Text and Data Mining –
Copyright and Data Protection Issues

5th International Summer School on Big Data and Machine Learning
21 August 2019

Prof. Dr. Anne Lauber-Rönsberg
Institute of Intellectual Property Law, Technology Law and Media Law

Data are the new oil…

but analogy is not accurate, as …

- data is not scarce, but abundant, can be replicated and infinitely used
- data becomes more useful the more it is used
- the use of data may cause (completely different) legal issues…

The Economist | Topics | Current edition | More | Sub

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Print edition | Leaders ›
May 6th 2017

A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants that deal in data, the oil of the digital era. These titans—Alphabet (Google's parent company), Amazon, Apple, Facebook and Microsoft—look unstoppable. They are the five most valuable listed firms

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

TDM:

- creation of the corpus: selection of relevant data, copying, extraction, transformation into machine-readable format, structuring the input data, adding metadata…
- Analysis: deriving patterns within the structured data
- evaluation and interpretation of the output

Legal Issues

Under which circumstances may material be used for the creation of a corpus and its analysis?

There may be legal restrictions by

Copyright law, if the material is copyright protected	Data protection law, if personal data is involved

Copyright Law in a Nutshell

- **Aim:** to incentivize creations.
- **Protected subject-matter:**
  - **works:** texts, photos, films, maps, databases, if: "author's own intellectual creation"
  - **not protected**: ideas, facts, principles, methods, ...
  - **related rights**: databases (EU); phonogram producers; photos (Germany), ...
- **Scope of Protection**: Exclusive rights granted to author.

➡ Acts of reproduction, distribution and of communications to the public only permitted, if either license or statutory permission (e.g. quotation).

➡ Furthermore moral rights, e.g. attribution right.

Why is TDM a Copyright Issue?

- Information ≠ copyright protection
- BUT: information is usually conveyed by some copyrightable “container”: text, picture, film, database, …
- When these “containers” are processed in order to extract the relevant (non-protectable) information, this usually requires reproductions of the copyrighted material: storing for further processing, digitisation, … .
- These acts of reproduction interfere with the copyright protection – even if full and legal access to the data body (e.g. right to read).

TDM is permitted if

the rightowner has granted a **contractual license**

BUT: licensed-based approach not satsfying, since

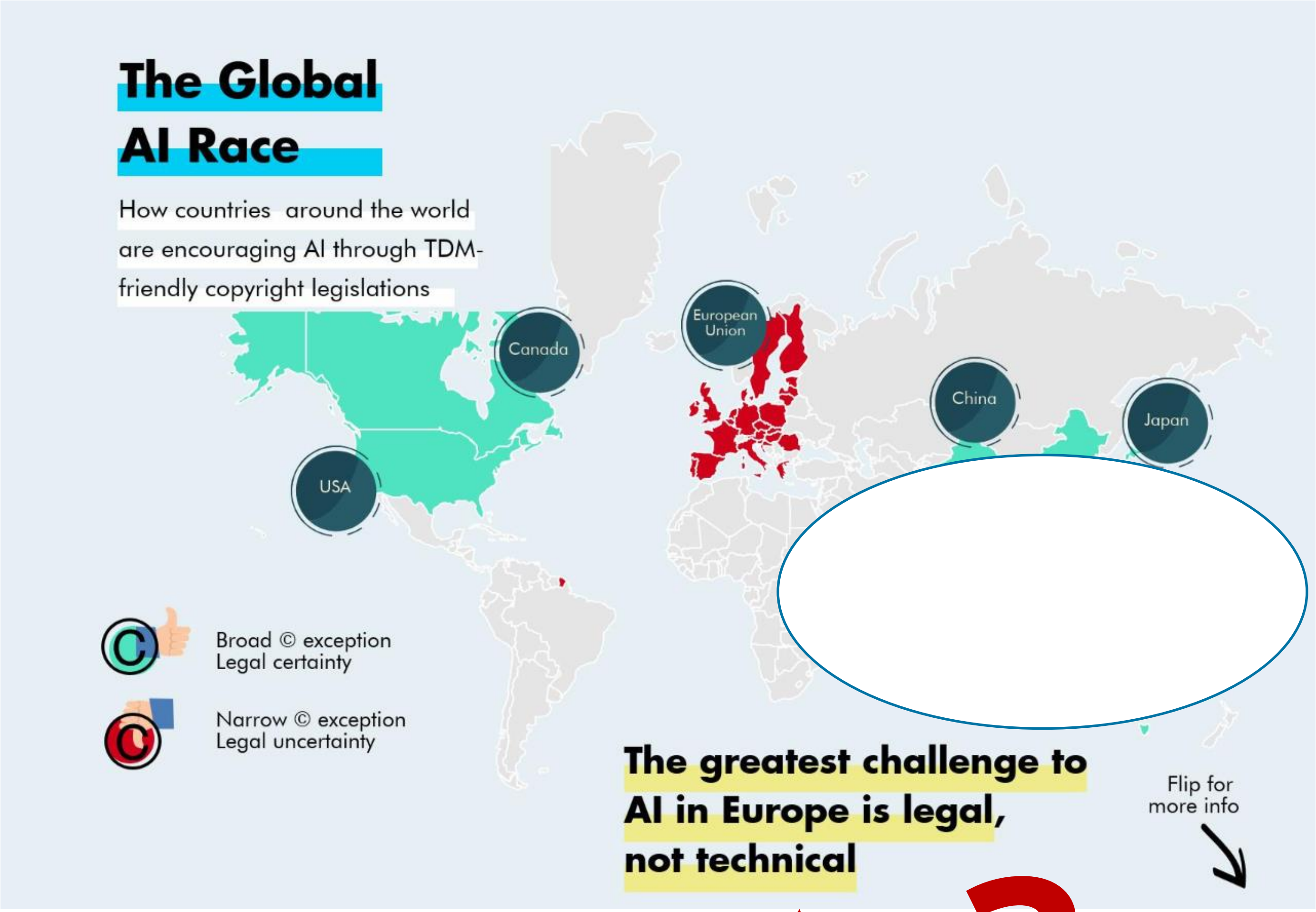
- there may be contractual restrictions
- fees, expenses for administering contracts

there is a **statutory provision** permitting this use

- temporary reproductions: Art. 5(1) Directive 2001/29

- Specific provisions very disparate until now

A Lobbying Chart (by companies represented in the „European Alliance for Research Excellence“)



?

http://eare.eu/assets/uploads/2018/06/Global-AI-Race.pdf

Heterogeneous Regulations:

- U.S.: fair use doctrine – allows TDM for commercial and non-commercial purposes
- Japan: has amended copyright law as of 1 January 2019
- EU: no fair use clause, but some EU Member States have specific provisions on TDM, e.g. Germany, UK.

Section 60d German Copyright Act: Text and data mining

(1) In order to enable the automatic analysis of large numbers of works (source material) **for scientific research**, it shall be permissible

1. to **reproduce the source material**, including automatically and systematically, in order to create, particularly by means of normalisation, structuring and categorisation, a corpus which can be analysed and

2. **to make the corpus available to the public** for a specifically **limited circle of persons for their joint scientific research**, as well as to individual third persons for the purpose of monitoring the quality of scientific research.

In such cases, the user may only pursue **non-commercial purposes**.

Section 60d German Copyright Act: Text and data mining

(…)

(3) Once the research work has been completed, the **corpus and the reproductions of the source material shall be deleted**; they may no longer be made available to the public. It shall, however, be permissible to transmit the corpus and the reproductions of the source material to (publicly accessible libraries, archives, museums and educational establishments, which neither directly nor indirectly serve commercial purposes) for the purpose of long-term storage.

Section 60d German Copyright Act: Text and data mining

(…)

(3) Once the research work has been completed, the **corpus and the reproductions of the source material shall be deleted**; they may no longer be made available to the public. It shall, however, be permissible to transmit the corpus and the reproductions of the source material to (publicly accessible libraries, archives, museums and educational establishments, which neither directly nor indirectly serve commercial purposes) for the purpose of long-term storage.

Section 60h: rightholders have right to remuneration.

New Copyright Directive!

- Directive of 17 April 2019 on copyright and related rights in the Digital Single Market
- To be transposed into national laws by 7 June 2021
- Requires reform of German Copyright Act and harmonizes regulatory framework within EU

Rec. 9 Directive 2019/790

Text and data mining can also be carried out in relation to mere facts or data that are not protected by copyright, and in such instances no authorisation is required under copyright law. There can also be instances of text and data mining that do not involve acts of reproduction or where the reproductions made fall under the mandatory exception for temporary acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC, which should continue to apply to text and data mining techniques that do not involve the making of copies beyond the scope of that exception.

This clarification corresponds to the current legal situation.

Art. 3 Directive 2019/790: TDM for the purposes of scientific research

1. Member States shall provide for an exception (...) for reproductions and extractions made **by research organisations and cultural heritage institutions** in order to carry out, for **the purposes of scientific research**, text and data mining of works or other subject matter to which they have **lawful access**. (*right to read = right to mine*)

2. Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results. *(i.e. permanently)*

Art. 3 Directive 2019/790: TDM for the purposes of scientific research

3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.

4. Member States shall encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively.

Art. 4 Directive 2019/790: Exception or limitation for TDM

1. Member States shall provide for an exception or limitation (…) for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.

Art. 4 Directive 2019/790: Exception or limitation for TDM

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

(So opting out possible)

4. This Article shall not affect the application of Article 3 of this Directive.

Legal Issues

Under which circumstances may material be used for the creation of a corpus and its analysis?

There may be legal restrictions by

Copyright law,
if the material is copyright protected

Data protection law,
if personal data is involved

Data Protection Law and the GDPR

- Aims: Protection of humans, not of data:
  protection of privacy and “informational sovereignty”
- EU: concerned with data processing by **private actors** as well as by **public actors**. Different approach in the U.S.
- General Data Protection Regulation 2016/679 (GDPR)
  - in force since 2016, applicable since 25 May 2018
  - GDPR is applicable to any processing of personal data, not just to cross-border activities!

Applicability: Personal Data

- Broad definition of personal data = any information relating to an identified or identifiable natural person (Art. 4 GDPR)
- not required that all the information enabling the identification of the data subject must be in the hands of one person; sufficient if controller can obtain additional information which may likely reasonably be used to identify the data subject
- Example: “dynamic” IP addresses (CJEU, C-582/14, ECLI:EU:C:2016:779 – Breyer/Germany)
- More restrictive: sensitive data (Art. 9 GDPR)

Personal Data?

- **anonymised data:**
  if irreversibly and effectively anonymized, so no re-identification possible ≠ personal data
  – GDPR not applicable!
  Still a realistic concept in the age of Big Data?

- **pseudonymised data**
  = personal data, if re-identification possible and likely reasonable

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study

Adam Tanner Contributor ⓘ
I write about the business of personal data.

A Harvard professor has re-identified the names of more than 40% of a sample of anonymous participants in a high-profile DNA study, highlighting the dangers that ever greater amounts of personal data available in the Internet era could unravel personal secrets.

Harvard Professor Latanya Sweeney

From the onset, the Personal Genome Project, set up by Harvard Medical School Professor of Genetics George Church, has warned

https://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/#7a07080692c9

Data Protection Principles (Art. 5 GDPR)

If the TDM includes personal data, the processing must be in line with the GDPR.

Core principles:

- Lawfulness, fairness and transparency
- Purpose limitation
- Data minimisation
- Accuracy
- Storage limitation
- Integrity and confidentiality
- Accountability

Some Challenges in relation to Big Data:

Lawfulness

- Processing only lawful, if based on one of the grounds listed in Art. 6 (1) GDPR.
- **Consent**: informed and freely given in relation to “one or more specific” purpose
- Data processing lawful also without consent e.g. according to Art. 6 (1) f), if necessary for **legitimate interests** pursued by the controller, except where overridden by the interests or fundamental rights and freedoms of the data subject

Some Challenges in relation to Big Data:

Purpose limitation

- Personal data must be collected for specified, explicit and legitimate purposes – requiring any processing to have a clearly defined purpose
- Challenge in big data contexts, because at the time when the data is collected, the purpose of processing may still be unclear.
- Provision also includes the prohibition to further process personal data in a way incompatible with the initial purposes (re-purposing). Still unclear, which processing is to be considered compatible / incompatible.

Some Challenges in relation to Big Data:

Data minimisation

- The type and scope of the processed data must be limited to what is necessary in relation to the purpose for which they are processed.
- Realistic scenario with regard to big data?

Some Challenges in relation to Big Data:

Storage Limitation

According to Art. 5(1) e) GDPR, “personal data must be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed;
personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes (...) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject”.

Conclusion

- As for **copyright law**, the European legislator has found a viable compromise, which has to implemented into national laws by 2021.
- Conditions for complying with **data protection laws** can only be judged on a case by case basis. But still many challenges to ensure that automatic data analyses compliant with GDPR.