fifteen days in scenario B and D) has much less impact on the

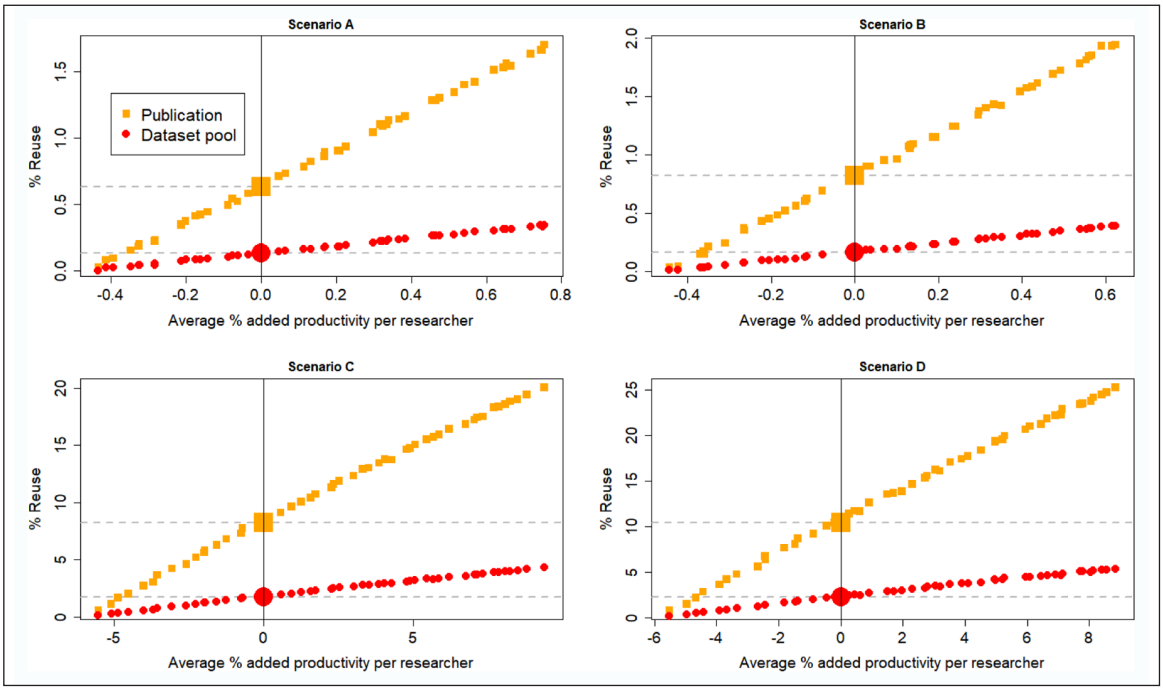

Figure 1: The influence of reuse, in percentages of either publications or available datasets, on average net efficiency (productivity) for the different scenarios at 50% sharing researchers. The large points represent the reuse at which a break-even point is reached. See Table 1 for an explanation of the scenarios.

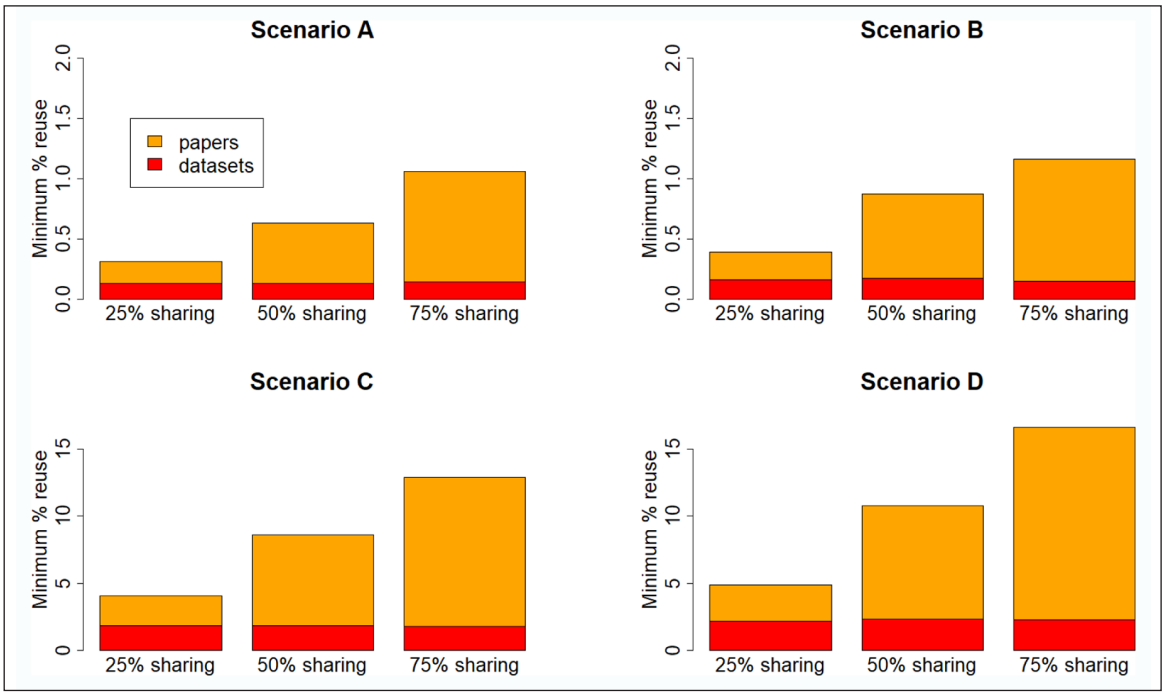

Figure 2: The minimum reuse, in terms of either publications or available datasets, needed for a break-even point in time invested in sharing and time gain by reuse, for different scenarios and for different percentages of researchers in a scientific community that are sharing. See Table 1 for an explanation of the scenarios.

reuse needed to reach the break-even point. Still, a small increase is observed in the scenarios where also reusers need to spend more time.

These scenarios were all calculated at a fixed 50% sharing researchers. Logically, community costs will increase if more researchers share. **Figure 2** shows how the break-even point for the scenarios (see **Table 1**) is altered at different numbers of sharing and non-sharing researchers.

In all scenarios in **Figure 2** it can be seen that an increased percentage of sharing researchers means more reuse is needed to reach the break-even point for efficiency.

Table 2: Net time efficiency of the scientific community in the case that 5% of papers are based on a reuse dataset, for different scenarios and different sharing rates. See Table 1 for an explanation of the scenarios.

	25% sharing	50% sharing	75% sharing
Scenario A	3.5%	3.2%	2.9%
Scenario B	2.6%	2.3%	2.1%
Scenario C	0.6%	−2.6%	−5.8%
Scenario D	0%	−3.2%	−6.6%

As an additional last analysis, the net efficiency at a fixed amount of reuse was investigated. The assumed reuse rate is 5%. The results are in **Table 2**. From **Table 2** it is clear that with this reuse rate of 5%, scenario A and B have a positive net efficiency, at all sharing rates. scenario C, on the other hand, has positive net efficiencies only at lower sharing rates. For scenario D there is no positive net effect at any of the analyzed sharing rates.

In short, given model simulations, sharing data can lead to a net benefit in terms of time efficiency (or productivity) for scientific communities where data is also reused. If either time for making data sharable and/or time to get to know and prepare the dataset for reuse is long, more reuse is needed to compensate for costs of sharing research data (**Figures 1** and **2**). If reuse is too little or doesn't occur, sharing will not be beneficial for scientific output (**Figure 1, Table 2**). If the sharing rate is high, more reuse is needed to break even. Lastly, scenario B in which the effort lies with the reuser is *more* efficient in terms of scientific output of the community than scenario C where the effort lies with the sharers themselves.

Discussion

Without any doubt, more productive, time efficient science is not the only benefit of a data reuse event. Data can be used by others than the scientific community, such as policy makers, teachers, or other members of the general public. Bishop and Kuula-Luumi (2017) show that in the case of a repository for qualitative data, most (85%) downloads are for teaching or learning purposes. Moreover, within the scientific community, available research data may have benefits other than actual reuse of the data itself for new products. Research data may allow researchers to optimally design their own study (i.e. Kreuzthaler et al., 2015). This is a valuable use of existing research data, but unfortunately not very traceable as it is not generally rewarded with a reference to the original study. The study in this paper is limited to sharing to enable reuse in a scientific context, so researchers can perform new research with existing data. This saves time by not having to produce research data, and as such increases scientific productivity.

The results of model simulations in this paper exemplify and make explicit several things. Firstly, model simulations show that if datasets take a disproportionate time to make ready for analysis (messy, incomplete, not interoperable data, i.e. not well prepared for reuse), then making these sets available will mean that a lot of reuse needs to occur before they contribute to scientific efficiency and output. Moreover, initially the more researchers share, the more reuse is needed to compensate for the time investment. That is why it is important to select data that is well worth the effort, is likely going to be reused, and to be critical of the ease of reusability. Statistics indicate that the reuse of datasets is indeed heavily skewed (i.e. Bishop and Kuula-Luumi, Piwowar and Vision 2013, Fear 2013) with some datasets being heavily reused, and some never. Fear (2013) investigates some characteristics of datasets that could be indicative of future reuse. Larger and more comprehensive datasets are, for instance, likely to be at an advantage. However, this does not imply that other research datasets do not need to be available. For transparent, verifiable science, it is always necessary that the dataset is preserved for the longer term. However, sharing with a view to verifiable research requires a different investment in preparing a dataset, than sharing for reuse does (see introduction). Sharing for verifiability means adhering to previously obtained results, to make sure they are repeatable and understandable.

Secondly, results stress the importance of stimulating researchers to make use of available and suitable datasets. If reuse is not enough, sharing will not lead to an efficiency revenue. There are strategies to increase reuse. Some funders already ask researchers to explain why existing data wasn't sufficient, to stimulate awareness of existing data. It would also be a good thing if more documentation on efficient search strategies for research data is available to researchers (e.g. Gregory, 2018). There have been several papers on what makes a good reuse experience (Faniel et al., 2013, Faniel et al., 2015, Zimmerman, 2008; Yoon, 2016). This information can be exploited to make reuse easier. Moreover, researchers tend to

trust data with reported earlier use (Curty et al., 2017), and this knowledge can be exploited to increase reuse.

Thirdly, in the simulations it proved more efficient overall to leave the effort to prepare the dataset for reuse to the reusers themselves, under the assumption that they are able to do so. This is because in the scenarios only a small part of datasets is reused, this will accumulate to a small investment, compared to the investment when all sharers having to spend time. Despite of this it seems logical that the person who created the dataset also is the best person to prepare it for reuse. Similarly in scientific publications, all efforts to make results readily interpretable are done by the publishing scientist. This helps other scientists towards new research. This is efficient as papers are only submitted if they are deemed to be of sufficient novelty and quality to pass peer review and in addition researchers reuse for their papers on average 1–6 scientific papers per page (Milojević, 2012) so there is a high reuse rate. Depending on the discipline, only 12% to 32% of publications remain uncited (e.g. officially unused) (van Noorden, 2017) and this number is decreasing in more recent literature (Lariviere et al., 2009). Measures to improve appropriateness and reuse in scientific papers might in this sense be equally applied to research data. In the current study it will become more efficient if sharers themselves prepare datasets for reuse when there are less shared datasets and more reusing researchers (results not shown).

Model results are always a simplification of the real-world situation. Pronk et al. 2015 already discuss shortcomings in the current model. For one, the model uses averages for all datasets. The outcomes are valid for disciplines or subfields where on average these numbers apply. Also, one particular characteristic of the model is that it has outcome variables '% datasets reused' and '% papers reusing a dataset' (**Figure 1**). This relation will differ in reality in cases where several datasets are reused in one paper. The number of reused datasets associated with the break-even point will be higher in cases where multiple datasets are reused for a single paper. For gene expression papers, typically more datasets in a single investigation were reused (Piwowar and Vision, 2013). For practical reasons the range of parameter values that were investigated was limited. Results may have seemed more positive if the time to prepare a dataset for reuse took up less than 15 days, and 15 days may be more appropriate to some disciplines than to others. To alleviate this and enable anyone to check model results to find the break-even point for different (average) parameter settings, an application is provided to enable anyone to change parameters via https://tessapronk.shinyapps.io/ReuseResearchDataMinimum/.

On another note, the question will remain how these calculated numbers relate to actual reuse. Is actual reuse enough to make sharing beneficial, given the calculated needed reuse? Despite of numerous very impressive and promising use-cases, quantifying research data reuse for scientific studies proves difficult (Piwowar et al., 2011). Whereas counting citations is a traditional way to measure the use of publications, citing research data is not yet common practice (Belter, 2014; Piwowar, 2011). There are no reliable methods as of yet to measure reuse. In general, the cases where citations to the dataset were taken as a measure for reuse, underestimation occurred. Piwowar et al. (2011) show for the three databases Pangaea, GEO, Treebase that the citing of a dataset is (more) frequently done by citing the data producing paper, instead of the dataset itself. Sometimes the dataset is mentioned exclusively in the methods and materials and is not officially cited (Piwowar et al., 2011). Indirect measures are available for measuring the use of datasets from repositories in downloads or views, which can give a clue as to the popularity of datasets. Bishop and Kuula-Luumi (2017) reported 15% of all downloads from the repository with quantitative data were intended for research purposes. Although these numbers are insightful, it is still unknown how many of these downloads resulted in actual reuse of the data for research. In disciplines where data is interoperable by the standards of measurement such as in gene expression profiling studies, reuse is known to be common. The reuse of microarray data was thoroughly quantified by Piwowar and Vision (2013). About 20% of the datasets deposited between 2003 and 2007 had been reused at least once by other scientists who published on it, which according to the authors is a very conservative estimate. Until the reuse of datasets can be reliably measured, it will be unknown whether reuse rates in different disciplines are enough to provide a time efficiency gain.

In conclusion, this study investigated and estimated the theoretical break-even point in efficiency gain for the scientific community in preparing research data for sharing (implicating a time cost) and its consecutive reuse (implicating a time gain). The results exemplify that sharing research data may indeed cause an efficiency revenue for the scientific community. The scientific community with the lowest reuse needed to reach a break-even point is one that has few sharing researchers and low time investments for sharing and reuse. Results stress the importance of stimulating and increasing reuse, and more so in the cases where the time needed by researchers to make datasets suitable for reuse is longer, or when the number

of sharing researchers is high. Moreover, results suggest overall efficiency can be increased if not too much effort is put into sharing datasets with low probability of reuse. Sharing data for reuse becomes worthwhile if the data is indeed sufficiently reused, contributing to more time efficient, productive scientific communities.

Acknowledgements

Conny van Bezu from Utrecht University Library has improved the use of English language in this paper. Jeroen Bosman, Marcel van Assen, and Igbal Safarov have taken the time to reflect on and discuss either the initial idea or (parts of) the findings in this paper.

Competing Interests

The author has no competing interests to declare.

References

Beagrie, N and **Houghton, J.** 2014. The Value and Impact of Data Sharing and Curation. A synthesis of three recent studies of UK research data centres. *Jisc report*. http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf.

Belter, CW. 2014. Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One*, 9(3): e92590. DOI: https://doi.org/10.1371/journal.pone.0092590

Bishop, L and **Kuula-Luumi, A.** 2017. Revisiting Qualitative Data Reuse: A Decade On. *SAGE Open*, 1–15. January–March 2017. DOI: https://doi.org/10.1177/2158244016685136

Bruland, P, McGilchrist, M, Zapletal, E, Acosta, D, Proeve, J, Askin, S, Ganslandt, T, Doods, J and **Dugas, M.** 2016. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. *BMC Medical Research Methodology*, 16: 159. DOI: https://doi.org/10.1186/s12874-016-0259-3

Curty, RG, Crowston, K, Specht, A, Grant, BW and **Dalton, ED.** 2017. Attitudes and norms affecting scientists' data reuse. *PLoS ONE*, 12(12): e0189288. DOI: https://doi.org/10.1371/journal.pone.0189288

Enke, N, Thessen, A, Bach, K, Bendix, J, Seeger, B and **Gemeinholzer, B.** 2012. The user's view on biodiversity data sharing. Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, 11: 25–33. DOI: https://doi.org/10.1016/j.ecoinf.2012.03.004

Faniel, I, Kansa, E, Whitcher Kansa, S, Barrera-Gomez, J and **Yakel, E.** 2013. The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. *JCDL 2013 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 295–304. New York, NY: ACM. DOI: https://doi.org/10.1145/2467696.2467712

Faniel, IM, Kriesberg, A and **Yakel, E.** 2015. Social Scientists' Satisfaction With Data Reuse. *Journal of the Association for Information Science and Technology*, June.

Fear, KM. 2013. Measuring and anticipating the impact of data reuse. Dissertation: University of Michigan. http://hdl.handle.net/2027.42/102481.

Figueiredo, AS. 2017. Data Sharing: Convert Challenges into Opportunities. *Front. Public Health*, 5: 327. DOI: https://doi.org/10.3389/fpubh.2017.00327

Fry, J, Lockyer, S, Oppenheim, C, Houghton, J and **Rasmussen, B.** 2008. Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes. https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/4600/1/JISC%20data%20sharing_final%20report.pdf.

Gregory, K, Khalsa, SJ, Michener, WK, Psomopoulos, FE, de Waard, A and **Wu, M.** 2018. Eleven quick tips for finding research data. *PLoS Comput Biol*, 14(4): e1006038. DOI: https://doi.org/10.1371/journal.pcbi.1006038

Kim, Y and **Yoon, A.** 2017. Scientists' Data Reuse Behaviors: A Multilevel Analysis. *Journal of the Association for Information Science and Technology*, 68(12): 2709–2719. DOI: https://doi.org/10.1002/asi.23892

Kolb, TC, Blukacz-Richards, EA, Muir, AM, Claramunt, RM, Koops, MA, Taylor, WW, et al. 2013. How to manage data to enhance their potential for synthesis, presernation, sharing and reuse: A Great Lakes case study. *Fisheries*, 38(2): 52–64. DOI: https://doi.org/10.1080/03632415.2013.757975

Koslow, SH. 2002. Sharing primary data: A threat or asset to discovery? *Nature*, 3: 311–313. April 2002. DOI: https://doi.org/10.1038/nrn787

Kreuzthaler, M, Schulz, S and **Berghold, A.** 2015. Secondary use of electronic health records for building cohort studies through top-down informatiaion extraction. *Journal of Biomedical Informatics*, 53: 188–195. DOI: https://doi.org/10.1016/j.jbi.2014.10.010

Larivière, V, Gingras, Y and **Archambault, E.** 2009. The decline in the concentration of citations, 1900–2007. *Journal of the Association for Information Science and Technology*, 60(4): 858–862. DOI: https://doi.org/10.1002/asi.21011

Milojević, S. 2012. How Are Academic Age, Productivity and Collaboration Related to Citing Behavior of Researchers? *PLoS ONE*, 7(11): e49176. DOI: https://doi.org/10.1371/journal.pone.0049176

Piwowar, HA, Carlson, JD and **Vision, TJ.** 2011. Beginning to track 1000 datasets from public repositories into the published literature. *ASIST 2011*, 9–13. October. New Orleans, LA, USA. DOI: https://doi.org/10.1002/meet.2011.14504801337

Piwowar, HA and **Vision, TJ.** 2013. Data reuse and the open data citation advantage, *PeerJ*, 1: e175. DOI: https://doi.org/10.7717/peerj.175

Pronk, TE. 2018. Replication data for: The time efficiency gain in sharing and reuse of research data. *DataverseNL*. Available at: https://hdl.handle.net/10411/4MB189 V2 [Version].

Pronk, TE, Wiersma, PH and **van Weerden, A.** 2015a. Replication data for: Games with research data sharing. *DataverseNL*. Available at: http://hdl.handle.net/10411/20328 V4 [Version].

Pronk, TE, Wiersma, PH, van Weerden, A and **Schieving, F.** 2015. A game theoretic analysis of research data sharing. *PeerJ*, 3: e1242. DOI: https://doi.org/10.7717/peerj.1242

Sprague, LA, Oelsner, GP and **Argue, DM.** 2017. Challenges with secondary use of multi-source water-quality data in the United States. *Water research*, 110: 252–261. DOI: https://doi.org/10.1016/j.watres.2016.12.024

van Noorden, R. 2017. The science that's never been cited. *Nature*, 552: 162–164. DOI: https://doi.org/10.1038/d41586-017-08404-0

Yoon, A. 2016. Data Reusers' Trust Development. *Journal of the Association for Information Science and Technology*, 68(4): 946–956. DOI: https://doi.org/10.1002/asi.23730

Zimmerman, AS. 2007. Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *Int J Digit Libr*, 7: 5–16. DOI: https://doi.org/10.1007/s00799-007-0015-8

Zimmerman, AS. 2008. New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology, & Human Values*, 33(5): 631–652. DOI: https://doi.org/10.1177/0162243907306704

How to cite this article: Pronk, TE. 2019. The Time Efficiency Gain in Sharing and Reuse of Research Data. *Data Science Journal,* 18: 10, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2019-010

Submitted: 29 July 2018 **Accepted:** 25 February 2019 **Published:** 19 March 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS