Dr. iur. Paweł Kamocki, IDS Mannheim

# Major developments in the legal framework concerning language resources

*Introductory Talk for the Workshop on Legal and Ethical Issues in Human Language Technologies,* LREC 2022, *Marseille, 24 June 2022*

The legal framework affecting access to and re-use of language data in the European Union has evolved very significantly since the last LREC conference (7-12 May 2018).

The General Data Protection Regulation (GDPR) entered into application on 25 May 2018, and although its content was already well-known and discussed at length during the last LREC, best practices and guidelines are still emerging. Today, we know much more especially about such aspects of the GDPR as Privacy by Design, the controller/processor dichotomy, or the data subject's right of access. In particular, specific guidelines on Virtual Voice Assistants were issued by the European Data Protection Board in 2021.

Among the directives adopted since 2018, two are of particular relevance for the language community: the Open Data Directive (of 20 June 2019) and the Directive on Copyright in the Digital Single Market (of 17 April 2019); both had to be implemented by mid-2021. The Open Data Directive has replaced the Public Sector Information Directive. Its scope is now significantly larger: while its predecessor facilitated access to and re-use of data held by public administrations as well as museums, libraries and archives, the new rules cover also data held by public undertakings and research data resulting from public funding. This opens a wealth of new data for use in language resources and language technology projects.

The Directive on Copyright in the Digital Single Market contains, among many other interesting provisions, a long-awaited copyright exception for text and data mining purposes. The mechanism is in fact two-fold, with one exception (Article 3) for research organisations and cultural heritage institutions, and another one (Article 4) for the general public. *Prima facie*, these rules allow for very wide re-use of copyright-protected material for language technology purposes, but they are in fact full of caveats and gray areas.

Finally, in 2020 the European Commission launched the European Strategy for Data. A series of proposals for Regulations (labelled, in the Anglo-Saxon way, "Acts") were adopted based on this consultation, including the Data Governance Act and the Artificial Intelligence Act. In particular, the Data Governance Act, which is now at the final stages of the legislative process and is expected to enter into application in mid-2023, contains interesting provisions on data altruism, a solution enabling individuals to 'donate' their data to registered organisations (legal entities established to meet objectives of general interest, operating on a non-profit basis and independently from any for-profit entities). The same Act also strengthens the rules concerning providers of data-sharing services.

The talk will discuss all the above-mentioned changes in the legal framework, and try to predict their impact on the language community.