University of Amsterdam

Institute for Information Law

Text & Data Mining: Barriers, Paths and Passable Roads

L. Guibault

Nexa Center for Internet and Society **27.06.2016**

Harvard Business Review

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | COMMENT 6 | TEXT SIZE | PRINT | $8.95 BUY COPIES

WHAT TO READ NEXT

What Every Manager Should Know About Machine Learning

What Separates a Good Data Scientist from a Great One

How to Start Thinking Like a Data Scientist

TDM today and tomorrow...

Using Big Data Analytics in all areas of life

Learning

Journalism

Science

eCommerce

IoT

Forces influencing the uptake of TDM

IPRs

Data Protection

Open Access Policies

Education and Skills

Compatibility of Licenses

Technical Interoperability

Data Management

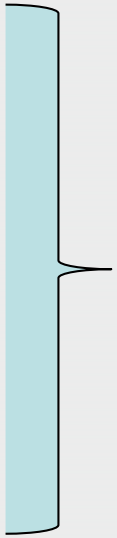
Intellectual Property Rights

THE PROBLEM?

TDM involves access to and usage of content (articles, sounds, images and data) in bulk which may infringe IP rights

TDM in terms of IPRs

- Crawling and scraping
- Create target datasets
- Analyze
- Evaluate
- Publish results



Acts of reproduction and communication to the public

Copyright

- Scope of protection:
  - Broad rights of reproduction & making available
  - Question: is every technical reproduction also a reproduction in the sense of copyright law?

Relevant copyright exceptions

- Transient and incidental act of reproduction
- Private copying
- Reproductions made by libraries & archives
- Educational use and research exception

Sui generis Database right

- Subject matter of protection

‘database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database’.

Sui generis Database right

- Right of extraction

  The permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form

- Right of re-utilization

  Any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission.

- Repeated and systematic extraction and/or re-utilization of insubstantial parts of the contents

Sui generis Database right

- Case C202-12, Decision 19 December 2013 (Innoweb v. Wegener)

  Dedicated meta search engine infringes database right since it constitutes “re-utilising” of the “whole or a substantial part” of a database.

Sui generis Database right

- Educational use and research exception (optional)

  (b) in the case of **extraction** for the purposes of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved;

Issues that remain unclear

- What about cross-border TDM activities?
- How much adaptation to the works or the database remains clear from infringement?
- What about snippets?

Summary of legal barriers

Restrictiveness:
Strict rules

Fragmentation:
Different rules and interpretations

Uncertainty:
Unclear rules

Copyright reform – where are we now?

The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014

"29A Copies for text and data analysis for non-commercial research

(1) The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that—

(a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and

(b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise).

(2) Where a copy of a work has been made under this section, copyright in the work is infringed if—

(a) the copy is transferred to any other person, except where the transfer is authorised by the copyright owner, or

(b) the copy is used for any purpose other than that mentioned in subsection (1)(a), except where the use is authorised by the copyright owner.

(3) If a copy made under this section is subsequently dealt with—

(a) it is to be treated as an infringing copy for the purposes of that dealing, and

(b) if that dealing infringes copyright, it is to be treated as an infringing copy for all subsequent purposes.

(4) In subsection (3) "dealt with" means sold or let for hire, or offered or exposed for sale or hire.

(5) To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable.".

UK exception

- For ‘research purposes’
- Limited to *non-commercial* research
- Prevails over contractual arrangement to the contrary
- Does not address the relationship with technological protection measures

At European Union level?

48. Stresses the need to properly assess the enablement of enable automated analytical techniques for text and data (e.g. '**text and data mining**' or 'content mining') for research purposes, provided that permission to read the work has been acquired;

Julia Reda, EU Copyright Evaluation Report (European Parliament, May 2015)

Towards a modern, more European copyright framework (EC Communication 9 December 2015)

The Commission is assessing options and will consider legislative proposals on other EU exceptions by spring 2016, in order to:

- allow public interest research organisations to carry out **text and data mining** of content they have lawful access to, with full legal certainty, for scientific research purposes;

Data protection

General Data Protection Regulation (2016/679)

- Definition of personal data & identifiable natural person
- Processing: lawful *and* justified

Stimulating policies

- Open Access Policies:
  - H2020, OpenAire, ERC,
  - National funding agencies
- European Cloud Initiatives:
  - 3O's (Open Access, Open Data, Open to the World)

Open Access (OA)

Unrestricted beneficiaries

No use restriction

No purpose restriction, unless non-commercial license

OA to publications

Required by research funders

OA copyright provisions

Increasing OA publishers

OA to data

More challenges:

- Confidential data
- Privacy and data protection
- Competition

Policy-wise in its infancy compared to OA to publications

Conclusion

- Imperative to make room for TDM in a broad manner
- EU competitiveness is at stake
- Good timing to review IP laws

Thank you for your attention!

For more information

L.Guibault@uva.nl