

“Hello ELSA, how are you?”

Legal and ethical challenges in RDM, current and future tasks of ELSA activities against the background of AI and Anonymisation

Franziska Boehm¹, Ulrich Sax² [<https://orcid.org/0000-0002-8188-3495>], Oliver Vettermann³ [<https://orcid.org/0000-0001-7393-1103>], Paweł Kamocki⁴ [<https://orcid.org/0000-0003-4881-7549>] and Vasilka Stoilova⁵ [<https://orcid.org/0000-0002-8961-509X>]

¹ FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe; Karlsruhe Institute for Technology; NFDI4Culture.

² Department of Medical Informatics, University Medical Center, Göttingen; NFDI4Health.

³ FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe; NFDI4Culture.

⁴ Leibniz Institute for the German Language, Mannheim; Text+.

⁵ University Library, Mannheim; BERD@NFDI.

Abstract. The proposed contribution will shed light on current and future challenges on legal and ethical questions in research data infrastructures. The authors of the proposal will present the work of NFDI’s section on Ethical, Legal and Social Aspects (hereinafter: ELSA), whose aim is to facilitate cross-disciplinary cooperation between the NFDI consortia in the relevant areas of management and re-use of research data.

Keywords: NFDI section; legal aspects; anonymisation; artificial intelligence; language models.

1. Introduction

Regarding the progress made in ELSA, we would like to present first results which are the outcomes of a workshop on anonymisation and of the ELSA Task Force aiming to put them forward for discussion. We discussed methods and tools around the anonymization of data spanning from tabular data [1] to free text data [2]. Another task force develops a decision tree on how to tackle privacy questions in research projects. Furthermore, we would like to engage in discussions on proposals for cross-consortium guidelines and legal standards. Specific topics we would like to discuss are data protection law with a focus on anonymisation as well as intellectual property law with a focus on legal status of AI-generated data. We also like to bring forward related ethical implications of using generative AI tools which were already discussed within ELSA.

2. Results

Moreover, new insights on questions arising out of the challenges of anonymisation and AI in a research data management context are demonstrated. Such challenges will affect all projects that (a) collect, process, or archive (personal) research data and/or (b) work with data or artefacts in which persons or organisations hold exploitation rights. By drawing on existing experi-

ence, we aim to enter into dialogue with rights holders in order to promote effective data protection and adhere to the research ethics guidelines (including the DFG principles of good scientific practice). Additional legal obstacles may arise from protecting research results that incur third-party rights. Solutions are to be prepared from various specialist communities. Lastly, ELSA also outlines training content for initial, further and continuing training in cooperation with section Training & Education (short: edutrain), as well as to involve external stakeholders in those areas where it seems appropriate. Addressing the privacy concerns and the accompanying legal uncertainties of NFDI consortia is the main goal of the newly created ELSA Task Force Data Protection. Task Force members focus on drafting standardized guidelines, which will enable researchers to adopt a FAIR approach to RDM while also protecting the personal data of their research subjects and adhering to the data protection regulations.

The rise of generative AI tools, such as Midjourney or ChatGPT, and their rapid development also created unexpected challenges for research data management, which were addressed by ELSA members. Especially in the field of text data, the so-called Large Language Models (or LLMs) like GPT-4 were a real gamechanger. Generative AI outputs are of increasing value for research, and are used as research data. Their legal status, in particular with regard to intellectual property rights, is difficult to determine, as illustrated by the recent Statement of Policy of the US Copyright Office (USCO). According to USCO, the mere fact of prompting an AI tool is not enough to claim copyright in the output. Still, works containing AI-generated material can be eligible for copyright protection, e.g. if such material is modified or arranged by a human author. In practice, the distinction is extremely difficult to make, especially that currently there is no reliable method to detect texts generated by ChatGPT.

Another pressing issue related to research data management and AI is to what extent data collected within the statutory exception for text and data mining for scientific research purposes (§ 60d UrhG) can be used to train AI models; under some approaches, the model itself, and at least some of its outputs, can be regarded as derived from the training data, which would fundamentally change their legal status. The discussion, closely monitored by ELSA, continues, and even more doubts were raised by the recent Getty Images lawsuit in the UK.

Finally, the use of generative AI raises important ethical questions, which also affect research data management. ELSA members proposed a taxonomy of ethical principles to be addressed at various stages of the research data lifecycle: Privacy, Property, Equality, Transparency and Freedom [3].

3. Discussion and Outlook

The section is open to the inclusion of further topics and willing to get an overview of newer questions such as the role of AI in RDM. A dialogue forum for the NFDI consortia is to be developed, which will serve to exchange legal, socio-scientific and research-ethical experiences and develop use-oriented, practical approaches to solutions. This is also explicitly aimed at the new consortia of the next NFDI approval rounds, for which the section is to offer a platform to constructively discuss overarching ethical and legal issues.

Competing interests

The authors declare that they have no competing interests.

Funding

The NFDI consortia where the authors are affiliated to are funded by the DFG.

References

1. Anna C. Haber, Ulrich Sax, Fabian Prasser, "Open tools for quantitative anonymization of tabular phenotype data: literature review", *Briefings in Bioinformatics*, Vol. 23, Issue 6, November 2022, bbac440, DOI: <https://doi.org/10.1093/bib/bbac440>.
2. Strathern et al., "QualiAnon - a Tool for anonymizing text data", Link: https://media-tum.ub.tum.de/doc/1575928/0hbkir8u5zldk4g2aw8icmnzo.TechReport_ws1.pdf.
3. Paweł Kamocki and Andreas Witt, "Ethical Issues in Language Resources and Language Technology – Tentative Categorisation", in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France. European Language Resources Association: 2022, pages 559–563.