TEXT AND DATA MINING IN PUBLIC RESEARCH

Julien Roche – University of Lille, ADBU
LIBER 2017 – Patras, Greece
1. Why does TDM matter?

2. Why isn’t it used more widely in public research?

3. How do we change this?
Study aims

Assess economic impact of TDM on public research in France via:

- Case studies (France, UK, Europe)
- Analysis of the relevance of a copyright exception for TDM – French context
- A final report for decision makers and lobbying

http://adbu.fr/etude-tdm/
€16 billion
Value of R&D performed within French universities and public research bodies (source: Eurostat)

6-fold return
€6 contribution to EU economy for each €1 directly generated by research universities (source: Biggar Economics)

20% per annum
Estimated rate of return to public investment in science and innovation (source: Frontier Economics)
2.5 quintillion bytes
Data produced each day

2.4 million
Scientific articles per annum

Zero
Number of researchers who can keep up
What is TDM?

Any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations.


=> TDM is machine reading, i.e. just another way to read
BASE CAMP

Where are we now, and how did we get here?
What is the problem?

...countries, in which academic researchers must acquire the express consent of rights holders to conduct lawful datamining, exhibit a significantly lower share of data mining research output relative to total research output

Handke, Guilbault and Vallbe IS EUROPE FALLING BEHIND IN DATA MINING? (2015)
What is the result?

The European ecosystem for engaging in text and data mining remains highly problematic... The end result: Europe is being leapfrogged by rising interest in other regions, notably Asia.

Filippov, S. & Hofheinz, P. Text and Data Mining for Research and Innovation (2016)
Legislative options

Loi pour une République Numérique (Loi LEMAIRE)  
N° 2016-1321 - 7 October 2016

1. Industry self-regulation
   - Non-commercial research only

2. 2014
   - Commercial research purpose, beneficiaries restricted

3. 2017-2018
   - Commercial research purpose, beneficiaries unrestricted

Mandatory exceptions to copyright
<table>
<thead>
<tr>
<th>Restriction</th>
<th>France</th>
<th>UK</th>
<th>EU</th>
</tr>
</thead>
<tbody>
<tr>
<td>No lawful access</td>
<td>🚫</td>
<td>🚫</td>
<td>🚫</td>
</tr>
<tr>
<td>Not scientific literature</td>
<td>🚫</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>Not public research</td>
<td>🚫</td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>Commercial purpose</td>
<td>🚫</td>
<td>🚫</td>
<td>🟢</td>
</tr>
<tr>
<td>Conservation not by designated body</td>
<td>🚫</td>
<td>🟢</td>
<td>🟢</td>
</tr>
</tbody>
</table>
The French context – Update, July 2017

- Bringing French and European frameworks into line;
- A MOU – best practices, waiting for the new EC copyright directive;
- A French decree, when the European framework is set.
In December 2016, the FutureTDM project released a policy framework document outlining the needs for a successful implementation of TDM.

Their work identifies a series of barriers that need to be overcome opposed to high-level principles that should be followed to address them.

Source: http://openminted.eu/future-tdms-policy-recommendations/
Barriers to TDM < FutureTDM

• The **barriers** that need to be overcome are as follows:

  • **Uncertainty**: this category includes uncertainties as to how, why and if TDM can be carried out, as well as the lack of awareness of different aspects of TDM.

  • **Fragmentation**: this refers to the fragmentation in the TDM landscape, which prevents TDM from being carried out across e.g. national borders, scientific domains, companies or fields of expertise.

  • **Restrictiveness**: the last category refers to direct limitations to the ability to carry out TDM, in the form of restrictive laws, lack of expertise, limited (financial) resources, etc.

The future of TDM < FutureTDM

- The **high-level principles** identified to overcome the barriers are:

  - **Awareness and Clarity**: Information and clear actions are crucial for a flourishing TDM environment in Europe.
  
  - **TDM without Boundaries**: boundaries should be broken down to reduce fragmentation in the TDM landscape.
  
  - **Equitable Access**: access to TDM tools, technologies and sources should take into account the need of both users and providers.

<table>
<thead>
<tr>
<th>BARRIERS</th>
<th>PRINCIPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertainty</td>
<td>Awareness and Clarity</td>
</tr>
<tr>
<td>Fragmentation</td>
<td>TDM without Boundaries</td>
</tr>
<tr>
<td>Restrictiveness</td>
<td>Equitable Access</td>
</tr>
</tbody>
</table>

Source: http://openminted.eu/future-tdms-policy-recommendations/
The long road to TDM

Summit: Researchers embrace TDM

Camp 1: Legal clarity
- EU Directive
- Copyright exception (Base Camp)

Camp 2: Access to content
- Camp 3: Technical infrastructure
- Camp 4: Skills and support

New route
Old route
Source: national geographic
1. ACHIEVING LEGAL CLARITY
The exception has made a massive difference...

Petr Knoth, Open University, UK
...the definition of commercial and non-commercial research is creating uncertainty

Petr Knoth, Open University, UK
What needs to happen?

• Communicate legal provisions for TDM with certainty and clarity
• Clarify the exception’s scope where public researchers collaborate with commercial partners
• Monitor the interaction of the copyright exception with digital rights management (DRM), licensing and other relevant legal regimes
2. SECURING ACCESS

China Halts All Climbing on Tibetan Side of Mt. Everest

Destinations - Asia
Martha Bellisle, Associated Press - Apr 30, 2015 3:30 am
I scaled down my TDM research, and had to exclude two publishers... I couldn’t do what I set out to do

Chris Hartgerink, Tilburg University, Netherlands
I had to ask too many publishers for the right to download ... it takes a lot of time and ... the publishers’ servers frequently block us.

Mathieu Andro, INRA, France
What is the problem with access?

- Technical protection measures (TPMs)
- Crawler traps
- Restricted access to application programming interfaces (APIs)
What needs to happen?

• Incorporate TDM clauses into model licence agreements
• Educate researchers on their rights
• Maintain dialogue with publishers
• Improve access through better infrastructure...
3. INFRASTRUCTURE & TOOLS

Image: National Geographic
...Every time you have a new project or data source... you hit issues about how the documents are structured, oddities of formatting, and so on.

Mark Greenwood, GATE, UK
The TDM Landscape

Text Mining Researchers

Computing Infrastructures

Content Providers

End Users

Source: OpenMinTED
What needs to happen?

- Invest in TDM infrastructure
- Make TDM accessible to non-specialists
- Streamline access
- Open standards and harmonised data formats
4. SKILLS & SUPPORT
...We have algorithms to answer questions, but we do not have algorithms to ask questions

François Rioult, GREYC Laboratory, Université de Caen, France
What is the role of the librarian?
The library needs to be able to say: ‘If you’ve got a question about TDM, come to us’

Danny Kingsley, Head of Scholarly Communications, University of Cambridge, UK
Library support for TDM

- Advocacy
- Copyright advice
- Access to legal expertise
- Skills development and training
- Advice on data sources and tools
Small steps, big progress

Ex. “WillO” tool of the university of Lille
https://decadoc.typeform.com/to/W2ZZMV

=> a tool for French researchers to know their rights and their obligations when they want to put their publications in Open Access, in the legal framework of the French law
5. EMBRACING TDM
"Because it's there"

Edmund Percival Hillary
There are so many obstructions in the way of doing this research, and doing it well. It is just too hard and so people do other things

Ross Mounce, University of Cambridge, UK
What needs to happen?

• Endorsement by senior research leaders
• Funding and incentives linked to TDM
• Alignment with moves to open science
1. Why does TDM matter?
2. Why isn’t it used more widely in public research?
3. How do we change this?
Why does TDM matter?

Public research is valuable

TDM makes research more efficient

TDM is worth investing in
1. Why does TDM matter?

2. Why isn’t it used more widely in public research?

3. How do we change this?
Summit: Researchers embrace TDM

Camp 1: Legal clarity
- EU Directive

Camp 2: Access to content

Camp 3: Technical infrastructure

Camp 4: Skills and support

Copyright exception (Base Camp)
1. Why does TDM matter?

2. Why isn’t it used more widely in public research?

3. How do we change this?
Making TDM a reality

Libraries

- Monitor researchers’ experience
- Develop case studies and guidance
- Involve the national libraries
- Invest in TDM support
- Incorporate TDM clauses into licence agreements
Full case studies from the TDM report

Legislators

- Provide certainty
- Enable public/private partnerships
- Monitor interaction with other legislation (e.g. DRM)

Institutions/research leaders

- Endorse TDM
- Invest in library services
- Explore knowledge exchange opportunities

Making TDM a reality

Research funders

- Invest in infrastructure
- Forum to improve access
- Link TDM to Open Science

Publishers & providers

- Cloud services for TDM
- Streamline access
- Open, harmonised standards
Thank you

J. Roche – R. Johnson

Full report available at:
http://adbu.fr/etude-tdm/

julien.roche@univ-lille1.Fr
rob.johnson@research-consulting.com

https://adbu.fr
www.research-consulting.com