Research Data: Incentive Structures in the Science System and the Use of Generative AI

Angelina Sophie Dähms[1*], Linda Nierling[1], Ralf Schneider[1], and Constanze Scherz[1]

[1]Karlsruhe Institute of Technology, Institute of Technology Assesment and Systems Analysis, Karlsruhe, Germany
[*]Corresponding author: angelina.daehms@kit.edu.

Introduction

The benefits of publishing research data, such as improved research integrity through data exchange, reproducibility and quality assurance, the potential for scientific careers (visibility) and increased efficiency [1–4] are offset by the following barriers: disciplinary, institutional and legal hurdles as well as uncertainties (lack of data standards, problems due to data protection and legal uncertainty); recognition problems in data transfer; operational challenges (time expenditure, resources) [1, 5, 6]. Potential measures for the publication of research data include, for example, impact metrics for research data, the establishment of recognised quality standards and metadata structures, and the allocation of resources for research data management (RDM) [1, 3, 5]. But what are the benefits, barriers and potential measures from a disciplinary perspective among the use case partners involved in the DiTraRe (Digital Transformation of Research) project? To this end, interviews were conducted with use case partners and, on this basis, a workshop was held to analyse both overarching developments and specific needs in the field of research data, what incentives and barriers exist in the scientific system with regard to research data, and what potential and limitations generative AI has. The interviews and the workshop were conducted by the Institute for Technology Assessment and Systems Analysis (ITAS) as part of the DiTraRe project.

Objective

The aim of the interviews and the workshop is to conduct a comprehensive reflection on the potential and challenges of research data, with a particular focus on incentive structures in the scientific system and the potential and challenges of generative AI.

Methods

For the interviews on research data, open data and open science, DiTraRe's use case partners were consulted on the following topics: understanding (types of data and their publication); evaluation of open data; framework conditions for open data; digital transformation of science and current developments (science-science interface, science-society interface). Based on the interview results, an internal workshop was held on the dimension of 'reflection and resonance'. The topics were: research data relating to incentive structures in the science system and the use of generative AI. Participants included use case partners and partners from DiTraRe dimensions, as well as external partners such as researchers and practitioners involved with research data from KIT and FIZ Karlsruhe. In the workshop, participants were introduced to the current state of research on the benefits, barriers and potential measures for publishing research data, particularly with the use of generative AI. The results from the interviews were then incorporated and used as the basis for two rounds of discussion. The first discussion round focused on incentive structures in the scientific system. The second discussion round was organised around the following topics: increasing the visibility of data sets; improving metadata information; increasing potential misuse of data sharing.

Results

The first round of discussions on incentive structures within the scientific system focused on subject-specific incentives and obstacles to the publication and uptake of research

data. Potentials and challenges of generative AI serving to support more and better open data and open science. Participants emphasised the tension between the development of impact metrics for research data and the perception that RDM is time-consuming. In addition, the debate revealed an ambivalence between promoting exchange, fostering an open culture of error and advancing scientific progress on the one hand, and concerns about idea theft and loss of prestige through the disclosure of errors and redundancies on the other.

The second round of discussions focused on three topics, first improving the visibility of datasets, second improving metadata information, and third potential misuse of data sharing, which were discussed individually, dealing with the potential of generative AI in relation to research data. To improve the visibility of data sets, data sets should be expanded in accordance with a clear definition using AI to develop new meaningful key figures. In addition, usage requirements and scenarios for the use of data sets by AI should be developed. In order to improve metadata information, it should be enhanced using generative AI with human-in-the-loop, metadata information should be standardised and clear rules should be established. In addition, automation using generative AI is perceived as an opportunity (increased efficiency and time savings), but also as a risk (without comprehensibility). Protective measures are to be developed to prevent potential misuse of data sharing before and after the use of generative AI. In addition, transparent use of AI in combination with human-in-the-loop is to be implemented for data-efficient and evaluable generation of new data, large data sets and from previously unconnected repositories.

The literature review, interviews with use case partners, and discussions during the workshop revealed that there are many benefits to using (generative) AI, e.g. the exchange of research ideas and cooperation through open data. Obstacles include a lack of anchoring and reputation in the scientific system, the high time expenditure for RDM, concerns about loss of reputation, and theft of paper ideas.

Discussion

An interdisciplinary cultural shift in science towards open science and open data is still necessary to fully exploit the potential of research data. This includes the publication of data as well as the motivations, advantages, and implementation of data management. This requires constant exchange of suitable formats and standards, even across disciplinary boundaries. The debate on generative AI in its use for research data is only just beginning [6,7]. Identifying potential and obstacles was therefore a fruitful exercise and focused on ways in which generative AI could facilitate the exploitation of research data potential. In the future, the research data community should hold specific discussions, for example to draw up guidelines for the specialist community. A holistic view of research data is required, from collection and publication to uptake and impact [8]. The holistic view and the use of generative AI is a good starting point for further discussions on new standards in open data in the times of AI.

References

[1] B. Fecher. (2013) Data sharing: Warum es sinnvoll ist, weshalb es trotzdem keiner tut und wozu es führen könnte. [Online]. Available: https://www.hiig.de/data-sharing-warum-es-sinnvoll-ist-weshalb-es-trotzdem-keiner-tut-und-wozu-es-fuhren-konnte-2/

[2] Goodey, Gregory et al, "The State of Open Data 2022," *Digital Science*, 2022. doi: 10.6084/m9.figshare.21276984 .

[3] Hahnel, Mark et al., "The State of Open Data 2023," *Digital Science*, 2023. doi: 10.6084/m9.figshare.24428194 .

[4] European Commission. (2025) Facts and Figures for Open Data. [Online]. Available: https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/open-science-monitor/facts-and-figures-open-research-data_en

[5] Tedersoo, Leho et al., "Data sharing practices and data availability upon request differ across scientific disciplines," *Scientific Data 8: 192*, 2021. doi: 10.1038/s41597-021-00981-0 .

[6] Gomes, Dylan G. E. et al., "Why don't we share data and code? Perceived barriers and benefits to public archiving practices," *Proceedings of the Royal Society B 289/1987: 20221113*, 2022. doi: https://doi.org/10.1098/rspb.2022.1113 .

[7] Chafetz, Hannah and Saxena, Sampriti and Verhulst, Stefaan G., "A Fouth Wave of Open Data? Exploring the Spectrum of Scenarios for Open Data and generative AI," TheGovLab, Tech. Rep., 2024. doi: 10.48550/arXiv.2405.04333 .

[8] Open Data Watch. (2024) The Data Value Chain: Moving from Production to Impact. [Online]. Available: https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact/