What Is a 'Research Organisation' and Why It Matters: From Text and Data Mining to AI Research

Quintais, J.P.

DOI
10.1093/grurint/ikaf030

Publication date
2025

Document Version
Final published version

Published in
GRUR International

GRUR International, 74(5), 2025, 397–398

https://doi.org/10.1093/grurint/ikaf030
Advance Access Publication Date: 19 March 2025
Editorial

OXFORD

JOÃO PEDRO QUINTAIS*

What Is a 'Research Organisation' and Why It Matters: From Text and Data Mining to AI Research

The concept of a 'research organisation' has significant implications across various domains of EU information law, including copyright and artificial intelligence (AI) regulation. Defined in the Copyright in the Digital Single Market Directive (Directive (EU) 2019/790, hereinafter CDSMD), this concept plays a crucial role in determining the legal obligations and rights of entities engaging in activities such as text and data mining (TDM) and AI research, or data access for research purposes (see, generally, European Commission 2024). By examining how this definition interacts with legislative frameworks like the CDSMD and the AI Act (Regulation (EU) 2024/1689), this editorial examines its critical role in EU digital regulation of research and highlights areas of legal uncertainty.

The concept of a 'research organisation' is defined in Art. 2(1) CDSMD.

A 'research organisation' is, first, 'a university, including its libraries, a research institute, or any other entity'. This includes a 'wide variety of entities', such as research institutes and hospitals that engage in research activities (Recital 12).

Second, the primary objective of the organisation must be to conduct scientific research – including research in the natural and human sciences (Recital 12) – or to carry out educational activities that also involve scientific research.

Third, while different legal forms and structures are permitted (Recital 12), such scientific research must be undertaken under one of the following conditions: (i) on a not-for-profit basis *or* with all profits reinvested in its scientific research, or (ii) pursuant to a public-interest mission recognised by a Member State. A public-interest mission might be reflected, for example, through public funding, provisions in national laws, or public contracts (Recital 12). Notably, this does not preclude research organisations from collaborating with private parties in the course of their scientific research. For instance, Recital 11 explains that existing EU research policy encourages partnerships between research organisations and the private sector. In such contexts, research organisations may rely on their private partners, including their technological tools, to carry out TDM.

Finally, the scientific research must be conducted in a manner ensuring that the results cannot be accessed on a preferential basis by an undertaking that exercises a decisive influence over the organisation. This concern applies, for instance, where a commercial entity holds a controlling position, such as through shareholder or membership status, and thereby gains preferential access to the organisation's research outcomes (Recital 12).

Overall, this is a complex definition with potential for legal uncertainty that opens the door to varying national implementations, even considering the interpretative guidance provided in Recitals 11 and 12.

But why does this definition matter?

The qualification of an entity as a 'research organisation' is pivotal in defining its legal obligations when engaging in TDM activities for the development of generative AI models, a subset of general-purpose AI (GPAI) models as defined in the AI Act (see Art. 3(63), Recitals 99 and 105).

Training and developing AI models often involve a range of activities (e.g., web scraping, pre-training, training, and finetuning), many of which entail copyright-relevant reproductions. Under EU copyright law, such activities frequently qualify as TDM, which is primarily regulated through the exclusive right(s) of reproduction and the two TDM exceptions in the CDSMD.

TDM is defined broadly in Art. 2(2) CDSMD as 'any automated analytical technique aimed at analyzing text and data in digital form in order to generate information, including but not limited to patterns, trends, and correlations'. It is now largely accepted that this definition encompasses the training activities (*lato sensu*) necessary for developing AI models, including generative AI models. The AI Act reinforces this understanding in Recital 105, even if it remains unclear whether the concept of TDM extends to all copyright-relevant acts involved in the training and development of AI models (Quintais 2025; contra, see Dornis 2024).

Articles 3 and 4 of the CDSMD provide two mandatory exceptions related to TDM. Article 3 applies to TDM activities conducted for scientific research purposes by 'research organisations' and 'cultural heritage institutions' on works or subject matter to which they have lawful access, subject to specific conditions (Margoni 2024; Szkalej 2024). Article 4 provides a broader exception for reproductions and extractions of lawfully accessed works or subject matter for the purposes of TDM. However, this exception allows rights holders to reserve their rights through 'machine-readable means' in cases where content is made publicly available online. Such reservations may be implemented via metadata or the terms and conditions of a website or service. This mechanism is commonly referred to as the 'opt-out' requirement (see e.g. Mezei 2024; Peukert 2024).

* Associate Professor, University of Amsterdam, Institute for Information Law (IViR), The Netherlands. Orcid: https://orcid.org/0000-0001-6462-2392. The author thanks Nuno Sousa e Silva and Kacper Szkalej for their comments.

Significantly, the opt-out mechanism does not affect the application of the TDM exception for scientific research under Art. 3. Consequently, entities that qualify for the Art. 3 exception are, by virtue of that qualification alone, exempt from most of the additional copyright-related obligations imposed on GPAI model providers by Art. 53(1)(c) AI Act. This provision requires GPAI model providers to establish policies to respect EU copyright law, including identifying and honouring opt-out requests expressed under Art. 4(3) CDSMD. At this stage, it is worth reminding that the copyright obligations in the AI Act are imposed on *GPAI model providers* in relation to all types of GPAI models. That is to say, these provisions do not directly apply to upstream players (e.g. an entity which only provides datasets) or – as a general rule – to downstream players, such as (non-vertically integrated) AI system providers or deployers.

In summary, an entity that qualifies as a research organisation and engages in TDM at various stages of AI model development faces significantly reduced legal exposure under copyright law. This principle was recently illustrated in the *Kneschke v LAION* judgment by the Hamburg Regional Court. In this case, the Court rejected photographer Robert Kneschke's request for an injunction against the German non-profit organisation LAION concerning the unauthorised use of his photograph. The photograph had been downloaded from a stock photo agency's website and included in an open-source dataset for AI training.

The court's decision hinged on the classification of LAION's preparatory copyright-relevant TDM activities (the creation of a dataset) as 'scientific research'. The non-commercial nature of LAION's research was a decisive factor in qualifying it as a research organisation under the national provision implementing Art. 3 CDSMD (Sec. 60d of the German Copyright Act). The Court held that the use of the dataset by commercial entities to train their AI models was irrelevant to the non-commercial character of LAION's research activities in creating the dataset. Likewise, the fact that individual LAION members engaged in paid work with such companies did not render those companies' activities attributable to LAION (Goldstein, Stuetzle & Bischoff 2024).

This qualification was crucial, as the court suggested in *obiter dictum* that if LAION were not considered a research organisation, it would be assessed under the national equivalent of Art. 4 CDSMD and likely found non-compliant with the opt-out requirement.

Moving beyond copyright, the legal definition of a research organisation may have broader implications under the AI Act. The Act applies to a wide range of stakeholders – providers, deployers, importers, distributors, product manufacturers, authorised representatives, and affected persons – who are involved in the placing on the market and putting into service of GPAI models and AI systems in the EU, or who are otherwise established or located within the Union (Art. 2(1)).

But some players and activities are excluded from the scope of application of the AI Act. For the purposes of this discussion, the most relevant exclusions are outlined in Arts. 2(6) and 2(8). Article 2(6) specifies that the AI Act does not apply to AI systems or models, including their outputs, that are specifically developed and *put into service* solely for scientific research and development purposes. Meanwhile, Art. 2(8) excludes the application of the Act to 'any research, testing, or development activity regarding AI systems or AI models prior to their being *placed on the market* or *put into service*.'

Some of the terminology used in these provisions requires clarification, as it may have significant consequences:

– 'Placing on the market' refers to the first instance of *making* an AI system or GPAI model *available on the Union market* (Art. 3(9)).
– 'Making available on the market' is defined as the supply of an AI system or GPAI model for distribution or use in the Union market in the course of a commercial activity, whether for payment or free of charge (Art. 3(10)).
– 'Putting into service' means supplying an AI system for first use directly to the deployer or for self-use in the Union for its intended purpose (Art. 3(11)).

The rationale for these exclusions, as outlined in Recital 25, is twofold. First, the AI Act aims to support innovation, uphold the freedom of science, and avoid impeding research and development activities. Second, it seeks not to affect scientific research and development activities related to AI systems or models prior to their placement on the market or being put into service.

This approach is further reinforced in the specific context of copyright-related obligations imposed on GPAI model providers. Recital 109 notes that compliance requirements are excluded for individuals or entities developing or using models for non-professional or scientific research purposes, although they are encouraged to voluntarily adhere to these requirements.

In my view, the net effect of these exclusions and definitions is as follows: if an entity qualifies as a research organisation engaged in TDM for purposes outlined in Art. 3 CDSMD, those activities are largely exempt from compliance with the obligations of the AI Act.

However, this does not amount to a blanket exemption for all activities of research organisations under the AI Act. For example, the exclusion does not cover fine-tuning modifications made to a scientific model that is subsequently commercialised (Recital 109) or models initially developed for scientific purposes that are later made available on the market (Recital 25). In these cases, the application of the AI Act obligations to research organisations and their research activities will need to be assessed on a case-by-case basis.

In conclusion, the concept of a 'research organisation' plays a pivotal role in EU digital regulation but is surrounded by significant legal uncertainties. Due to its origin in the CDSMD, the concept is susceptible to divergent national implementations leading to fragmentation, which conflict with its application in directly applicable Regulations like the AI Act.

These inconsistencies create ambiguities that affect compliance obligations under copyright law, AI regulation and other data and platform regimes that rely on this concept. To ensure legal coherence and avoid fragmentation, the concept should be regarded as an autonomous concept of EU law. Clarification from the Court of Justice of the European Union may ultimately be required to establish its definitive interpretation and harmonise its application across Member States while accommodating the diverse contexts in which the term operates.