

Governing Research Data in the Age of Generative AI

Interdisciplinary Perspectives

Angelina Sophie Dähms, Felix Bach, Christian Bonatto Minella,
Tobias Kerzenmacher, Linda Nierling, Lea Sophie Singson

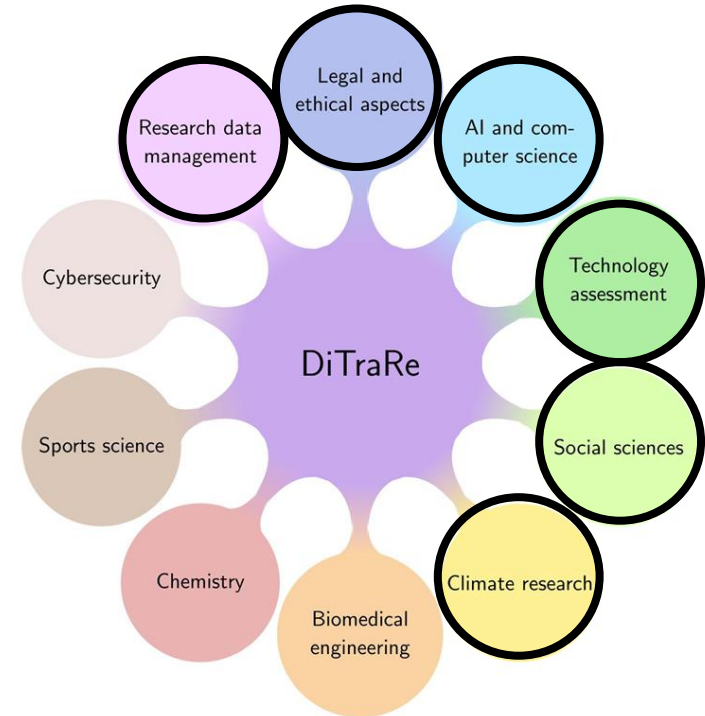
STS Conference Graz, 4 - 6 May 2026
ID 10: Governing Digital Research Infrastructures

Agenda

1. Project DiTraRe
2. Problem orientation & motivation
3. State of the art: research data
 1. Research data - potential benefits through GenAI
 2. Research data - potential barriers through GenAI
 3. Use of GenAI to overcome potential barriers
4. Use case: climate science & meteorology
5. Conclusion

1. Digital Transformation of Research (DiTraRe)

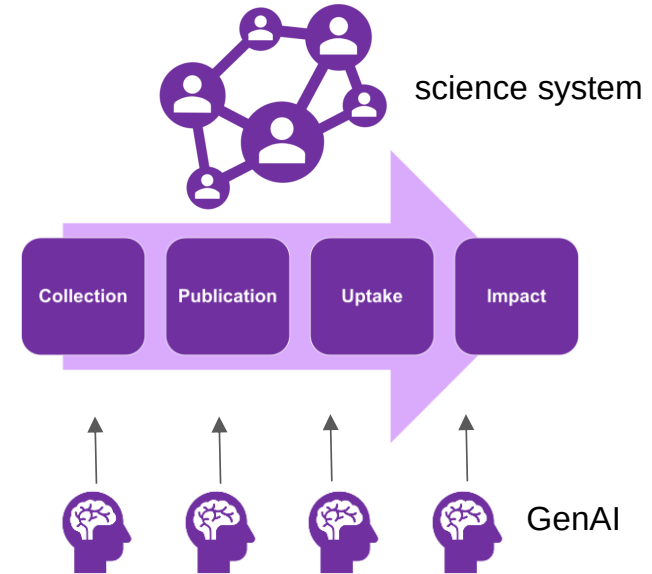
- Leibniz ScienceCampus “Digital Transformation of Research” is a joint project of FIZ Karlsruhe & KIT, funded by the Leibniz Association (2023-2027)
- Investigating the effects & potential of the increasing digitalization of scientific work in four interdisciplinary research clusters
- Includes the theoretical-conceptual level, methodological issues as well as consequences for science & society



Leibniz ScienceCampus
**Digital Transformation
of Research**

2. Problem Orientation & Motivation

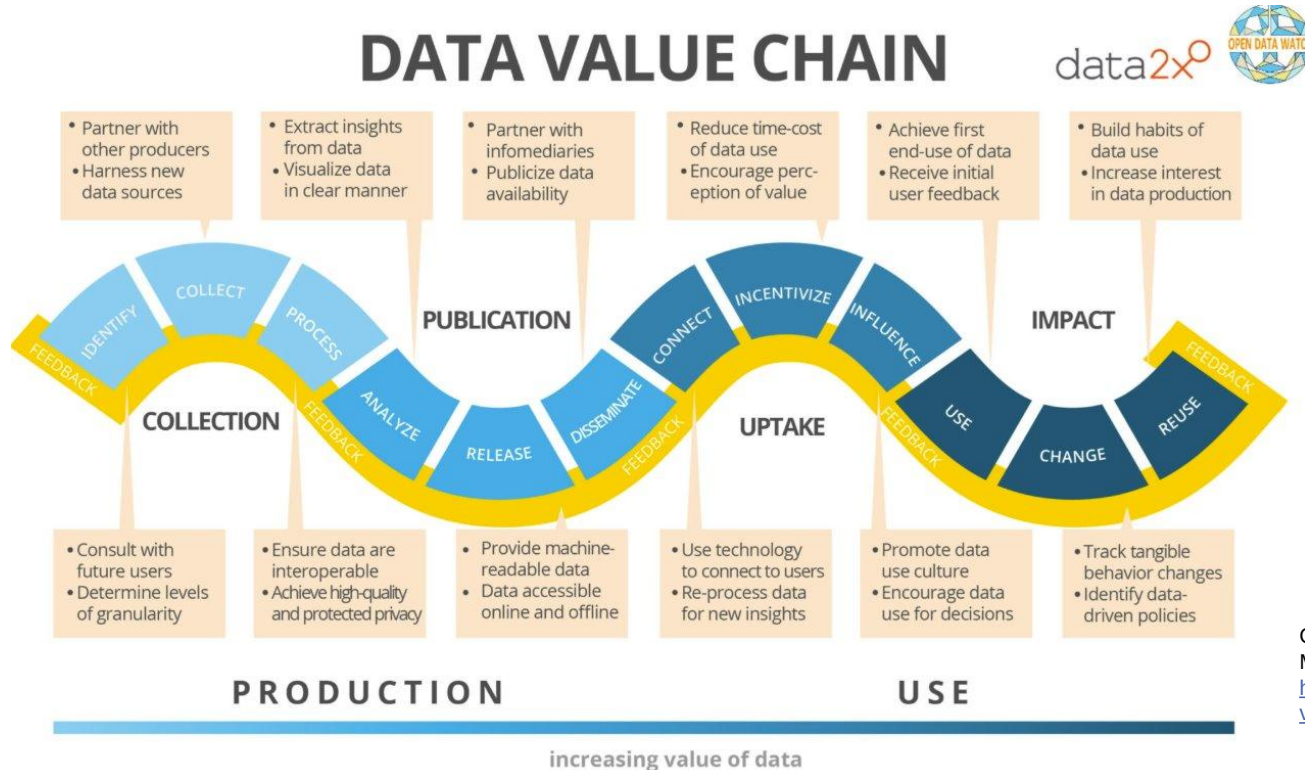
- Motivation: to provide a reflective perspective on research data in the era of digital transformation in research
- Aim: to move from discipline-specific starting points & challenges toward a general & pragmatic approach to the publication of research data with regard to GenAI
- Interdisciplinary approach & discipline-specific perspectives: TA, social science, computer science, law, climate science & meteorology



2. Problem Orientation & Motivation

- Challenges: data management arise in governing AI-supported research infrastructures
 - Ensuring data provenance
 - Traceability
 - Versioning of AI-processed datasets to maintain transparency & reproducibility
- GenAI with potential to change the relationship between data producers & data infrastructures
 - Automating metadata generation
 - FAIRness assessment
 - Persistent identifier assignment
- Legal regulation with a pacing problem
 - Fast technological development of GenAI
 - Anticipatory governance of the AI Act leaves regulatory gaps, particularly regarding rapidly evolving GenAI applications

2. Problem Orientation & Motivation



Open Data Watch. 2024. The Data Value Chain: Moving from Production to Impact.
<https://opendatawatch.com/publications/the-data-value-chain-moving-from-production-to-impact/>

3.1 Research Data - Potential Benefits through GenAI

- **Improved research processes & visibility** through data sharing & the credibility of research via AI-supported validation, error detection & transparent use of open data for training & analysis
- **Quality assurance** through best practices through GenAI by automating adherence to best practices (e.g., FAIR principles, standardization & metadata completeness)
- **Shift in scientific careers:** data publications & curated datasets might gain importance e.g. AI-enabled of open datasets & increased visibility of data contributions
- **Increased efficiency** through AI-driven data integration, synthesis & reuse, reducing redundant data collection & enabling faster knowledge generation

3.2 Research Data - Potential Barriers through GenAI

- **Increasing potential misuse of data sharing:**
 - Generation of new data using GenAI with potential for forgery (paper mills, data manipulation)
 - Political disinformation campaigns
 - Legal uncertainty for authorship
 - Biases in training data (e.g. only clear sky measurements are used for training)
 - Untransparent use of GenAI
 - Fabrication or distortion of structured metadata
- **Paradox developments in climate science:**
 - Ecological footprint of GenAI vs. climate research
 - Efficiency vs. CO₂ production
 - Data reuse benefits vs. compute cost of GenAI
 - High energy usage

3.3 Potential Use of GenAI to Overcome Barriers

- **Operational challenges:** time pressure, resources → reduction through GenAI by automating documentation, metadata creation, data curation & data management planning
- **Lack of discipline-specific, recognized data standards** → support via AI-assisted standardization, schema alignment & translation between heterogeneous data formats
- **Lack of recognition** → increasing through AI-enabled tracking of data reuse, automated citation practices & visibility of data contributions in research ecosystems
- **Idea theft & damage to reputation** → risk mitigation through AI-supported data licensing, provenance tracking & controlled access mechanism for data usage

4. Case: Climate Science & Meteorology



https://www.rdm.kit.edu/forschungsdaten_fdm_zyklus.php

Collection:

- **Process:** GenAI-support for cleaning data, recognition of anomalies, standardizing formats

Publication:

- **Analyze:** AI supported synthesis of heterogeneous climate data sets; high risk of hallucinations
- **Release:** obligatory labelling of AI generated datasets
- **Disseminate:** open, machine-readable, accessible and findable by users using adequate channels; GenAI can personalize dissemination, & augment or partially automate quality assurance

Uptake:

- **Connect:** link similar datasets; GenAI can recognize semantic similarities
- **Incentivize:** GenAI-supported citation tracking could incentivize data sharing

Impact:

- **Reuse:** data enrichment and curation are good reuse activities; but GenAI increases the risk of misuse

4. Case: Climate Science & Meteorology

- Increasing the visibility of datasets:
 - New key figures through GenAI
 - Usage requirements & scenarios
 - Linking similar datasets
 - Data set enrichment
 - Curating data sets into high-quality datasets
 - Mandatory labeling of AI-generated datasets
- Improving metadata information:
 - Use GenAI & human-in-the-loop approach to extract, process & standardize metadata from existing datasets & publications with clear & transparent rules
 - Improved documentation
 - Standardized information, voting in the communities = community validation?

4. STS & Use Case: Climate Science & Meteorology

- Automation redefines data work
 - Shifting responsibilities from human curators to algorithmic systems
 - Altering who is credited accountability
 - Altering how responsibility can be credited if at all (explainability)
 - Qualified to perform core data curation tasks
- GenAI reallocates responsibilities between human actors & technical systems
 - Balancing the need to work efficiently with the responsibility to manage research data ethically & with integrity
 - Modified quality assurance & validation procedures are necessary to verify AI-enhanced datasets & metadata
 - Potentially creating accountability gaps and diffusion of responsibility
 - Is human expertise for AI generated outputs still possible on all scales?

5. Conclusion

- Delegation of data curation tasks to AI systems introduces new governance & legal challenges:
 - Accountability, liability & compliance
 - Responsibility, when algorithmic outputs compromise reproducibility, degrade data quality, or shift interpretative authority away from human experts
- Potential measures & incentives:
 - Linking published data to publicly available research profiles such as ORCID or Google Scholar
 - Recognised quality standards & metadata information for GenAI
 - Guidelines for the use of GenAI in research data
 - Development of skills, training & support structures for researchers, data stewards are crucial components for responsibly integrating GenAI into RDM

5. Conclusion

- To assess the role of GenAI a holistic approach is needed: from data collection to societal impact across individual, organizational & legal levels
- Disciplinary differences and specific data settings have to be taken into account
- Debate on GenAI for research data is just beginning → joint reflection, shared community guidelines & promising use cases needed



Literature

- Ahern, Deirdre. 2025. The New Anticipatory Governance Culture for Innovation: Regulatory Foresight, Regulatory Experimentation and Regulatory Learning. In *European Business Organization Law Review* 26, 241-283. <https://doi.org/10.1007/s40804-025-00348-7>
- Bach, Felix, Linda Nierling, Christian Bonatto Minella, and Angelina Sophie Dähms. 2025. The Impact of Digitalisation on Science: Workflows, Outputs, and Trust. In *E-Science-Tage 2025: Research Data Management: Challenges in a Changing World*, Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger, Cora F. Krömer (ed.), 365-385. Heidelberg: heiBOOKS. <https://doi.org/10.11588/heibooks.1652.c23935>
- Chafetz, Hannah, Sampriti Saxena, and Stefaan G. Verhulst. 2024. *A Fourth Wave of Open Data? Exploring the Spectrum of Scenarios for Open Data and Generative AI*. <https://doi.org/10.48550/arXiv.2405.04333>
- Dähms, Angelina Sophie, Nierling, Linda, Scherz, Constanze, & Schneider, Ralf (2025). Research Data: Incentive Structures in the Science System and the Use of Generative AI. DiTraRe Symposium on Digitalisation of Research (DiTraRe Symposium), ZKM Karlsruhe. *Zenodo*. <https://doi.org/10.5281/zenodo.17898606>
- European Commission. 2025. Facts and Figures for open data. 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
- Fecher, Benedikt. 2013. Data sharing: Warum es sinnvoll ist, weshalb es trotzdem keiner tut und wozu es führen könnte. <https://www.hiig.de/data-sharing-warum-es-sinnvoll-ist-weshalb-es-trotzdem-keiner-tut-und-wozu-es-fuehren-konnte-2/>
- Gomes, Dylan G. E., Patrice Pottier, Robert Crystal-Ornelas, Emma J. Hudgins, Vivienne Foroughirad, Luna L. Sánchez-Reyes, Rachel Turba, Paula Andrea Martinez, David Moreau, Michael G. Bertram, Cooper A. Smout, and Kaitlyn M. Gaynor 2022. Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proceedings of the Royal Society B* 289: 20221113. <https://doi.org/10.1098/rspb.2022.1113>
- Goodey, Gregory, Mark Hahnel, Yuanchun Zhou, Lulu Jiang, Ishwar Chandramouliswaran, Amy Hafez, Taunton Paine, Susan Gregurick, Samuel Simango, Juan Miguel Palma Peña, Holly Murray, Matt Cannon, Rebecca Grant, Kate McKellar, and Laura Day. 2022. The State of Open Data 2022. *Digital Science*. <https://doi.org/10.6084/m9.figshare.21276984>

Literature

- Jahnel, Jutta, Michael Nentwich, Pauline Rioussel, Christine Milchram, Christoph Kehl, and Steffen Albrecht. 2025. Generative AI in policy-oriented technology assessment. ITA-manu:script No.: ITA-25-01, Vienna. <https://doi.org/10.1553/ita-ms-25-01>
- Hahnel, Mark, Graham Smith, Niki Scaplehorn, Henning Schoenenberger, and Laura Day. 2023. The State of Open Data 2023. Digital Science. <https://doi.org/10.6084/m9.figshare.24428194>
- Hofmann, Stefan, Kerstin Soltan, and Felix Bach. 2025. RADAR: Enhancing FAIR Research Data Management with AI Support. Open Science Conference 2025 (OSC2025), Hamburg. Zenodo. <https://zenodo.org/records/17222396>
- Jacyszyn, Anna M., Mahsa Vafaie, Felix Bach, Tobias Kerzenmacher, Kerstin Soltan, and Stefan Hofmann. 2025. "Streaming Data Publication: Automatic Metadata and Large Datasets in the Age of AI." Discussion Session at Open Science Conference 2025. Open Science Conference 2025 (OSC2025), Hamburg. Zenodo. <https://zenodo.org/records/17281939>
- KIT-Bibliothek. 2025. Guideline zum Publizieren von Forschungsdaten für Forschende am KIT. <https://forschungsdienste.bibliothek.kit.edu/guideline-publikation-forschungsdaten/>
- Novelli, Claudio, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. In *Computer Law & Security Review* 55(106066) <https://doi.org/10.1016/j.clsr.2024.106066>
- Tedersoo, Leho, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, Karin Kogermann, and Tuul Sepp. 2021. Data sharing practices and data availability upon request differ across scientific disciplines. In *Scientific Data* 8: 192. <https://doi.org/10.1038/s41597-021-00981-0>
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific Data* 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Yang, Wenli, Rui Fu, Muhammad Bilal Amin, and Byeong Kang. 2025. The Impact of Modern AI in Metadata Management. In *Human-Centric Intelligent Systems* 5, 323–350. <https://doi.org/10.1007/s44230-025-00106-5>

Thanks for your attention!

Further questions:
angelina.daehms@kit.edu



www.ditrare.de/en

