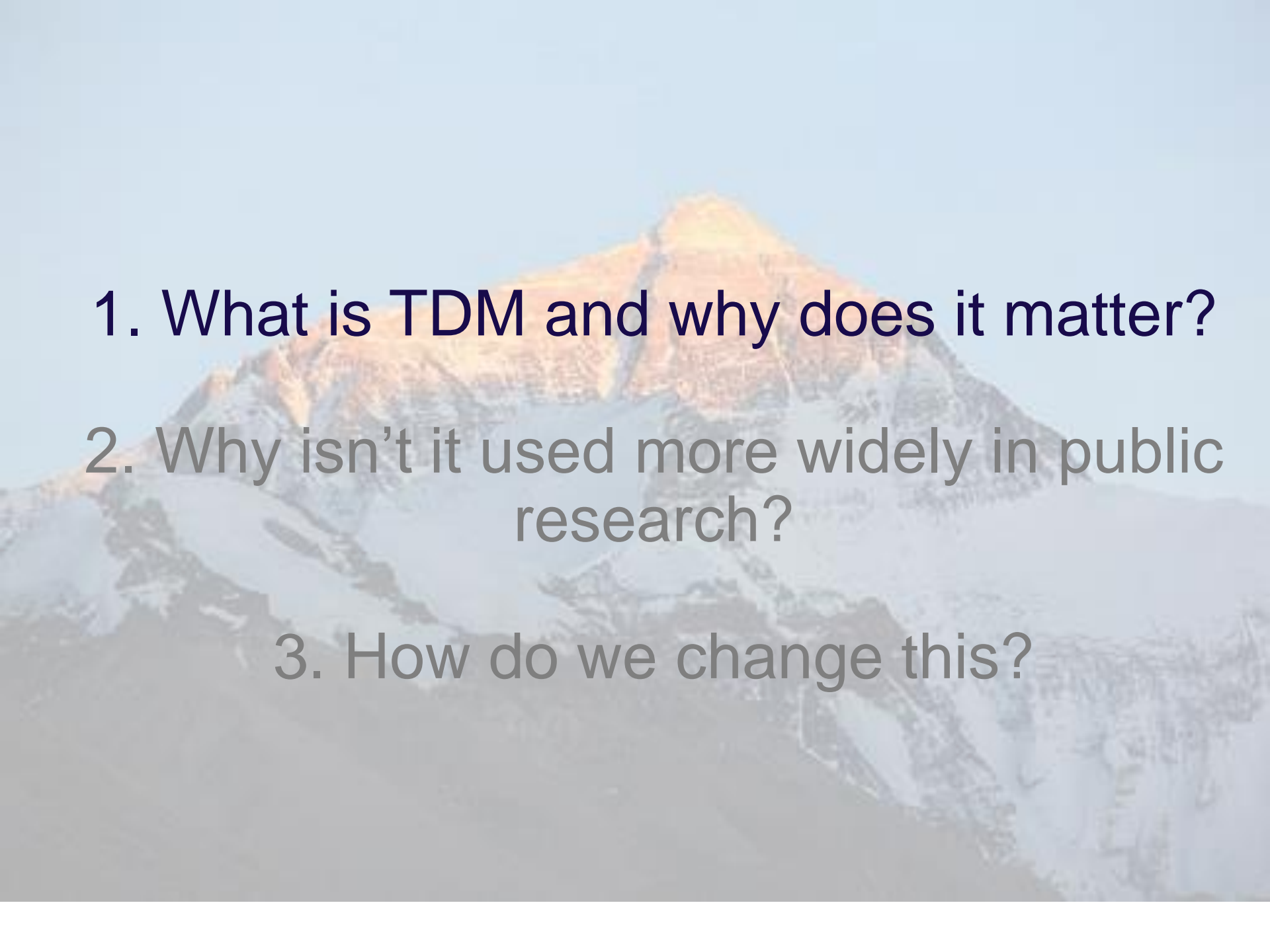


Text and Data Mining in the  
context of the reform of the EU  
directive regarding Copyright

-

Fiesole Retreat  
Barcelona  
April 2018

- 
1. What is TDM and why does it matter?
  2. Why isn't it used more widely in public research?
  3. How do we change this?

# 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.*

European Commission. [Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market](#)

=> This is nothing new, we already do that for years...

The screenshot shows the LILLIAD website interface. At the top, there is a navigation bar with the LILLIAD logo and menu items: DÉCOUVERTE, Disciplines, Services, and Vivre l'innovation. A search bar is located on the right side of the navigation bar. Below the navigation bar, there is a secondary menu with options: NOUVELLE RECHERCHE, LISTE DE REVUES @, and AIDE. The main content area features a large search bar with the text "Rechercher" and "Rechercher sur le site". To the right of the search bar, there is a button labeled "RECHERCHE AVANCÉE". Below the search bar, there are two main sections:

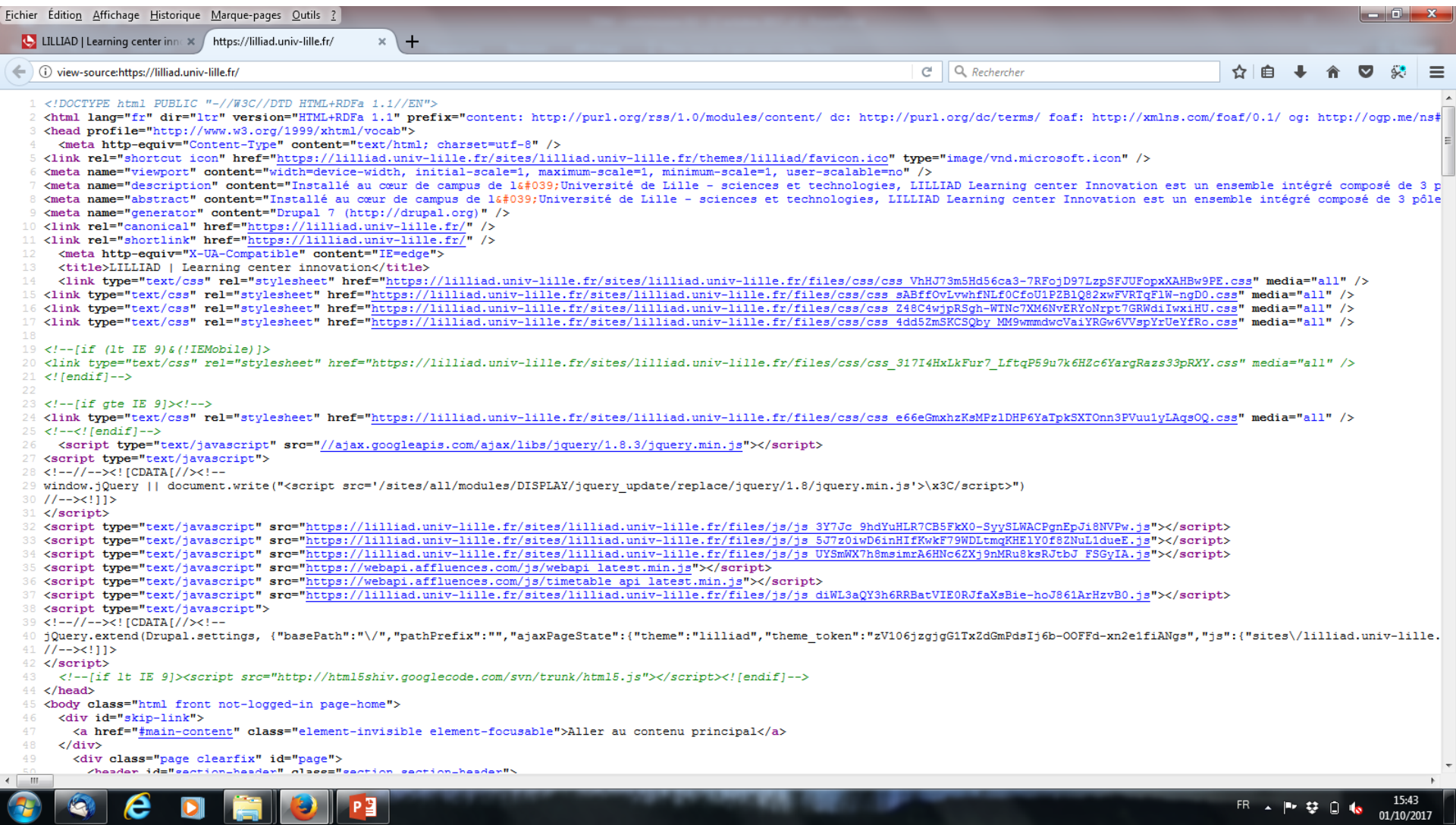
- Le moteur de recherche documentaire**: This section describes the search engine's capabilities, stating it provides access to the majority of collections from LILLIAD libraries, including books, journals, articles, databases, and theses. It also mentions that certain resources are not referenced in the search engine.
- Vous ne trouvez pas ce que vous cherchez ?**: This section offers three options for users who cannot find what they are looking for: "Faites venir un document d'une autre bibliothèque", "Suggérez un achat d'ouvrage", and "Signalez un problème technique".

Below these sections, there is a large banner with the text "Nouvelle année, nouvelle interface !". Underneath this banner, there are three promotional cards:

- 26/09/2017**: Nouvelle année, nouvelle interface ! EN SAVOIR PLUS
- 25/09/2017**: Donnez votre avis sur la signalétique LILLIAD EN SAVOIR PLUS
- 18/09/2017**: Fête de la Science : visitez Xperium EN SAVOIR PLUS

The bottom of the screenshot shows the Windows taskbar with various application icons and the system tray displaying the date and time: 15:41, 01/10/2017.

What we « see »....

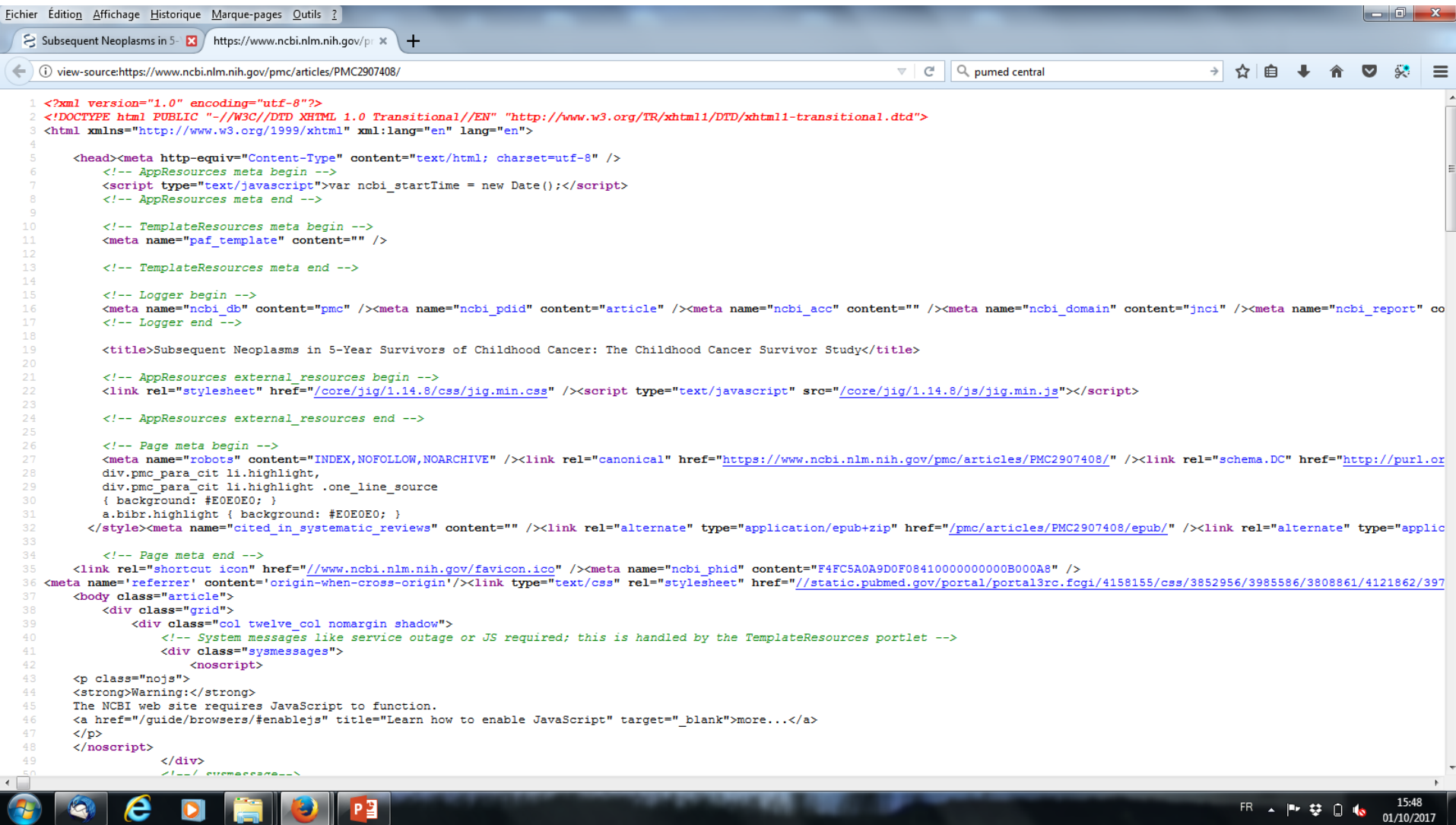


... and the reality.

The screenshot shows a web browser window with the following elements:

- Browser Tabs:** LILLIAD | Learning center inn... and Subsequent Neoplasms in 5-...
- Address Bar:** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2907408/>
- Page Header:** NCBI Resources How To Sign in to NCBI
- PMC Logo:** US National Library of Medicine National Institutes of Health
- Search Bar:** PMC [Search]
- Journal List:** J Natl Cancer Inst > PMC2907408
- Article Banner:** OXFORD JOURNALS JNCI Journal of the National Cancer Institute
- Article Info:** J Natl Cancer Inst. 2010 Jul 21; 102(14): 1083-1095. PMID: PMC2907408. doi: 10.1093/nci/djq238
- Title:** Subsequent Neoplasms in 5-Year Survivors of Childhood Cancer: The Childhood Cancer Survivor Study
- Authors:** Debra L. Friedman, John Whitton, Wendy Leisenring, Ann C. Mertens, Sue Hammond, Marilyn Stovall, Sarah S. Donaldson, Anna T. Meadows, Leslie L. Robison, and Joseph P. Neglia
- Links:** Author information, Article notes, Copyright and License information
- Yellow Box:** This article has been cited by other articles in PMC.
- Abstract Section:** **Abstract** (with a 'Go to:' link), **Background**, **Methods**
- Abstract Text:** The occurrence of subsequent neoplasms has direct impact on the quantity and quality of life in cancer survivors. We have expanded our analysis of these events in the Childhood Cancer Survivor Study (CCSS) to better understand the occurrence of these events as the survivor population ages.
- Methods Text:** The incidence of and risk for subsequent neoplasms occurring 5 years or more after the childhood cancer diagnosis were determined among 14 359 5-year survivors in the CCSS who were treated from 1970 through 1986 and who were at a median age of 30 years (range = 5-56 years) for this analysis. At 30 years after childhood cancer diagnosis, we calculated cumulative incidence at 30 years of subsequent neoplasms
- Formats:** Article | PubReader | ePub (beta) | PDF (621K) | Citation
- Share:** Facebook, Twitter, Google+
- Save items:** Add to Favorites
- Similar articles in PubMed:** Secondary sarcomas in childhood cancer survivors: a report from the Childhood Cancer Survivor Study. [J Natl Cancer Inst. 2007]; New primary neoplasms of the central nervous system in survivors of childhood cancer: a report from the C [J Natl Cancer Inst. 2006]; Second neoplasms in survivors of childhood cancer: findings from the Childhood Cancer Survivor Study cohort. [J Clin Oncol. 2009]; Risk of selected subsequent carcinomas in survivors of childhood cancer: a report from the Childhood Cancer St [J Clin Oncol. 2006]; Late mortality among 5-year survivors of childhood cancer: a summary from the Childhood Cancer Survivor [J Clin Oncol. 2009]
- Cited by other articles in PMC:** Somatic and germline TP53 alterations in second malignant neoplasms from pediatric cancer [Clinical cancer research : an ...]; Therapeutic radiation for childhood cancer drives structural

What we « see »....



... and the reality.

⇒ TDM is only machine reading,  
i.e. just another way to read

Like we did for years browsing the  
Internet for content



# What does the research community expect?

---

*A copyright reform at the EU to perform TDM, because TDM will / is leading to major scientific innovations*

-

*Writing -> scriptorium -> printing -> peer-reviewing -> ... machine reading.*

# What is the situation?

*Member States pay each year hundreds of millions euros for academics to have the right to read the scientific literature we the academics published.*

*Whatever might be the evolution of the economical model for scientific publication, Member States will keep paying for scientific literature (traditional subscription models, APCs, gold, green AO)*

*=> We already pay – a lot – to read scientific literature.*

Europe has tremendous  
ambitions regarding innovation...

Two horizontal bars are positioned below the first text block. The left bar is split into a dark blue segment on the left and a red segment on the right. The right bar is split into a teal segment on the left and a light green segment on the right.

...and public research is a  
powerful player.

Why is HER concerned about the  
Copyright reform?

-

because the future will be heavily  
innovation driven...

...and Text and Data Mining is a powerful way to achieve disruptive innovation

# The scope of TDM...

---

- Artificial Intelligence,
- Biology,
- Medicine,
- Political sciences,
- Economics,
- History,
- Linguistics...

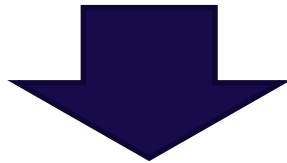
# Why does TDM matter?



Public research is valuable



TDM makes research more efficient



TDM is worth investing in



**2.5 quintillion bytes**

Data produced each day



**2.4 million**

Scientific articles per annum



**Zero**

Number of researchers who can keep up



# TDM 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*

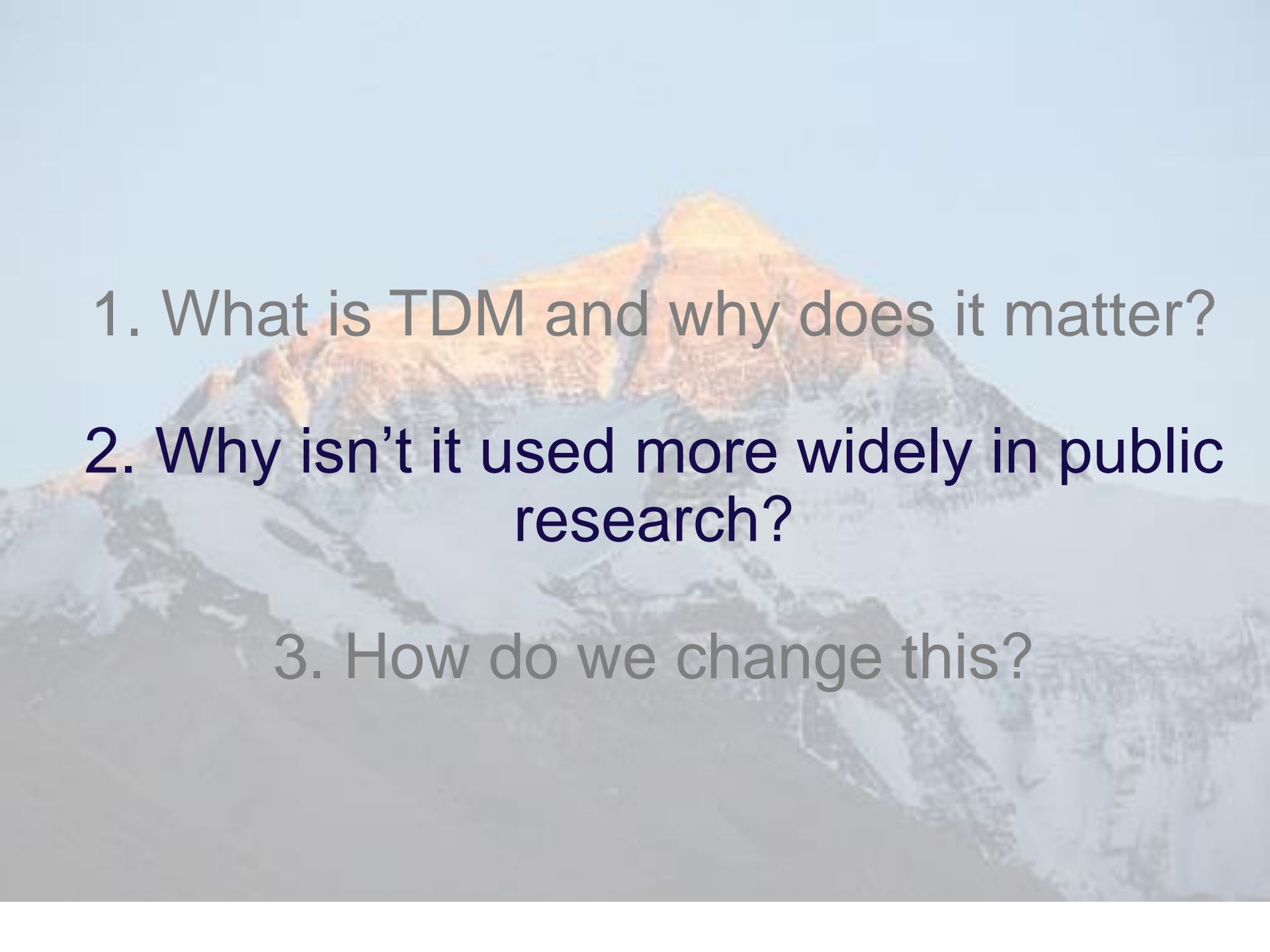
Handke, Guilbault and Vallbe [IS EUROPE FALLING BEHIND IN DATA MINING?](#) (2015)

# 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)

- 
1. What is TDM and why does it matter?
  2. Why isn't it used more widely in public research?
  3. How do we change this?

# Barriers to TDM < *FutureTDM*


- 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.

# Barriers to TDM < *Future TDM*

- **Three 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:

BARRIERS		PRINCIPLES
Uncertainty		Awareness and Clarity
Fragmentation		TDM without Boundaries
Restrictiveness		Equitable Access

- **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.

Summit: **Researchers embrace TDM**

Camp 4: **Skills and support**

Camp 3: **Technical infrastructure**

Camp 2: **Access to content**

Camp 1: **Legal clarity**

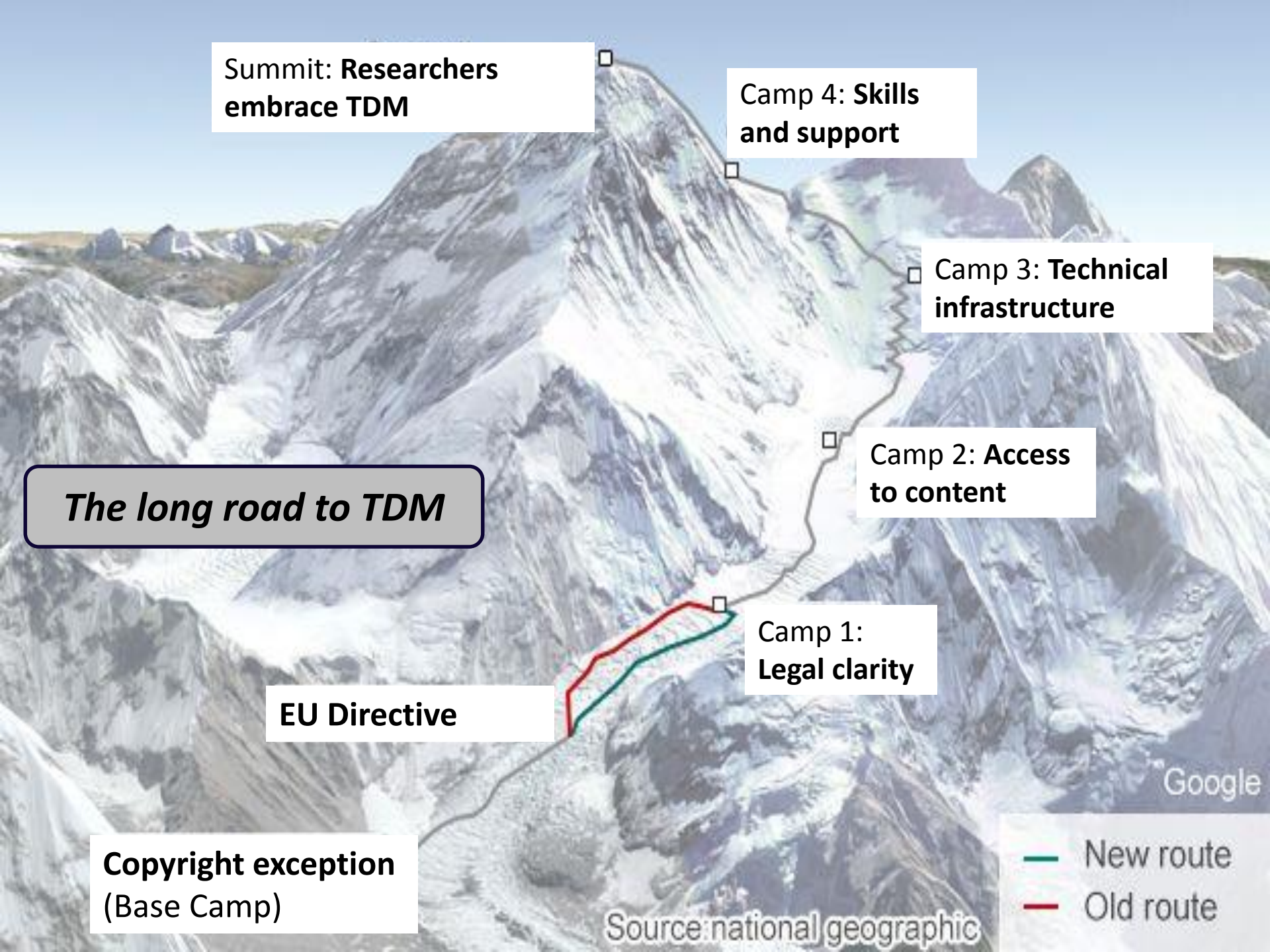
*The long road to TDM*

**EU Directive**

**Copyright exception (Base Camp)**

— 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



*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.

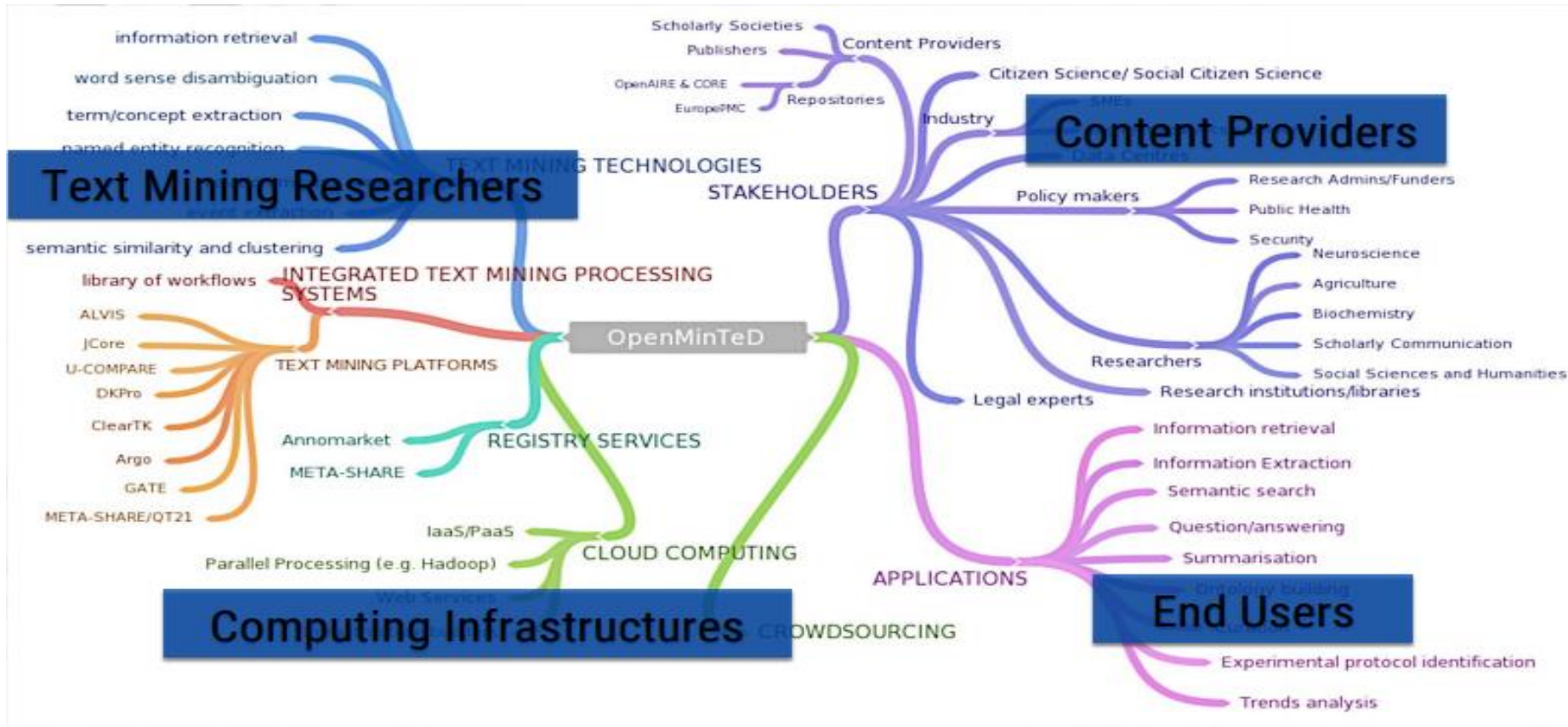
# INFRASTRUCTURE & TOOLS



*...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



# 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?



Photo: REUTERS





*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
- All kind of TDM services



# 5. EMBRACING TDM

Why?

“

*"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?**

# What do we expect?



A simple, clear, full and European wide exception for  
TDM for Research  
to be able to mine content lawfully accessible

Researchers believe that “the right to read is the right  
to mine”



# Making TDM a reality

## Libraries

- Help decision makers in their own institution
- Monitor researchers' experience
- Develop case studies and guidance
- Involve the national libraries and other stakeholders
- Invest massively in TDM support
- Incorporate extra TDM clauses into licence agreements, in the context of a large, simple and strong exception at the EU level



# *Making TDM a reality*



## **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



## **Research funders**

- Invest in infrastructure
- Create fora to improve access and sharing
- Link TDM to Open Science



## **Publishers & providers (private and public sector)**

- Develop cloud services for TDM
- Give streamline access
- Provide open, harmonised standards

# So let's summarize the core needs

- Communicate legal provisions for TDM with certainty and clarity
- Clarify the exception's scope where public researchers collaborate with commercial partners
- Let the researchers do their research with as little impedimenta as possible
- Monitor the interaction of the copyright exception with digital rights management (DRM), licensing and other relevant legal regimes



# Full case studies from the TDM report

<http://adbu.fr/competplug/uploads/2016/12/Annex-1-Full-case-studies-Final-11-Dec-16.pdf>

# Thank you

Full TDM report – in English - available at :

<http://adbu.fr/etude-tdm/>



Research consulting :

<https://www.research-consulting.com/reports/>



LIBER and TDM :

<http://libereurope.eu/text-data-mining/>



Ligue des Bibliothèques Européennes de Recherche  
Association of European Research Libraries

*LIBER is Europe's largest network of research libraries,  
with over 400 members.*



[julien.roche@univ-lille.fr](mailto:julien.roche@univ-lille.fr)

