twitter.com/openminted_eu

Dr. Thomas Margoni
Senior Lecturer in Intellectual Property and Internet Law
CREATe Centre – University of Glasgow

Text and Data Mining and Exceptions and Limitations to Copyright under EU law

Seoul Copyright Forum 2016

Why TDM

- **1,8 billion websites** & **3,46 billion internet users**, on 25 September 2016.
- **24 million wireless sensors and actuators** worldwide (553% up, between 2011 and 2016)
- **16 zettabytes** of useful data (**16 Trillion GB**) by 2020
- YouTube claims to upload **24 hours of video every minute**, making the site a hugely significant data aggregator.
- Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, **>500 million tweets per day** and around 200 billion tweets per year.
- Over **1,5 million** scientific publications every year
- **74,200,000 pages** existed on Facebook, with **7 million apps** and websites integrated with Facebook on 30/5/2016

From: OpenMinTeD 2016

CREATe

TDM a working definition

Text and data mining (TDM) is a powerful tool to discover value in (old and new) data

TDM: analysing structured and unstructured datasets and content to discover concepts and entities in the world, patterns they may follow and relations they engage in, and on this basis annotate, index, classify and visualise such content.

From: OpenMinTeD 2016

CREATe

Legal barriers to TDM

Copyright and rights related to copyright (e.g. Sui generis database right (SGDR))

- These rights usually restrict the reproduction (copy) and distribution of protected works and databases with substantial investment (e.g. Art 2 InfoSoc Directive and Arts. 5 and 7 Database Directive)
- Problem: reproduction is defined very broadly by EU law (any temporary or permanent copy of the whole or part of a work, etc); SGDR restricts copies of **substantial parts** and repeated copies of insubstantial parts
- Therefore any TDM (or any other act) which requires any temporary copy of the original work or DB or part thereof infringes protected works and/or SGDR

  - **Privacy/data protection**
    Protects personal data (e.g. databases containing names, addresses, age, sex, etc).
    One of the most important elements is the concept of **consent:** data subject can give consent for treatment of his/her data (e.g. in a DB). But such consent needs to be specific for a purpose. Consent cannot be given for any type of use (like e.g. copyright licences). Therefore, all data subjects may have to give their consent for every new use, something difficult to foresee in an open research environment (Open Science)
  - **PSI**
    Public Sector Information legislation is based on a different paradigm than other approaches (e.g. U.S. where works of Federal Government are not protected in the U.S.). PSI 2013 has an “open by default” approach but copyright and other similar rights and personal data are object of specific exclusion and therefore PB are under no obligation to make them accessible and/or reusable. Plus, FoA remains MS competence.
  - **Contracts/terms of use**
    Even when no rights exist on a specific BD (because there is no originality, no substantial investment, no personal data, etc) terms of use of data provider may restrict use and redistribution of DB. This limitation is based on a contractual relationship but is still an enforceable obligation (although there are differences). See ECJ in Ryanair v PR Aviation

Exceptions to legal barriers:

- **Copyright and rights related to copyright**
  - Exception and limitations to copyright (ELC), fair dealing, fair use. ELC are only partially harmonised (e.g. in EU 1 mandatory plus 20 at discretion of MS). Internationally, even more differences.
  - For TDM in EU possible exception for research and teaching. Problem: it is not uniformly implemented in all MS and it is often limited to partial copies. It is also limited to non commercial activities and only for illustration for teaching and research. Art. 5(1) is mandatory but limited in scope. Absence of general open norm (e.g. US fair use; UK fair dealing is narrower)
  - Recently, UK introduced a limitation to copyright and related rights for acts of TDM for non commercial purposes and for legally accessed sources on the basis of the EU ELC for research. In draft for a Directive for Copyright in DSM EC has introduced a mandatory TDM exception, not limited by contracts (but yes by TPM) which is only available to research organisations (contrast this with e.g. US where most TDM are considered "transformative" uses, therefore covered by fair use).

- **Privacy/data protection**
  Anonymisation of data (removal of personal data) but this is time/money consuming and may reduce the usefulness of DB
- **PSI**
  PSI legislation does not affect FoA (Freedom of Access) legislation which is MS power. But if MS empower FoA legislation then PSI "reusable by default" rule applies. However, limitation regarding copyright and personal data still applies
- **Contracts/terms of use**
  These are private agreements so there are no real exceptions. However, certain regulations (antitrust, abusive clauses, consumer protection) could under certain circumstances invalidate specific terms. This is however a case per case issue and does not seem to constitute a sound course of action.

CREATe

Licences and licence compatibility

- Licences are permissions/authorisations (contract or otherwise based) that allow one or more parties to perform certain activities.

- Licences (so called esp. in the field of copyright) may be directed to a plurality of subjects and be drafted in standard forms or had hoc

- Some licences are usually called public licences (e.g. CCPL = Creative Commons Public Licence, GPL = General Public Licence, etc).

- In certain fields Open Content Licences (e.g. CCPL, CC0, EPL, etc) are used to grant a permission to perform acts (copy, redistribute, modify, etc) in relation to a work of authorship or other subject matter (e.g. a DB), under certain conditions (Attribution, Non Derivatives, Share Alike, Non Commercial, etc).

- A possible problem is “licence proliferation”, i.e. too many (and possibly incompatible) licences. Therefore, there is a general consensus that new licences should not be created unless really necessary.

- Some projects (e.g. OpenMinTeD) promote Legal Interoperability through analysis of legal documents and compatibility matrix.

Inner limits of licences

- Licences are a powerful instrument but not perfect…
  - “Private ordering tool” i.e. can we entrust a private law tool with a function that should be a matter of public interest/intervention (wider access to knowledge)?
  - Licences are a voluntary tool, i.e. only if the owner of Work/DB is willing to grant you access, licences work. If work owner says no, there is no remedy based on contracts that can force him/her to deal with you.
  - Even if DB is willing to employ licences, very often there are problems of correct labelling (legal code, metadata, etc) of resources. This is a very serious issue faced in many projects in TDM and in science/academia.

CREATe

Policy recommendations best practices

- Through proper policy choices some of other disadvantages can be fixed.

  - Recommending 1 or a very limited no. of licences which are **compatible** (fixing problem of licence incompatibility)
  - Crucial importance that **data providers, funding agencies, scientific and public institutions** require use of correct licences and subject grants or funding to the **correct implementation** of those licences (fixing problems of “voluntarity” and “labeling”)
  - **Influence public debate** so that legislative intervention in the field is appropriate (e.g. definition of right of reproduction, harmonisation of ELC, need of a broader standard for ELC, limit of non commercial exception such as in UK).
  - Many projects in EU (e.g. OpenMinTeD) focus on OA resources given the complex legal issues (market failure?) connected with TDM.

CREATe

Example: OpenMinTeD

- The global research community generates over 1.5 million new scholarly articles per annum.

  The STM report (2009)

- ... some 90% of papers … are never cited.
  ... 50% of papers are never read by anyone other than their authors, referees and journal editors

  Lokman I. Meho, The rise and rise of citation analysis, 2007

- … one paper published every 30 seconds

Spangler et al, Automated Hypothesis Generation based on Mining Scientific Literature, 2014

From: OpenMinTeD 2016

CREATe

Example: OpenMinTeD

Machine reading

process textual sources, organise and classify in various dimensions, extract main (indexical) information items,

… and "understanding"

identify and extract entities and relations between entities, facilitate the transformation of unstructured textual sources into structured data

… and predicting

enable the multidimensional analysis of structured data to extract meaningful insights and improve the ability to predict

From: OpenMinTeD 2016

Example: OpenMinTeD

- The Open Mining Infrastructure for Text and Data (openminted.eu)
- Horizon2020 (innovation, delivering economic growth faster and delivering solutions) funding scheme (E-Infrastrucure) 2015 – 2018
- Building a registry of TDM resources, software and services to ease researchers' activity
- Focus on interoperability: technological (service and software), semantic (metadata, workflows) and legal
- WG3 works on legal interoperability focusing on rights, exceptions and licences at the content, software and service level
- Overall objective: moving Open Access towards Open Science using TDM as example

Thank you for your attention!

Gam sa hap ni da

thomas.margoni@glasgow.ac.uk

CREATe