

TEXT AND DATA MINING, PUBLIC RESEARCH AND LIBRARIES

Who I am

-

- I used for years TDM for my research without knowing it was “TDM”
- I’m the director of a university library network delivering services to support research – and TDM support is in the portfolio,
- I managed a study regarding TDM in Higher Education and Research in France and UK – end of 2016.

A study regarding TDM

Assess the impact of TDM on public research (in France) via:

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

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

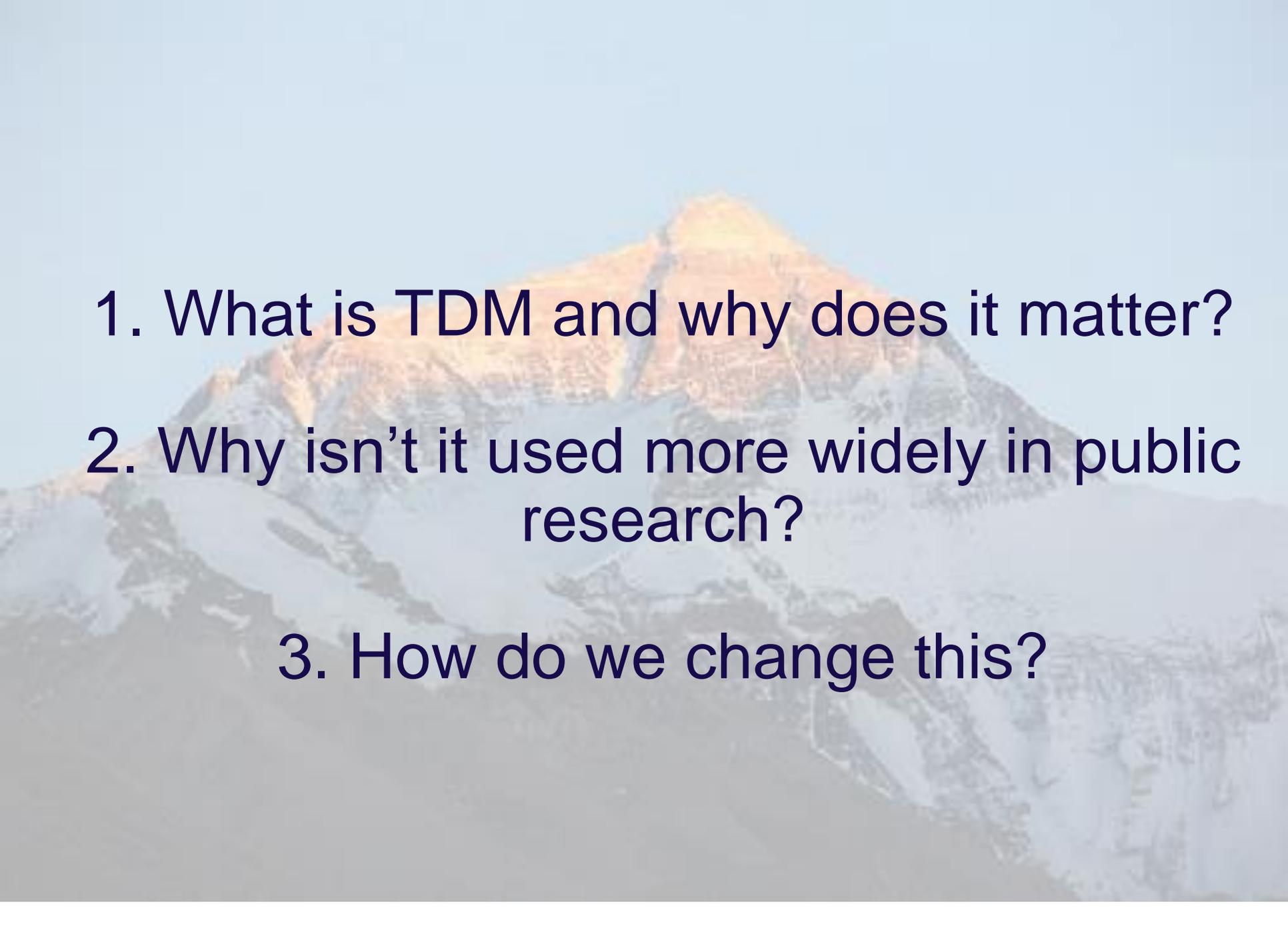
Text and data mining in higher education and public research

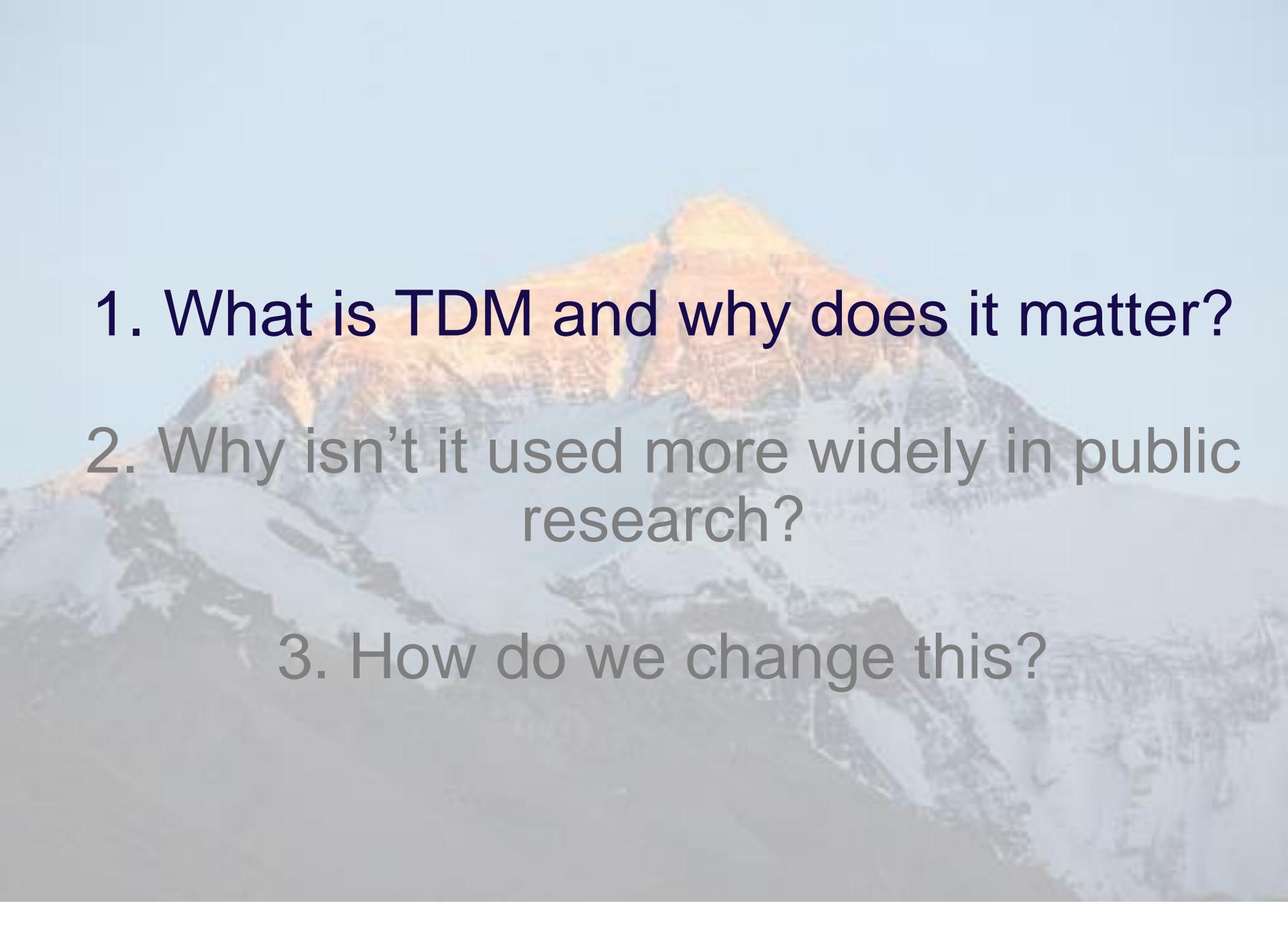


An analysis of case studies from the United Kingdom and France

PREPARED ON BEHALF OF THE ADBU

December 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?

- 
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...

Fichier Édition Affichage Historique Marque-pages Outils ?

LILLIAD | Learning center inn x +

https://lilliad.univ-lille.fr

Rechercher

LILLIAD LEARNING CENTER INNOVATION

DÉCOUVERTE ▼ Disciplines Services Vivre l'innovation

PUBLIC ▼

Université de Lille

EN / NL

NOUVELLE RECHERCHE LISTE DE REVUES @ AIDE

Rechercher sur le site

RECHERCHE AVANCÉE

Le moteur de recherche documentaire

Le moteur de recherche documentaire donne accès à la majeure partie des collections des bibliothèques de LILLIAD : livres, revues, articles, bases de données, thèses...

Les ressources suivantes ne sont pas référencées dans le moteur de recherche:

Vous ne trouvez pas ce que vous cherchez ?

- Faites venir un document d'une autre bibliothèque
- Suggérez un achat d'ouvrage
- Signalez un problème technique

Nouvelle année, nouvelle interface !

26/09/2017

Nouvelle année, nouvelle interface !

EN SAVOIR PLUS

25/09/2017

Donnez votre avis sur la signalétique LILLIAD

EN SAVOIR PLUS

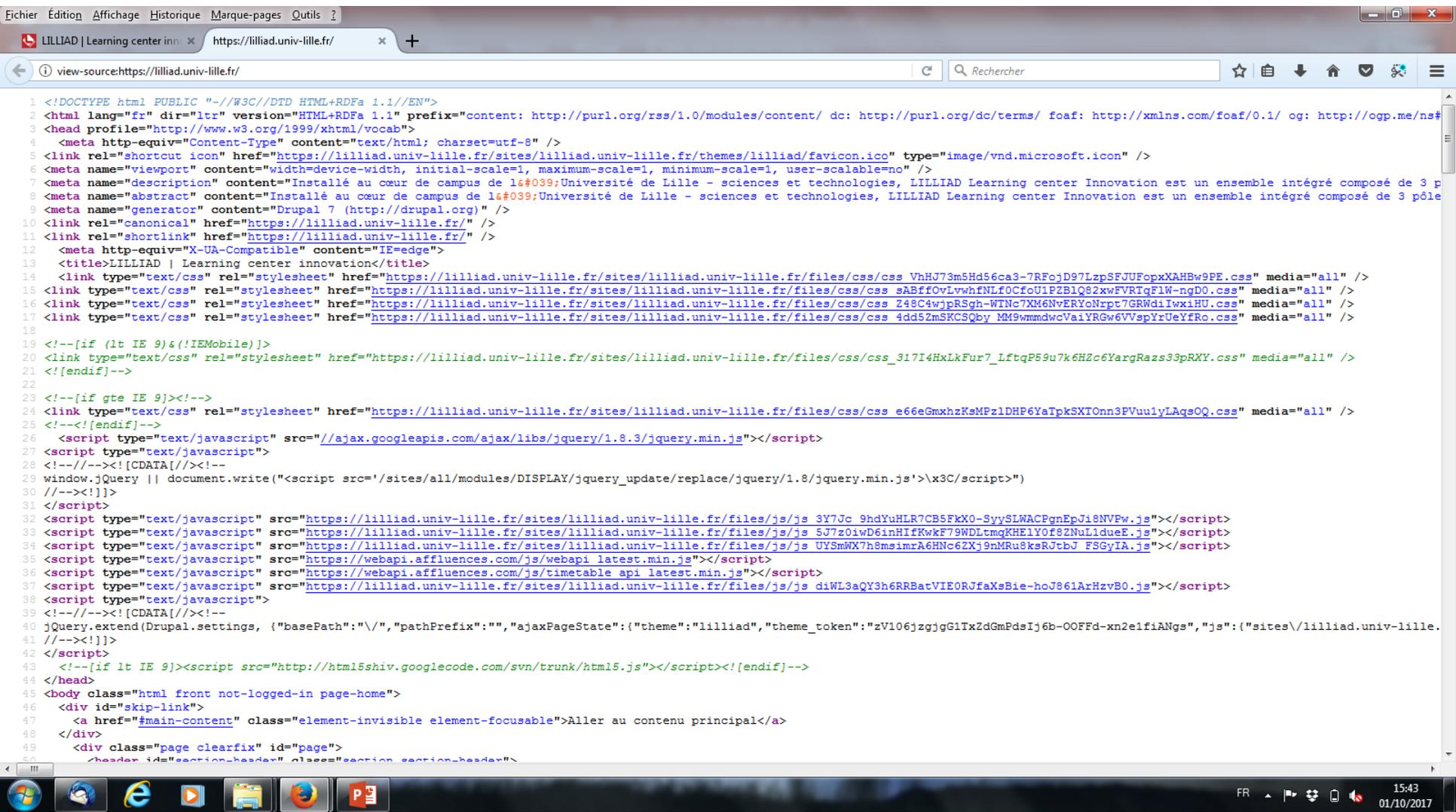
18/09/2017

Fête de la Science : visitez Xperium

EN SAVOIR PLUS

FR 15:41 01/10/2017

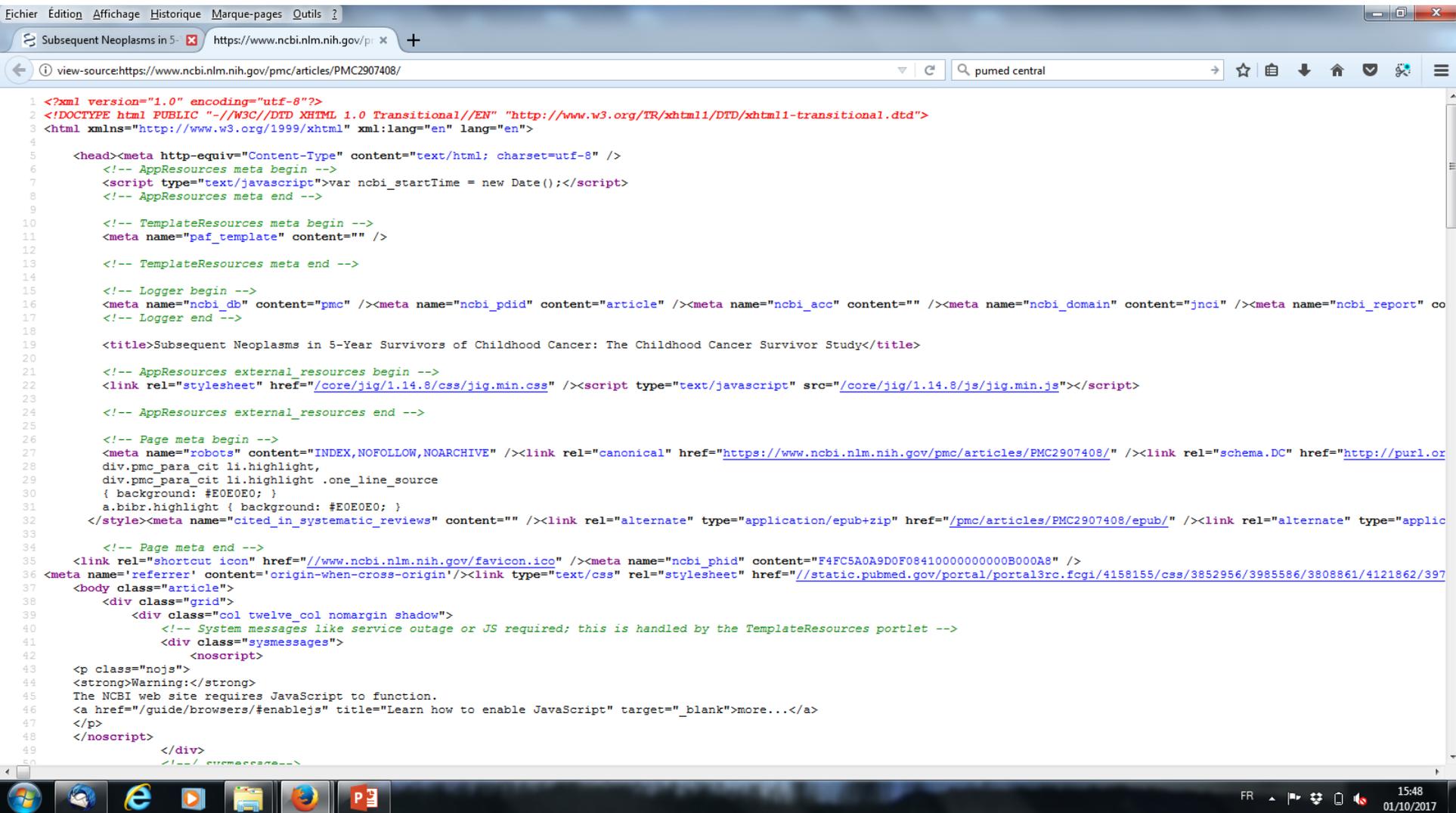
What we « see »....



... and the reality.

The screenshot shows a web browser window with the address bar displaying the URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2907408/>. The page is from the National Center for Biotechnology Information (NCBI) and features the PubMed Central (PMC) logo. The article title is "Subsequent Neoplasms in 5-Year Survivors of Childhood Cancer: The Childhood Cancer Survivor Study" by 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. The article is published in the JNCI Journal of the National Cancer Institute, 2010, 102(14): 1083-1095. The abstract section is visible, starting with "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." The right sidebar contains sections for "Formats" (Article, PubReader, ePub, PDF, Citation), "Share" (Facebook, Twitter, Google+), "Save items" (Add to Favorites), "Similar articles in PubMed" (listing related studies), and "Cited by other articles in PMC" (listing studies that cite this article).

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.

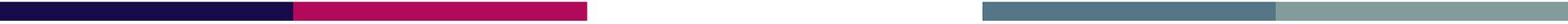
What is the situation?

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

=> 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.

The scope of TDM...

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

Why does TDM matter?



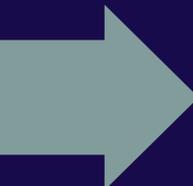
Public research is valuable



TDM makes research more efficient

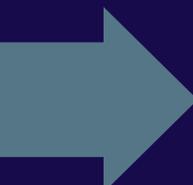


TDM is worth investing in



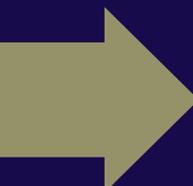
2.5 quintillion bytes

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

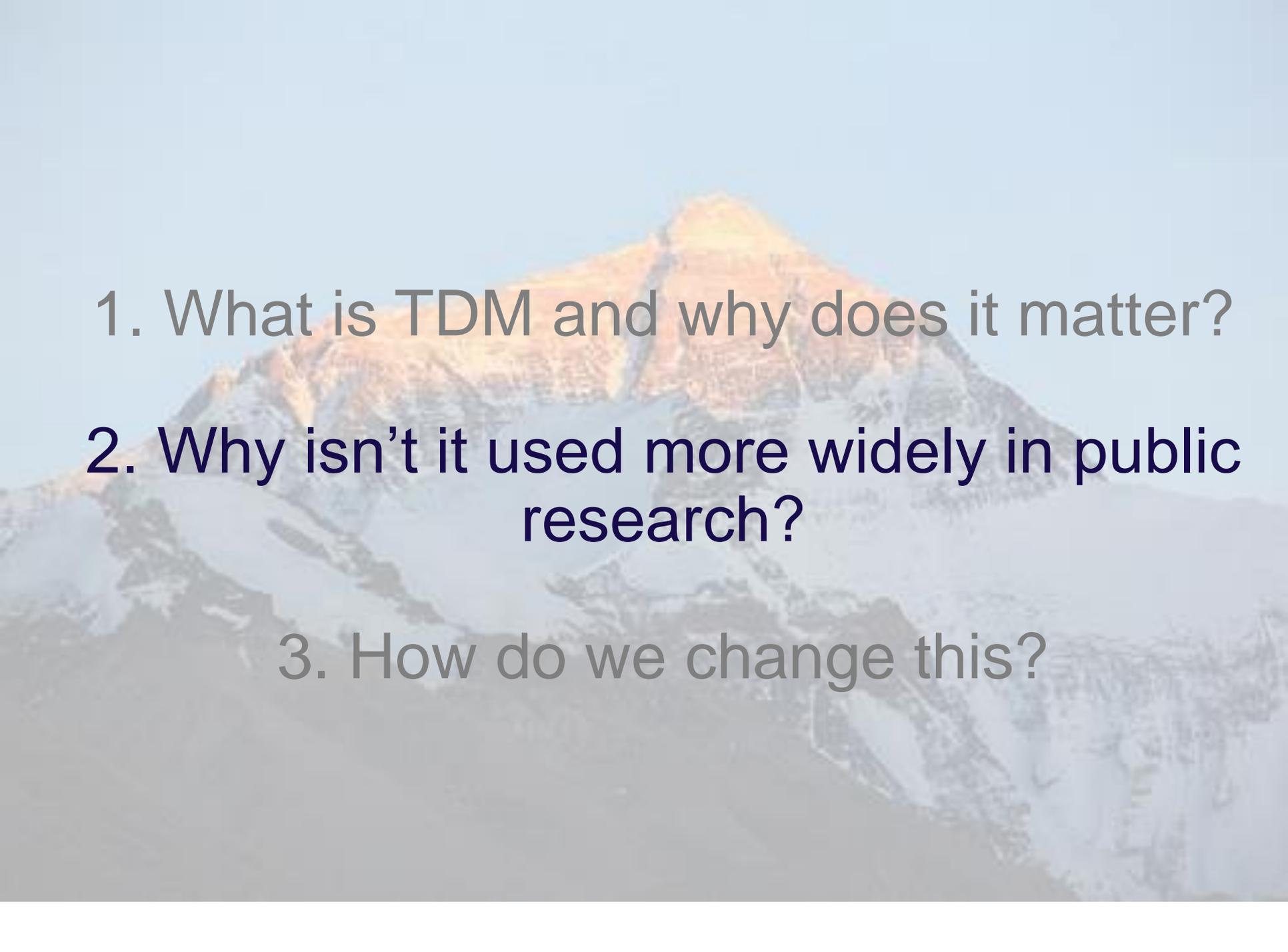
Handke, Guilbault and Vallbe [IS EUROPE FALLING BEHIND IN DATA MINING?](#) (2015)

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**

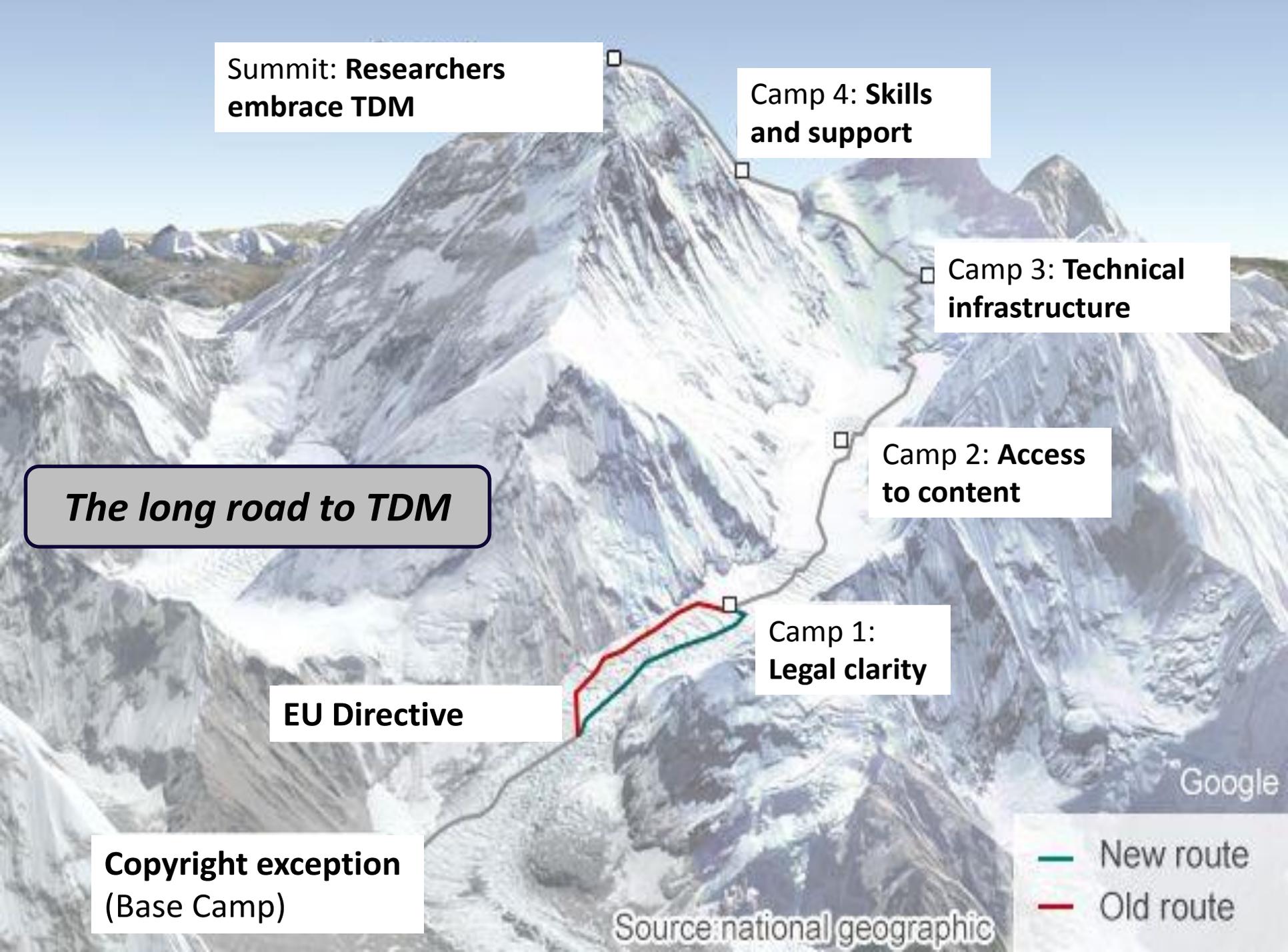
EU Directive

The long road to TDM

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.

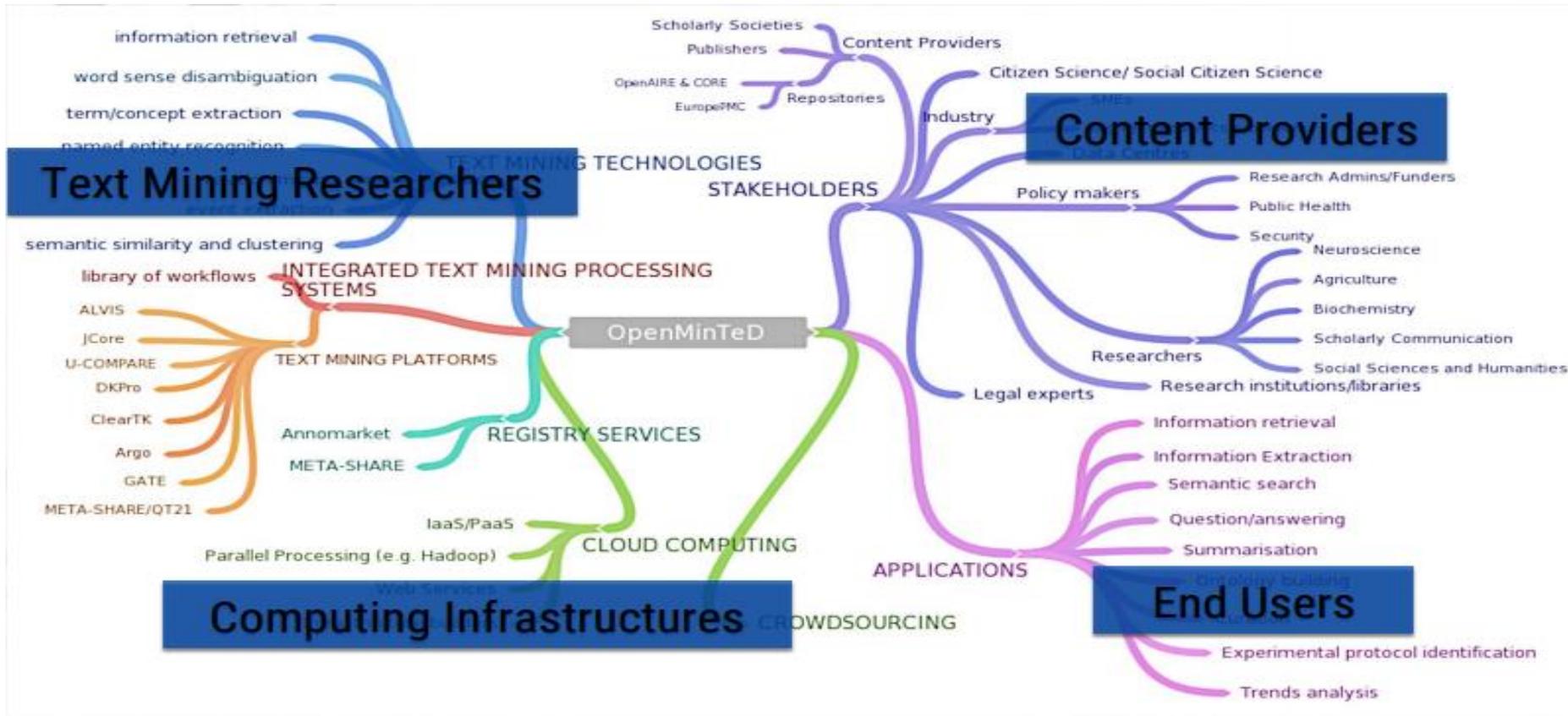
INFRASTRUCTURE & TOOLS



...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



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

Why?

“

"Because it's there"

Edmund Percival Hillary



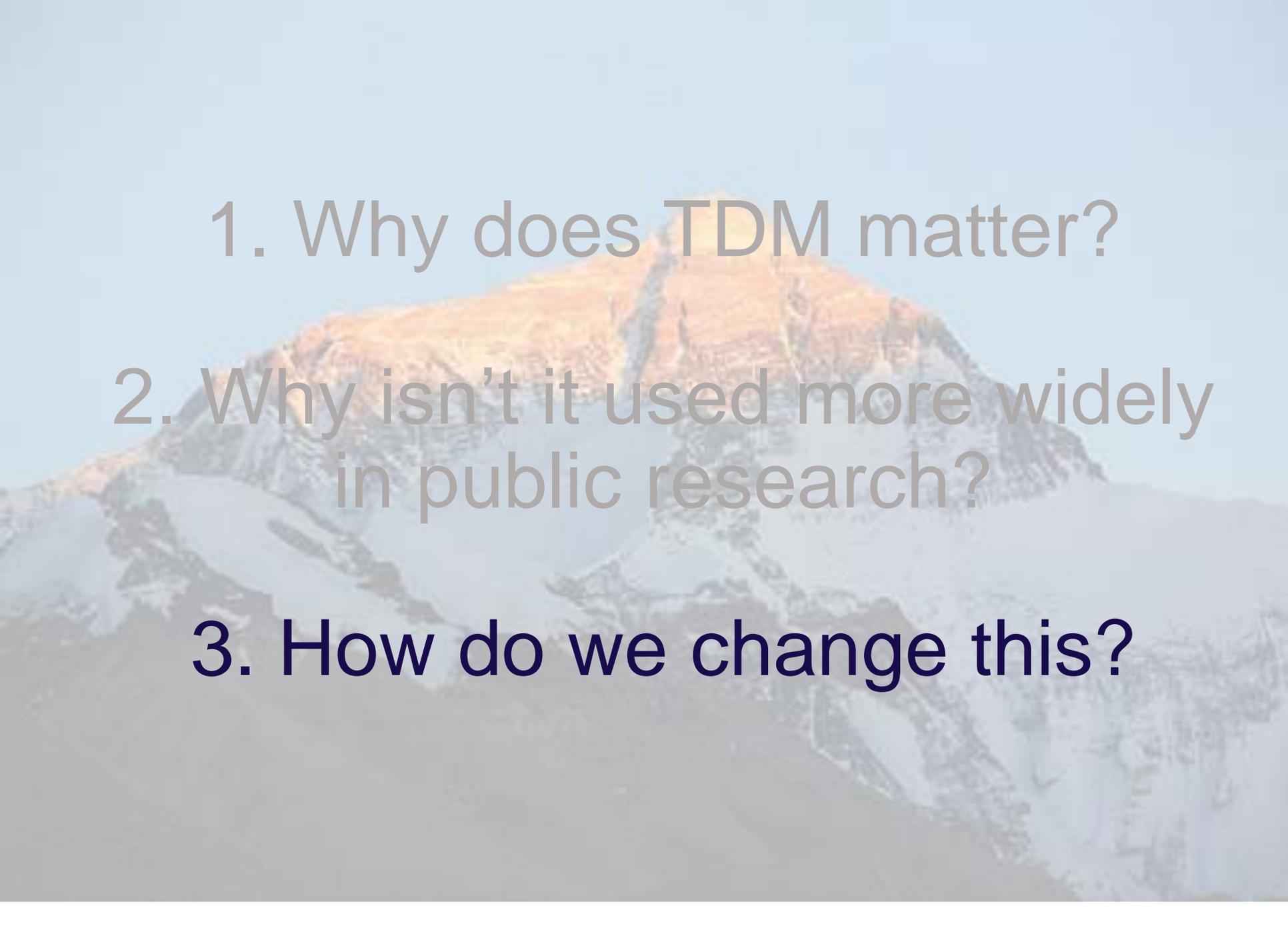
There are so many obstructions in the way of doing this research, and doing it well. It is just too hard and so people do other things

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?

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



Creating data

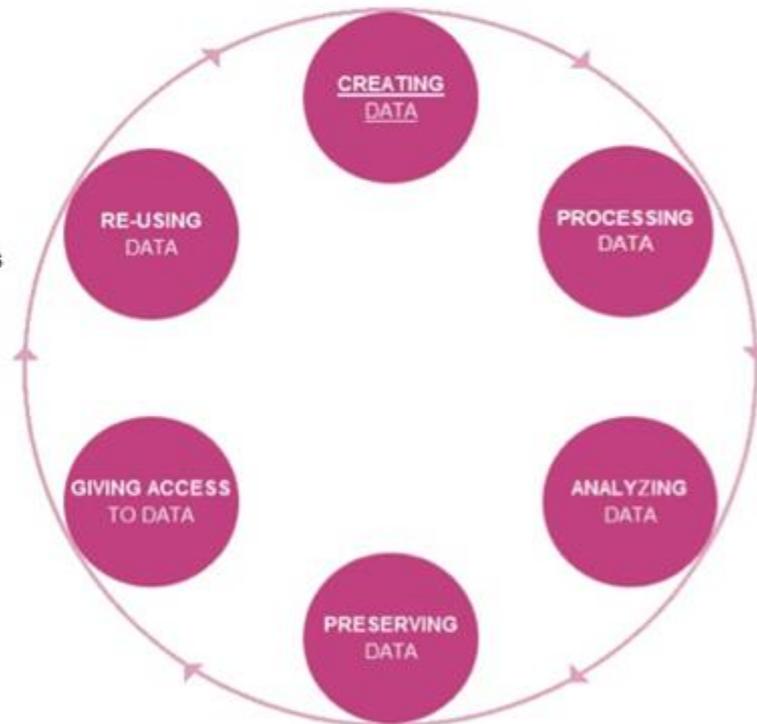
- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

Re-using data

- follow-up research
- new research
- undertake research reviews
- scrutinise findings
- teach and learn

Giving access to data

- distribute data
- share data
- control access
- establish copyright
- promote data



Processing data

- enter data, digitise, transcribe, translate
- check, validate, clean data
- anonymise data where necessary
- describe data
- manage and store data

Analysing data

- interpret data
- derive data
- produce research outputs
- author publications
- prepare data for preservation

Preserving data

- migrate data to best format
- migrate data to suitable medium
- back-up and store data
- create metadata and documentation
- archive data



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
- 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

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/>

LIBER and TDM :

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



julien.roche@univ-lille1.Fr

