The regulation of Text and Data Mining

Melanie Dulong de Rosnay

French National Center for Scientific Research (CNRS)

Institute for Communication Sciences

@melanieddr

Liber conference, Riga, 03/07/2014

Which regulations?
Which role for libraries?

Explain: Legal clarity for research activities

- **Advise**: Copyright chilling effect + other legislations
  - Research and education
  - Public sector information
  - Data protection
  - Sectorial domains (environment, health)

Negotiate: Terms of use

Develop

- Technical infrastructure
- Machine-readable licence
- Interface for access and reuse

Restrictive terms of use

- Elsevier
- http://libereurope.eu/blog/a-scientist-s-take-on-the-new-elsevier-tdm-policy/
- Contract override
- Unfair contracts
- Monopoly
- Copyfraud: commodification of unregulated areas

Private enclosure of data produced by the public

Twitter Data Grant

« we'll select a small number of proposals to receive free datasets »

« opportunities for the selected institutions to collaborate with Twitter engineers and researchers »

The legal framework for data

- **Data** vs databases and compilations
- *Sui generis* rights to the **producer of a database**: Europe, Mexico, Korea
  - Exclusive right against repeated and systematic extraction
  - Except non substantial use, NC, teaching & research
  - Update? +20
  - Law for unfair competition or **terms of use**
- Scientific data
  - **Copyrightable elements** text, photos, notices, observations, results and comments
  - **Metadata**, taxonomy, ontology structure: database law or terms of use
  - PSI, spacial information, environmental data
    - Additional options to restrict access

Possible reforms of the copyright directive

- Access, reuse = scanning, crossing, detecting patterns
- *Licensing for Europe* Text and data mining Working Group
  - Additional remuneration
  - Clarification of the right to perform TDM after lawful access (2001 EUCD silent)
    - Terms of use prohibiting the lawful right to perform data mining on a content accessed legitimately should be considered an abuse of exclusive rights and declared illegal
    - Technical restrictions to download the content in an open format will also be considered a barrier to data mining and their circumvention should not be considered a breach of terms of use, of copyright, or of effective rights protection measure

      (Communia international association on the public domain, EC consultation)

The problem with exceptions

_ They expire

_ A *de facto* recognition that TDM are not legitimate usages

- Should not be covered by copyright and database right
- Should be considered as an extension of the right to read

_ Non commercial?

- publicly funded research carried out by a private company, for instance a member of an FP7 project
- research conducted by a university leading to a patent application?
- project with a university and a big pharma?

Policy arguments for sharing data

- Better, quicker research
  - Reproducibility, detect errors, extract, parse and analyse data collected by others (open science)
  - Avoid duplication of funding to collect similar dataset (open access)
  - Citizen science, innovation by startups and NGO (open data)

- Archiving should be performed by libraries
  - chances to find the dataset falls by 17% every year from the third year after publication (Vines et al, 2013)
  - data related to studies of the 1990s would be permanently lost (format or affiliation change)
  - impossible to produce long-term or comparative studies (Melis et al, 2011)

Recommendations for sharing

- Legal OA
  - CC0
  - CC4.0: attribution stacking
- Technical OA
  - Restrictions to download
  - Registration
  - Absence of policy
  - 20% of 200 life science “public domain” databases (Science Commons and Dulong de Rosnay, 2008 and 2010)
  - 1% of 11000 datasets of the Global Biodiversity Information Facility with Open Data licence (Desmet, 2013)

OA mandates

- Applicable to all *vs* voluntary, imperfect effort
- No open data legislation in the world requiring authors to share their data
- What initiatives in the right direction?
- How to encourage deposit?
- Create datajournals
- Extend institutional mandates for publications to data

Open Data pilot of the European Commission Horizon 2020 (december 2013)

- Underlying data defined as the *data necessary to validate the results presented in the scientific publications, including the metadata which it should be possible to access, **mine**, exploit, reproduce and disseminate free of charge*
- Side effect of the inclusion of data mining in list of exclusive rights
- Acknowledgment that mining is not always an activity outside of the scope of copyright
- Opt-out: confidentiality or security reasons, for personal data, if there is an obligation to protect results if they can be commercially exploited, if the principal objective of the project is jeopardised, or also *for any other legitimate reason*

Legislations mandating open access

Spain

Self-archiving requirement

without prejudice of the agreements

which can have transferred to third parties

the rights on the publications (publishers)

or when the results are susceptible of protection

Peru

A central national repository

for OA to publications, data and statistics

The information should be in OA

free to read, reuse, mine and all necessary acts

but for a non-commercial purposes

and with respect to copyright law

Argentina

requires public research institutions to develop repositories

and publicly funded research to be made available in OA repositories

within 6 month after publication for the article

and 5 years after collection for the primary data

so that other researchers might reuse them

exceptions in the case of intellectual property

prior agreements with third parties

confidentiality

Germany

mandate of self-archiving

for non-commercial purposes

of the author's final version of articles

published in journals issued at least twice a year

and funded for at least 50% publicly

declares contradictory publishers' agreements void

good, but may apply to national publishers only

Italy

18 months after first publication

for scientific, technical, and medical disciplines

24 months for humanities and social sciences

longer than acceptable recommendations

by the Open Access scientific community

Crucial elements

Definition of the scope of the research results covered

What is Open Access

Pre-existing copyright agreements

Risk of reidentification by big data / confidentiality

Implementation means and technical support

Questions?

Thanks!

melanie.dulong@cnrs.fr