All European Academies

July 2026

ALLEA STATEMENT ON THE IMPORTANCE OF THE TEXT AND DATA MINING EXCEPTION FOR SCIENTIFIC RESEARCH

The European Federation of Academies of Sciences and Humanities (ALLEA) welcomes the opportunity to contribute to the ongoing review of the Directive on Copyright in the Digital Single Market (Directive (EU) 2019/790, "CDSM Directive"). In this Statement, ALLEA underscores the importance of text and data mining ("TDM") for conducting scientific research and emphasises that scientific TDM is distinct from, and not to be conflated with, training of generative AI models. ALLEA therefore urges the EU legislature to at least preserve, and possibly improve, the mandatory research exception of Article 3 of the Directive.

Article 3 CDSM

Article 3 of the CDSM Directive provides a mandatory exception permitting research organisations and cultural heritage institutions to reproduce and extract works and other subject matter to which they have lawful access for the purposes of text and data mining (TDM) carried out for scientific research. The exception reflects a fundamental principle: where researchers already have lawful access to works (or other subject matter) containing data and information, they should be able to use computational methods to analyse that data and information and to derive information from such analyses, such as patterns, trends and correlations, without requiring additional authorisation from rights holders.

The provision was adopted in 2019 in recognition of the increasing importance of digital research methods across all scientific disciplines. It ensures legal certainty for research organisations operating across the European Research Area (ERA) and enables researchers to employ computational analysis on (inter alia) scientific publications, archival materials, databases, texts, images, audiovisual content, and other resources suitable for large-scale data analysis. Article 3 is mandatory for all Member States of the European Union (EU) and cannot be overridden by contractual provisions, thereby ensuring that researchers can effectively exercise the rights granted by the exception.

Text and Data Mining is an essential tool for contemporary science.

TDM has become a core research method in contemporary science. It enables researchers to identify patterns, relationships, trends, and anomalies within large volumes of data and information that cannot realistically be analysed through manual methods alone. As the volume of digital data continues to expand, computational analysis is increasingly indispensable to the production of new knowledge.

In the life sciences and medicine, TDM is routinely used to analyse millions of scientific publications and biomedical datasets to identify relationships between genes, diseases, biological pathways, and potential treatments. Literature mining has contributed to drug discovery, drug repurposing, epidemiological research, and the identification of novel research hypotheses. During the COVID-19 pandemic, researchers relied extensively on TDM to accelerate their understanding of the virus and potential therapeutic responses.

In the environmental and earth sciences, researchers use TDM to analyse large collections of climate data, satellite imagery, biodiversity records, environmental reports, and scientific publications. These methods help identify long-term environmental trends, monitor ecosystem changes, and improve climate modelling.

In the social sciences, TDM enables the systematic analysis of parliamentary debates, policy documents, surveys, historical records, newspapers, social media postings, etc. Political scientists use these techniques to study policy development, while sociologists employ them to analyse social networks, public discourse, and societal change at scales that would otherwise be impossible.

The humanities have likewise benefited enormously from TDM-based research. Historians analyse large corpora of digitised newspapers, archives, and correspondence to trace social, cultural, and political developments across centuries. Literary scholars use computational methods to study language change, genre evolution, and patterns across large collections of texts. Such approaches complement, rather than replace, traditional close reading and qualitative analysis.

Across all disciplines, TDM enhances the efficiency, scope, and quality of research. It allows researchers to formulate new hypotheses, identify previously unnoticed connections, conduct more comprehensive literature reviews, and improve the reproducibility and transparency of scientific inquiry.

Scientific TDM is not synonymous with training generative AI models.

Recent public debate has often associated TDM primarily with the training of generative AI models. While AI training may involve certain forms of TDM, it is important to recognise that scientific TDM is a much broader and long-established research practice that predates contemporary generative AI by many years. At the same time, the training of generative AI involves many other techniques and methods that must be distinguished from TDM.

Most research uses of TDM are not directed toward the development or training of generative AI systems. Researchers frequently use computational analysis simply to extract information, identify patterns, test hypotheses, conduct systematic reviews, classify documents, map scientific fields, or analyse historical and social phenomena. In such cases, the objective is not to create a new AI model but to extract and generate relevant information and scientific knowledge.

For example, a medical researcher mining thousands of journal articles to identify possible links between biomarkers and disease outcomes is conducting scientific TDM, not training a generative AI system. A historian analysing millions of digitised newspaper pages to study public discourse during a particular period is conducting TDM, not AI training. Similarly, an environmental scientist examining large collections of climate reports and observational data to identify long-term trends is engaging in TDM as a research method rather than developing an AI model.

Conflating scientific TDM with generative AI training risks misunderstanding the central role that computational analysis now plays in contemporary research. The legal framework governing scientific TDM should therefore be assessed on its own merits and with due regard to the needs of the research community.

Preserve or improve Article 3 CDSM.

ALLEA considers Article 3 one of the most important research exceptions introduced by the CDSM Directive as it provides legal certainty to research organisations, and strengthens Europe's scientific competitiveness. The exception enables researchers to make effective use of lawfully accessible knowledge resources and helps ensure that European science remains at the forefront of global research.

Nevertheless, ALLEA is of the opinion that the scope of Article 3 CDSM in its present form is too narrow since it designates only "organisations" as beneficiaries of the exception. As ALLEA already pointed out in its 2017 Statement on The Text and Data Mining Exception and the Enhancement of Access to Scientific Information in Europe[1], this excludes a broad range of researchers that should similarly be permitted to conduct TDM, such as researchers and scholars not affiliated to research organisations, citizen scientists, and investigative journalists. ALLEA therefore advises the EU legislature to extend Article 3 to all persons conducting bona fide scientific research, regardless of institutional status.

In any case, any narrowing of the exception would have severe consequences for research activities across a wide range of disciplines. Requiring additional permissions or licensing arrangements for scientific TDM would impose significant administrative and financial burdens on researchers and their organisations, while frustrating scientific progress. Such narrowing down would be detrimental to EU-funded research projects and collaborations, to publicly funded research institutions and, in particular, to smaller research organisations with limited resources.

The continued success of the ERA depends upon researchers' ability to analyse data and information using the best available methods. TDM is no longer a niche activity but a foundational component of contemporary scholarship across the sciences, social sciences, and humanities.

1 'The Text and Data Mining Exception and the Enhancement of Access to Scientific Information in Europe', Statement prepared by the ALLEA Permanent Working Group Intellectual Property Rights. November 2017. Available from: https://www.allea.org/wp-content/uploads/2017/11/PWGIPR_Statement_TDM_2017.pdf.

For these reasons, ALLEA urges EU lawmakers to preserve the Article 3 TDM exception in its current form or, preferably, improve it to extend to all researchers conducting bona fide scientific research, regardless of institutional status. Maintaining a robust mandatory exception is essential to safeguarding scientific freedom, promoting innovation, and ensuring that European researchers can continue to generate knowledge for the benefit of society.

About ALLEA

ALLEA is the European Federation of Academies of Sciences and Humanities, representing 61 academies from 40 countries in Europe. Since its foundation in 1994, ALLEA speaks out on behalf of its members on European and international stages, promotes science as a global public good, and facilitates scientific collaboration across borders and disciplines. Learn more: http://www.allea.org.

About the Statement

This ALLEA statement has been prepared by ALLEA's Task Force on Protecting and Sharing Data. With its work, ALLEA seeks to ensure that science and research in Europe can excel and serve the interests of society. Read more about the ALLEA's Task Forces and their members: https://allea.org/task-forces/.

ALLEA | Jägerstr. 22/23, 10117 Berlin, Germany
www.allea.org
secretariat@allea.org