1st Conference on Research Data Infrastructure

Humanities and Social Sciences

https://doi.org/10.52825/CoRDI.v1i.301

Published: 07 Sept. 2023

Open Science and Language Data: Expectations vs. Reality

The Role of Research Data Infrastructures

Paweł Kamocki[1][https://orcid.org/0000-0003-4881-7549], Erhard Hinrichs[2][https://orcid.org/0009-0006-4192-7779], Sabine Springer[3][https://orcid.org/0009-0004-8395-4279], Peter Leinen[4][https://orcid.org/0000-0002-3014-000X], Andreas Witt[5][https://orcid.org/0000-0002-0299-5713], and Dorothea Zechmann[6][https://orcid.org/0009-0008-2704-2135]

[1; 2; 5] Leibniz-Institut für Deutsche Sprache, Mannheim, Germany

[3; 4; 6] Deutsche Nationalbibliothek, Germany

Abstract. Language data are essential for any scientific endeavor. However, unlike numerical data, language data are often protected by copyright, as they easily meet the threshold of originality. The role of research infrastructures (such CLARIN, DARIAH, and Text+) is to bridge the gap between uses allowed by statutory exceptions and the requirements of Open Science. This is achieved on the one hand by sharing language data produced by research organisations with the widest possible circle of persons, and on the other by mutualizing efforts towards copyright clearance and appropriate licensing of datasets.

Keywords: Research infrastructures, Text data, Open science

Language data are essential for any scientific endeavor. Since natural language is a primary tool for human communication, the importance of language data reaches beyond language science and the humanities, also to the STEM. The importance of language data has been highlighted by the recent boom in generative chatbots (such as Chat GPT), based on Large Language Models (LLMs). However, unlike numerical data, language data are often protected by copyright, as they easily meet the threshold of originality (in 2009, the Court of Justice of the European Union ruled that texts as short as 11 words can be protected by copyright [1]). This means that, in principle, language data cannot be reproduced (copied) and communicated to the public (shared) without permission from the rightholders, unless in cases expressly authorised by statutory provisions (known as exceptions or limitations). The development of LLMs by large and mostly US-based companies was allowed under legal frameworks that do not apply in Europe (such as the US fair use doctrine).

At the beginning of copyright history, and for a relatively long time, research activities were regarded as irrelevant from the point of view of copyright, or exempted under the 'de minimis' principle. It is only in the second half of the 20th century, with the development of reproduction technologies (such as Xerox machines) that the conflict between the interests of copyright holders and the needs of academia became apparent, and research exceptions were introduced in national, and in 2001 also in European copyright law. For years, research exceptions could not keep up with technological transformation of science, and they were considered outdated and insufficient. Things gradually changed first at the national level, with the adoption of the German *Gesetz zum Urheberrecht für die Wissenschaft* (UrhWissG, entered into force in 2018)[2], and then also at the European level, with the 2019 Directive on Copyright in the Digital Single Market (transposed in Germany in 2021)[3].

The German text contained a relatively robust exception for Text and Data Mining (TDM) for non-commercial research purposes; it also allowed the German National Library to build so-called citation archives of publicly available works for research purposes (§16a DNBG).

The European text then harmonized exceptions for TDM for scientific research purposes. Under current German law (§60d UrhG), research organisations can make reproductions of any material that they have lawful access to for TDM purposes. Moreover, the resulting corpus can be shared with a limited circle of persons for joint scientific research (which, rather regrettably, does not seem to be the case in all EU Member States).

Although the TDM exception was a welcome development, it accentuated the gap between what is allowed by copyright law, and what is required by the principles of Open Science, which today is of fundamental importance for any well-managed and ethical research project. Open Science requires open availability of research data, i.e. the possibility for everyone to reuse the data for any purpose[4]. *Vis-à-vis* this requirement, the possibility to share research data with 'a limited circle of persons for joint scientific research' is clearly insufficient. In that sense, the TDM exception failed to meet the needs of the research community; it may even have an adverse effect of creating isolated data ponds rather than robust knowledge commons by disincentivizing proper licensing of research data. Such proper licensing requires time and effort (which does not always produce the desired result), and it may be tempting for researchers to just rely on the statutory exception and settle on making their data available only to a limited circle of project partners instead.

The problem has certainly been noticed by both the German and the European legislators, as attested by proposed legislation such as the Research Data Act (*Forschungsdatengesetz*) or, at the European level, the Data Act, both intended to grant researchers access to data held by the private sector. The creation of Common European Data Spaces (including the Common European Language Data Step) is already a step in this direction. In this new paradigm the weight is shifted from property interests in data (such as copyright, database right, trade secret) to data governance and public interest (rights of access, portability), with the intention to rebalance the data market. One can point out that these efforts are not necessarily a step towards real Open Science, as the beneficiaries of the access right will certainly be limited (e.g., to public research institutions).

While these proposed developments are welcome they are still quite far from becoming a reality. For now, research infrastructures such as CLARIN, DARIAH, and Text+ are essential for providing researchers (and not only) with access to text data.

Text+ is a consortium of the national research data infrastructure (*Nationale Forschungsdateninfrastruktur*, NFDI). Text+ is not limited to existing data, but will systematically expand its portfolio in close consultation with the expert communities involved. This also includes tools to support researchers in the FAIR creation, use and provision of data throughout the entire data lifecycle.

Although the amount of data held by language data infrastructures is small compared to data used to develop the latest LLMs, their comparative advantage lies in the quality and heterogeneity. Data used to train LLMs are mostly obtained via web scraping, so their quality is low compared to carefully selected, annotated and curated multilingual language resources held by European infrastructures. One can hypothesise that this superior quality could counterbalance quantitative limitations, and allow to build leaner, less energy- and compute-hungry, but equally performant language models.

From the perspective of intellectual property rights, the role of research infrastructures is to bridge the gap between uses allowed by statutory exceptions and the requirements of Open Science. This is achieved on the one hand by sharing language data produced by research organisations with the widest possible (yet still 'strictly limited') circle of persons, and

on the other by mutualizing efforts towards copyright clearance and appropriate licensing of datasets.

Competing interests

The authors declare that they have no competing interests.

References

1. Court of Justice of the European Union, Case C-5/08, Infopaq International A/S v. Danske Dagblades Forening, 16 July 2009.
2. Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft vom 1. September 2017, Bundesgesetzblatt Jahrgang 2017 Teil I Nr. 61, pp. 3346-3351.
3. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, PE/51/2019/REV/1, OJ L 130, 17.5.2019, p. 92–125.
4. UNESCO Recommendation on Open Science, UNESCO 2021.