Text and Data Mining

Prof. em. Frank GOTZEN
President ALAI International

Conference: ***Challenges of the Directive 2019/790 on Copyright and Related Rights in the Digital Single Market***

24-25 October 2019

Law Faculty, Jagiellonian University, Larisch Palace, Bracka St.12, Cracow, Poland

Fourth Industrial Revolution

- **High-speed mobile internet**
- **Cloud technology**
- **Big data analytics**
- **Internet of things**
- **Artificial Intelligence**

TDM as a tool in the Fourth Industrial Revolution

Automated analytical tools are necessary to master the unlimited amounts of data

Since 2003 more data are created in two days time than in the whole history before

Datamining searches for significant relations and combinations through a predefined simple algorithm or through an evolving machine- learning algorithm.

Noticing patterns that would escape the human eye.

Definition in the EU

Art. 2 (2) of DIRECTIVE (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ 17.5.2019, L130/92

‘text and data mining’ means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations

Scientific research needs data

- **Analysing correlations: medical data – disease mapping, clinical trial; economic data, patent mapping; climate change**
- **Discovering patterns and trends: mobility, economic decisions, identify and remove fake content**

Business needs data

Data are the new gold mine:

Turning raw data into smart data through selection and combination with trends

- **Analysing customers' general behaviour and defining individual profiles and preferences, predictive marketing**
- **Discovering trends: trends in social media before product launch**
- **Predict movements in financial markets**
- **Finding talent in the media and in sports**
- **Journalism**

The Government needs data

Good government depends on millions of informations

Smart use of Big Data is an instrument of Power

- **Social Security**
- **eHealth-platforms**
- **Tax policy, detecting fraude**
- **Economic decisions**
- **Justice and Police, crime detection**

AI needs data

The development of increasingly intelligent computational systems for the analysis of information in digital form is dependent on the input of large amounts of data.
This is certainly the case for smart robots .

Data need AI

The sheer scale of massive data input needs a structure that allows insight

The data are not always free to use

- **Privacy, personal data and image problems**

- **IP protection problems:**

- Copyright: images, music, text, graphics
- Database rights: structured data sets

How to solve the copyright problem?

- **Getting round the obstacle by interpretation**

- ***TDM is a temporary technical step (art.5.1 of the Infosoc Directive)***

Objection: If any of the copies is permanent it does not fall under art. 5.1 Infosoc Directive

- ***The purpose of TDM is to extract information and not to duplicate the data***

Objection: Extracting the information through TDM requires previous copying

- ***A TDM copy is not intended to appeal to a human consumer. It is a mere tool in a learning process for a machine***

Objection: Unvisable technical copies inside a machine have been considered a reproduction in principle (exempted only if under art.5.1 Infosoc Directive)

- ***There is no harm to the exploitation of the original work because of no competition***

Objection: competition is not the core concern of Copyright

- **Removing the obstacle by introducing a new exception**

New exceptions in the EU

Art. 3 and 4 of DIRECTIVE (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ 17.5.2019, L130/92

TDM for scientific research

Art. 3 Member States **SHALL** provide for an exception to the rights provided for in:

- Article 5(a) of Directive **96/9**/EC: **reproduction** of copyrighted database
- Article 7(1) of Directive **96/9**/EC: **extraction** from sui generis database
- Article 2 of Directive **2001/29**/EC: **reproduction** of author's work, performer's performance, phonogram producer's phonogram, producers of first fixations of films, broadcasting organizations' fixations of broadcasts
- Article 15(1) of Directive **2019/790**: press publications online

TDM for scientific research

- By a **research organisation**:

Art. 2 (1): this means a **university**, including its libraries, a **research institute** or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research, including hospitals:

(a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or

(b)pursuant to a public interest mission recognised by a Member State;

in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organization

Quid individual researchers ?

- By a **cultural heritage institution**:

According to Art. 2 (3): this means a publicly accessible **library** or **museum**, an **archive** or a film or audio **heritage** institution

TDM for scientific research

- For the **purposes of scientific** research

- Concerning works or other subject matter to which they have **lawful access**. This means (recital 14):

- access to content based on an **open access** policy
- access through **contract**ual arrangements between rightholders and research organisations or cultural heritage institutions, such as subscriptions. Persons attached thereto should be deemed to have lawful access.
- access to content that is freely **available online**.

TDM for scientific research

It is allowed to:

- Make copies
- Store copies : 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.

Potential harm to rightholders being minimal,

no compensation regime needed (rec.17)

TDM for scientific research

- 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. This may be necessary in case of a high number of access requests.

  e.g. through IP address validation or user authentication (rec.16), CAPTCHA challenges, limiting downloading range

- Member States shall encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning secure storing and the preservation of security and integrity of networks and databases

TDM in general

Art. 4 Member States **SHALL** provide for an exception *or limitation* to the rights provided for in:

- Article 5(a) of Directive **96/9**/EC: **reproduction** of copyrighted database
- Article 7(1) of Directive **96/9**/EC: **extraction** from sui generis database
- Article 2 of Directive **2001/29**/EC: **reproduction** of author's work, performer's performance, phonogram producer's phonogram, producers of first fixations of films, broadcasting organizations' fixations of broadcasts
- *Article 4(1)(a) and (b) of Directive **2009/24**/EC: reproduction of **computer program***
- Article 15(1) of Directive **2019/790**: **press** publications online

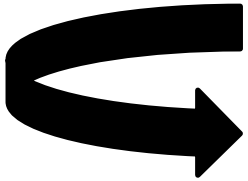
TDM in general

Not just by a research organisation or a cultural heritage institution:

but also by “the private sector” and “public entities” (recital 18)

TDM in general

Not just for the purposes of scientific research, but also “for **various purposes**, including for government services, complex business decisions and the development of new applications or technologies” (rec.18)

 for wider **commercial purposes** ?

TDM in general

It is allowed to

- Make copies
- Retain copies : “for as long as is necessary for the purposes of text and data mining”

TDM in general

There are conditions !

1. It must concern **lawfully accessible works**, “including when it has been made available to the public online” (recital 18)

⇀ Thus not for secured documents (password etc.) ?

2. “that the use of works and other subject matter **has not been expressly reserved** by their rightholders in an **appropriate manner**”

⇀ Opt out ?

TDM in general

What is appropriate?

- **publicly *available online* content**:

  by the use of machine-readable means (art. 4.3), including metadata and terms and conditions of a website or a service (recital 18).

- **In *other cases***:

  by other means, such as contractual agreements (art.7.1 a contrario) or a unilateral declaration (recital 18).

Rightholders should be able to apply measures to ensure that their reservations in this regard are respected.