Projektbericht
Research Report

September 2024

Perceptions of data waste among researchers
First survey evidence

Robert Braun,
Elisabeth Frankus,
Katharina Gangl,
Sabine Neuhofer

Institut für Höhere Studien
Institute for Advanced Studies
Vienna

1. Introduction

"Dark Data" refers to unknown, unused, unnecessary, or redundantly stored data that consumes resources like storage, energy, and computing power, often posing a security risk without offering corresponding benefits. In research, dark data also opposes the goal of transparent and comprehensive knowledge production. Estimates suggest that up to 95% of all stored data might be Dark Data (Hewitt, 2022). There is an emerging field of research called Critical Data Studies (Dalton et al., 2016; Dalton & Thatcher, 2015; Kitchin & Lauriault, 2018; Lucivero, 2020), however their approach, findings and critical methods have not (yet) permeated discourses outside of their own academic field. Although some papers problematize dark data and issues related to sustainability, privacy and other risks (Corbett, 2018; Grimm, 2019; Schembera & Durán, 2020; Seehusen & Maldonado, 2020; Shetty, 2017) there has been a limited engagement with dark data risks and potential governance or management practices to address such challenges. Technical solutions like AI might worsen the problem by making it easier to create ever new file versions (Pohl et al., 2019). Thus, actively managing Dark Data, such as deleting redundant data is necessary to reduce the amount of dark data. This research report provides first empirical evidence on researchers' awareness about data waste as well as willingness to participate in such data waste management courses. After conducting a literature review and framing the problem, we defined the research gap as follows: data waste or dark data is under-researched, as well as vaguely defined, and mostly dealt with in grey and business-oriented literature. We decided that in order to define a research question a small exploratory project is advised, mapping the awareness and practices of academic researchers working with various forms of data. Therefore, we opted to run an online survey focusing on researchers and their data management practices.

Data collection: An online survey was disseminated between November 2023 and Mai 2024 via snowball method among researchers across Europe (e.g., via mailing lists[1], flyers at scientific conferences in Vienna[2]). The questionnaire was developed based on a literature review and five qualitative interviews with administrative and scientific personnel of an Austrian research institution.[3]

Sample: In total 214 participants started the questionnaire, 124 completed it. The sample consists of 51% women and 46% men; half of the sample is 40 years or younger;

[1] The survey was spread in personal and professional networks. This includes researchers from BOKU Vienna, TU Vienna, University of Vienna, VetMed, TU Graz, FH OÖ, It has also been shared in international professional networks' mailing lists in STS and Critical Data Studies as well as sent out to two currently running research projects' consortia (funded by Horizon Europe).

[2] Conferences visited were took place in Vienna, the fields include different disciplines of the Humanities and Biology.

[3] The five respondents are employed at the Institute of Advanced Studies, Vienna. The respondents were sampled from different organizational units: two people from IT, one person from the library services, and researchers.

64% are researchers in the social sciences, 36% in other sciences (natural sciences, humanities, formal sciences, applied sciences).

2. Spotlight of results

Rather small response rate. Considering the effort and time invested in distributing the questionnaire, the response rate is comparatively low. This already shows a relatively

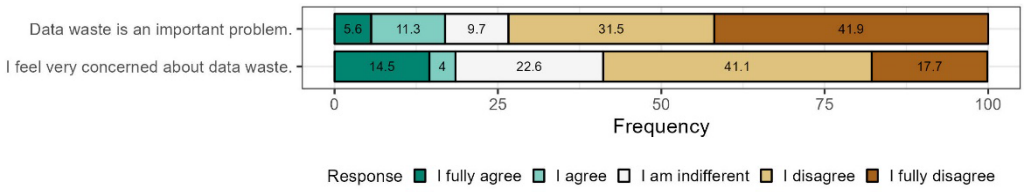

Figure 1 Plot of relative frequencies. N = 124. Question "Please indicate how much you agree or disagree with these statements:" Response options: fully agree, agree, indifferent, disagree, fully disagree (5)"

low level of awareness about the topic.

Researchers do not consider data waste as a big problem: Approximately 17% of respondents (fully) agree that "data waste is an important problem" and 19% "feel very concerned about data waste" (see Figure 1).

However, researchers' knowledge about data and data waste is limited: Only 22% estimated the (correct) share of data waste to be between 61-90% (cf. https://www.logikcull.com/blog/what-is-dark-data), and 75% think less than 60% is waste.

Also, researchers have no coherent understanding of "data"[1]. In total, 214[2] respondents produced 129 different associations with the word "data". The most frequent association is "information", provided by 15% of respondents (see Figure 2 for count of the 15 most frequent associations). Less frequent associations include more qualitative

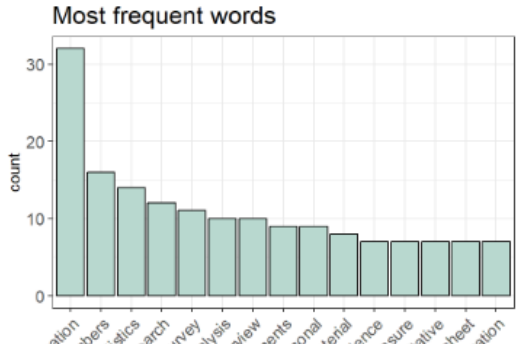

Figure 2 The 15 most frequent associations with "data" (N = 214)

[1] We asked in the questionnaire "*What do you think of when you think of "data"? Please just tell us your first three spontaneous associations*". Two research assistants sorted responses into categories.

[2] The questionnaire offered three fields to input associations. Not every respondent offered three associations.

types of data, such as “interviews”, “text” and, furthermore, the process, including “storage” or “collection”, among other domains.

Nonetheless, 60% of the surveyed researchers are willing to attend a data waste management course, 40% have no interest to attend a course.[1]

3. Implications of results

Our survey was likely filled in by researchers interested in the topic, thus, it can be assumed that the general population of researchers has an even more divers understanding of data, less concerns of data waste and thus, might also be less willing to participate in a data waste management course. Nonetheless, these first results highlight the potential success concerning the uptake of offering data management courses to researchers: Even though few researchers consider data waste as a problem, many would be willing to participate in a data waste management course. Such courses would ensure professional data management at research institutions and can reduce data waste. From a research perspective our exploratory research confirmed the need for a more thorough research agenda related to data waste and dark data management practices. It is also important to bring learning from Critical Data Studies literature to a more general research audience; to engage researchers and research funders in dialogue about data, data management and data waste/dark data challenges.

The present results should encourage research managers in offering relevant qualifications among research to handle data waste. In addition, the results can be used for future research projects on data handling in science and beyond such as in private and public organizations.

References

Corbett, C. J. (2018). How sustainable is big data? Production and Operations Management, 27(9), 1685-1695. https://doi.org/https://doi.org/10.1111/poms.12837

Dalton, C., Taylor, L., & Thatcher, J. (2016). Critical data studies: A dialog on data and space. Big Data & Society. https://doi.org/10.1177/2053951716648346

Dalton, C., & Thatcher, J. (2015). Inflated granularity: Spatial “big data” and geodemographics. Big Data & Society, 2, 1–15.

Grimm, D. (2019). The dark data quandary. American University Law Review, 68(3), 761-821.

Hewitt, N. (2022). What is dark data and how can we find it? [Blog Post] https://www.imperva.com/blog/what-is-dark-data-and-how-can-we-find-it/

Kitchin, R., & Lauriault, T. (2018). Towards critical cata ctudies: Charting and unpacking data assemblages and their work. In J. Eckert, A. Shears, & J. Thatcher (Eds.), Geoweb and Big Data.

Lucivero, F. (2020). Big data, big waste? A reflection on the environmental sustainability of big data initiatives. Science and Engineering Ethics, 26, 1009–1030. https://doi.org/https://doi.org/10.1007/s11948-019-00171-7

Pohl, J., Hilty, L. M., & Finkbeiner, M. (2019). How LCA contributes to the environmental assessment of higher order effects of ICT application: A review of different approaches. Journal of Cleaner Production, 219, 698-712. https://doi.org/https://doi.org/10.1016/j.jclepro.2019.02.018

Schembera, B., & Durán, J. M. (2020). Dark data as the new challenge for big data science and the introduction of the scientific data officer. Philosophy of Technology, 33, 93–115.

Seehusen, V., & Maldonado, E. (2020). Using a roadmap in the back alleys of dark data. Journal of Technology Research.

Shetty, S. (2017). How to tackle dark data. Gartner. https://www.gartner.com/smarterwithgartner/how-to-tackle-dark-data