

XXIV Seminário Internacional sobre Sociedade Informacional e Propriedade Intelectual

Marek Salamonowicz

University in Bydgoszcz

Declarations of intent by entities holding rights to intangible assets regarding the conditions for the admissibility of text and data mining (TDM)

Marek Salamonowicz Department of Civil Law Kazimierz Wielki University in Bydgoszcz





Content of the exposition

- The essence of TDM and legal basis.
- Technical forms of submitting an opt-out statement.
- Analysis of current case law.
- Determining the legal nature of an opt-out statement.



TDM – concept and legal significance

- Any automatic technique for analyzing digital data to generate information (patterns, trends, correlations) – Article 2(2) of the DSM.
- Recital 8: “automatic mathematical analysis of information in digital form (text, sound, images, data).”

Activities within the framework of TDM

- Identification and access to previously stored materials (usually in databases)
- Copying significant fragments
- Converting them into a format suitable for the technology used
- Extracting and analyzing data to detect relationships and patterns
- Use of results: presentation, verification (sometimes with text fragments)



TDM – functions

- The use of computers (statistical and computational analysis) to find patterns, new trends, and gain knowledge based on large data sets.
- This method is used in astronomy, mechatronics, medicine, linguistics, musicology, marketing, finance, meteorology, AI, business decisions, and the creation of new technologies, among other fields.


International aspect




- In many EU countries (United Kingdom 2014, France 2016, Estonia 2017, Germany 2017), regulations legalizing TDM were already in force before the DSM Directive came into effect. Poland – 2024.
- In the US, Canada, and Israel, such activities are permitted on a broader scale than provided for in the Directive.
- Japanese copyright law (Article 30-4), Brazilian AI bill No. 2338/2023.




Problem



legal uncertainty – whether certain actions (e.g., copying works, downloading databases) infringe copyright or database producer rights. (Recitals 8, 10, 11 of the DSM Directive).

A large orange circle on the left side of the slide, partially cut off by the edge.

TDM performed for commercial purposes

- It is possible for copyright holders to expressly prohibit the use of TDM for commercial purposes by appropriate means, for example **by means of machine-readable measures in the case of content that has been made publicly available on the internet** (Article 4(4) DSM).
 - Reproduction and retrieval for TDM purposes “**may only be stored for the time necessary for the purposes of the text or data mining operation concerned**” (Article 4(2)).
 - This introduces a time limit on the duration of such storage, which does not exist when TDM is used for research purposes.
- 
- Three blue curved lines of varying lengths and orientations, located in the bottom right corner of the slide.

Polish Act on Copyright and Related Rights Reproduction of published works for the purpose of text and data mining

- Art. 26(3). 1. It is permitted to reproduce disseminated works for the purpose of text and data mining, unless the rightholder has expressly stated otherwise.
- 2. The reservation referred to in paragraph 1 shall be made explicitly and in a manner appropriate to the way in which the work is made available. In the case of **works made publicly available in such a way that anyone can access them at a place and time of their choosing**, the reservation shall be made in a machine-readable format within the meaning of Article 2(7) of the Act of August 11, 2021 on open data and the reuse of public sector information (Journal of Laws of 2023, item 1524) together with metadata.
- – a file format structured in such a way that computer programs can identify, recognize, and retrieve specific data and its internal structure.



Metadata in files

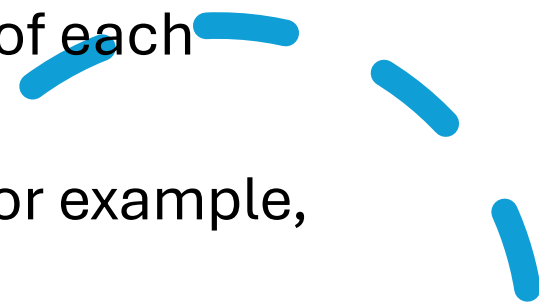
- This is additional information stored inside the file, but next to the actual content.
- Example: a JPG file contains not only an image, but also EXIF/IPTC data, e.g., who took the photo, the date, copyright information.
- Practical application A photographer can save the following in a JPG file:
- “Author: John Smith” “Rights: © 2025, use only with permission”
- Graphics programs and browsers can read this.
- In the context of TDM, you can enter the designation “no TDM” there, which the bot recognizes as a restriction.

HTTP headers

- These are messages exchanged between the user's browser and the server each time a web page is opened.
- The server can add additional information to them.
- Example: When a user visits a website, the server may send the following header:
- X-Robots-Tag: noai, notdm
- This means: “this content must not be used for AI/TDM training.”
- A normal user cannot see this, but a bot (e.g., a data-collecting crawler) will read it.



• Robots.txt

- It is a simple text file (like a notepad) located in the root directory of each website.
 - You can view it by entering, for example, www.onet.pl/robots.txt.
 - How it works: It informs bots which parts of the site they can and cannot download.
 - Example: User-agent: GPTBot Disallow: /
→ says: “a bot called GPTBot is completely prohibited from downloading content from this website.”
 - Similarly: It's a bit like a sign stuck on a door: “No entry – private property.”
- 

Comparison



Differences between them (for lawyers)



Metadata → related to the file itself (goes with it).



HTTP headers → only work on the Internet, when downloading.



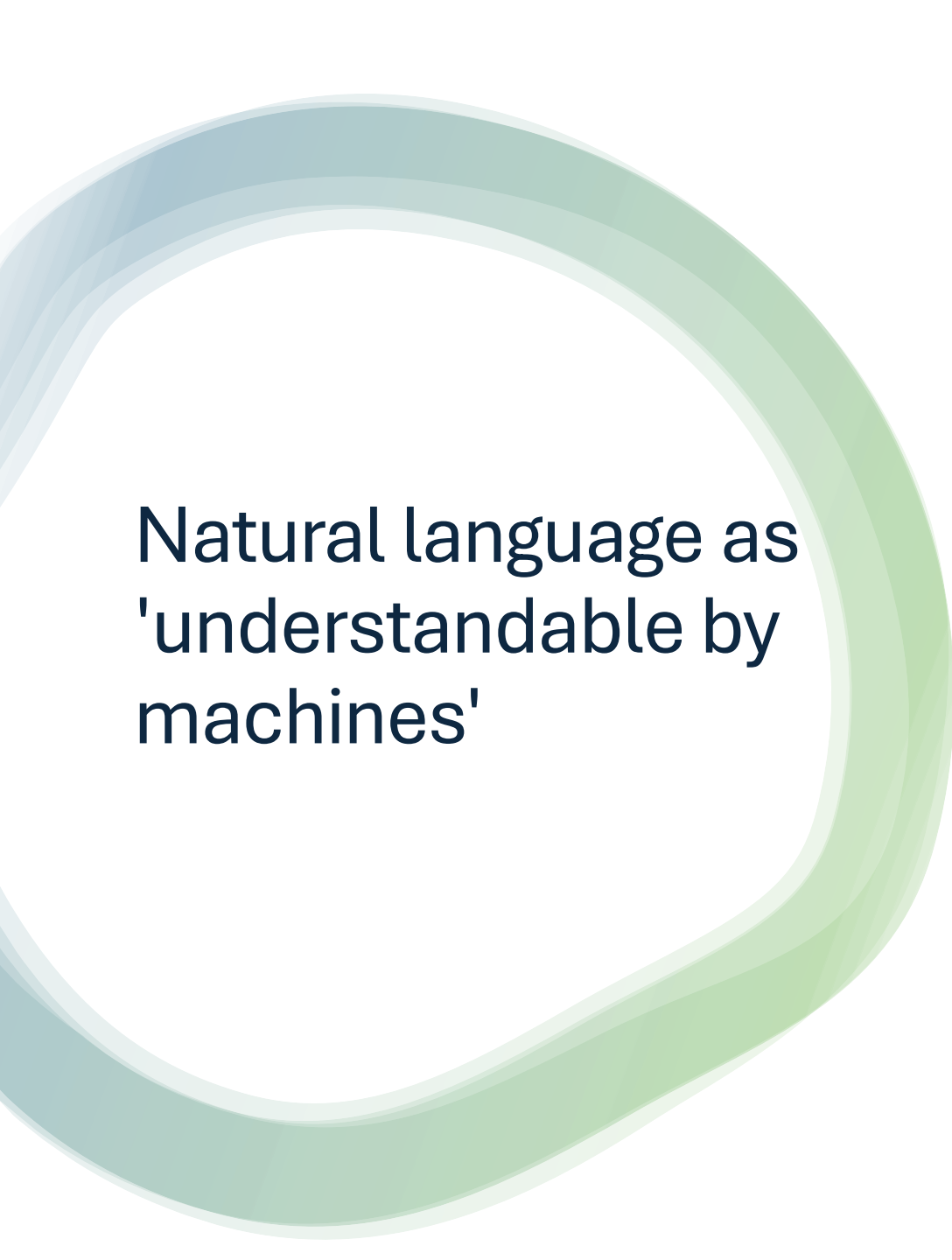
Robots.txt → declaration for the entire website, instructions for bots.

Case: Kneschke v. LAION – LG Hamburg (310 O 227/23)

- Plaintiff: photographer Dirk Kneschke, owner of the rights to the photographs.
- Defendant: LAION association, which creates datasets for AI training.
- Issue: use of photographs in the LAION database for training AI models.
- Main legal issue: effectiveness of the opt-out statement (reservation of use) in accordance with Article 4(3) of the DSM Directive and Section 44b of the UrhG.
- Key question: Is a reservation in natural language (“do not use for AI training”) sufficiently clear for machines and legally effective?
- Judgment: The Hamburg Regional Court ruled that such a reservation is effective and binding on LAION.

Machine readability of opt-out

- Contrary to the defendant's position, the notice was also machine-readable within the meaning of Section 44b(3) sentence 2 UrhG. **The requirements in this regard are no higher than those for human machine readability;** however, the notice was written in capital letters. Furthermore, the text was also recognizable as a notice by a computer program.



Natural language as 'understandable by machines'

- "However, the Board is inclined to treat a claim drafted solely in 'natural language' as 'machine-understandable' ... Nevertheless, the question of whether and under what specific conditions **a claim expressed in 'natural language' can also be considered 'machine-understandable'** will always have to be decided depending on the technical development existing at the time of use of the work."

Conclusions from the case AI Act and 'most advanced technologies'

- "In this regard, the European legislator has also provided in the AI Regulation that **AI model providers must have a strategy in place**, in particular to identify and respect legal reservations notified in accordance with Article 4(3) of the DSM Directive 'also using the most advanced technologies' (Article 53(1)(c) of the AI Regulation).
- However, these 'most advanced technologies' clearly include AI applications that are capable of recognizing the content of **text written in natural language.**"

C-250/25 Like Company v Google — background to the case

- Plaintiff: Like Company Kft. (Hungary), local publisher.
- Defendant: Google Ireland Ltd.
- Dispute: use by AI chatbots of press content (including an article about “dolphins in Lake Balaton”) generated in responses to users.
- National court: Budapest Környéki Törvényszék — referred questions for a preliminary ruling to the CJEU.
- Key issue: whether LLM training and responses infringe the rights of press publishers (Article 15 DSM) and whether they can benefit from the TDM exception (Article 4 DSM).
- Status: case pending (preliminary ruling request of April 3, 2025).

C-250/25 Like Company v Google — question 1 (CJEU)

- "Should Article 15(1) of Directive (EU) 2019/790 [...] and Article 3(2) of Directive 2001/29 [...] be interpreted as meaning that **the display in the responses of a chatbot based on an LLM model of text that is partially identical to the content of press publishers' websites, the length of which is such that it is already protected under Article 15 of Directive 2019/790, constitutes a case of making available to the public?** If the answer to that question is in the affirmative, is it relevant that [the responses in question] are the result of a process in which the chatbot merely predicts the next word on the basis of observed patterns?"
- Source: CURIA, Summary of the request for a preliminary ruling – Case C-250/25 (April 3, 2025)

C-250/25 — significance for TDM i LLM (PL)

- First case before the CJEU concerning chatbots/LLMs and the rights of press publishers (Article 15 DSM) and TDM (Article 4 DSM).
- • May decide: whether LLM training constitutes ‘multiplication’ (Article 2 InfoSoc) and when it falls within the TDM exception.
- It will answer whether displaying content in chatbot responses constitutes ‘making available to the public’ (Article 3 InfoSoc).
- Direct impact on: the scope of permitted commercial TDM, the significance of the opt-out in Article 4(3) DSM, and AI business models.
- For publishers: potential remuneration/consent for uses beyond ‘single words or very short extracts’.
- Status as of August 21, 2025: case pending — no judgment (CJEU, C-250/25).

+

•

○

„Declaration of intent" in practice – how to submit it correctly

- Goal: create a legal and technical signal that crawlers will actually see and that can be proven.
- Layers (recommended package):
- Technical layer (online): robots.txt (directives for relevant bots), HTTP headers (e.g., X-Robots-Tag), metatags in and/or in files (EXIF/IPTC with “no-TDM” flag), site map with metadata/license.
- Legal layer: clause in the terms and conditions/license (clear, highlighted, assigned to specific content), message in the footer/on content pages,
- in databases: separate provisions on extraction and reuse, prohibition on building datasets.
- Evidence layer: archiving of page versions (hash, timestamp), server logs showing bot access.

Conclusions

- Legal nature of the opt-out statement
- The opt-out statement is an explicit and unambiguous reservation by the right holder that limits the right to use works within a specific scope of TDM.
- This statement is an instrument for exercising an exclusive right (control right).
- Legal effect: Negative → excludes the statutory TDM exception (Article 4 of the DSM).
- The issue of the territorial scope of the reservation – EU law, Polish law.

Thank You for attention.

