

# An International Workshop on Research Data Management

November 21, 2014, Hacettepe University Culture Center, Sıhhiye, Ankara, Turkey



HACETTEPE  
UNIVERSITY



GOETHE  
INSTITUT

## *Is There a Right to Mine (r2m) or Is It a Subject to Licences?*



**Rainer Kuhlen**

**Professor of Information Science**

**Department of Computer and Information Science -  
University of Konstanz, Germany**

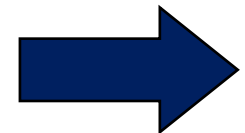
**[www.kuhlen.name](http://www.kuhlen.name)**



TÜBİTAK



Deutsch-Türkisches Jahr der  
Forschung, Bildung und Innovation 2014  
Türk-Alman Araştırma,  
Eğitim ve İnovasyon Yılı 2014



**TDM – data analysis**

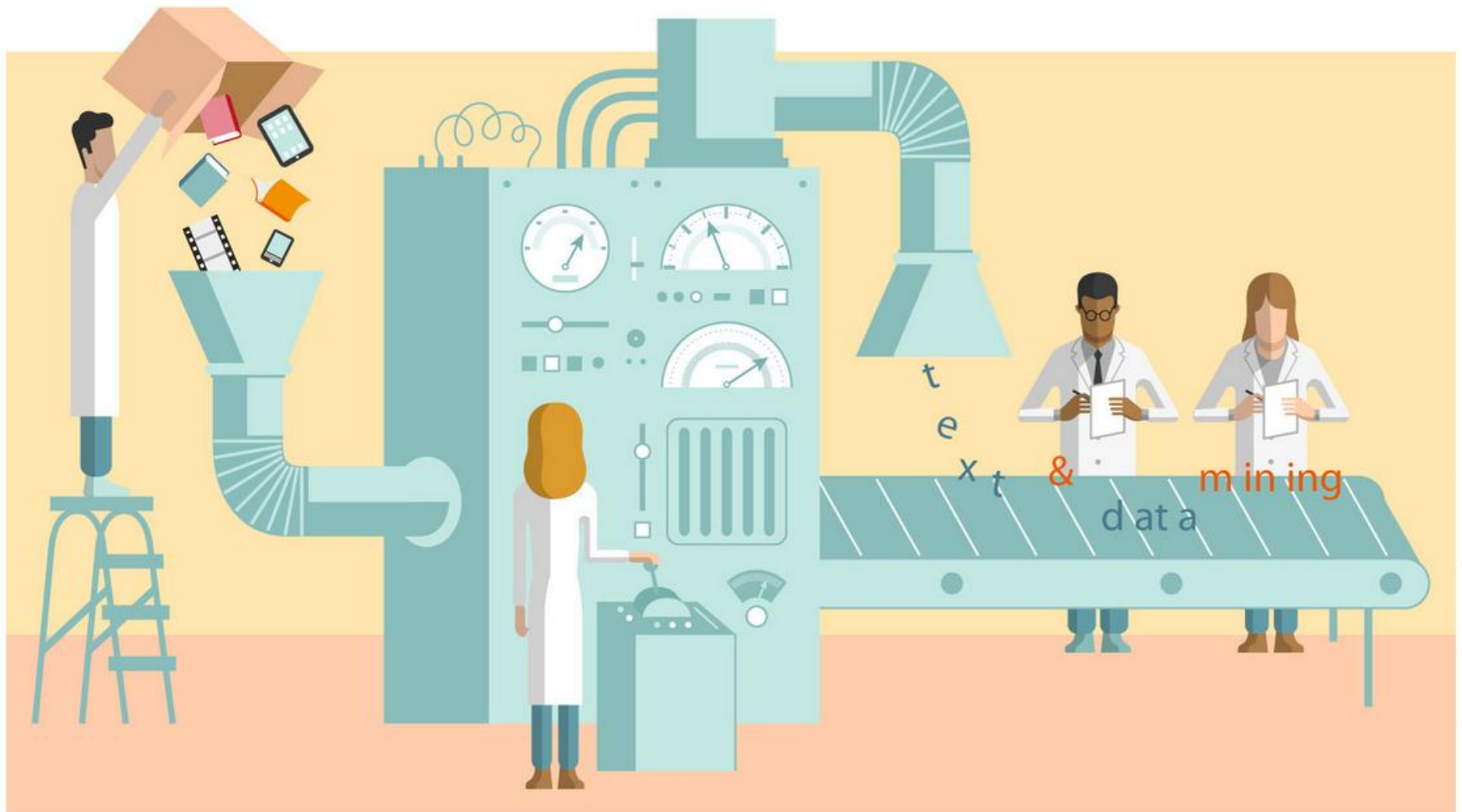
**The intellectual property right / copyright challenge to TDM**

**How publishers try to cope with the TDM challenge**

**How national regulation cope with the licenses challenge**

**How the EU Commission try to cope with the TDM challenge**

**Conclusion – a need for a new exception to copyright**



Source: <http://copyrightuser.org/topics/text-and-data-mining/>

# Main question

Should text and data mining (TDM) (data analysis) be a free domain in scientific research

or should it be a field dominated by commercial exploitation interests?

enabled by

**copyright regulation**

protected by

granted by legally binding copyright exceptions

no need for right holders' agreements

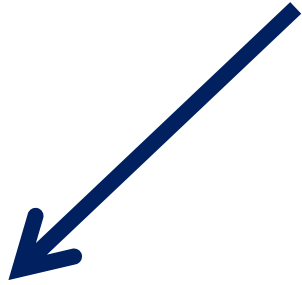
exclusive rights of right holders

right holders' contractual agreements (licenses)

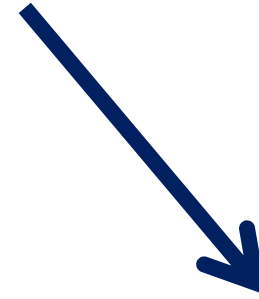
# TDM

## Data analysis

# What is TDM?

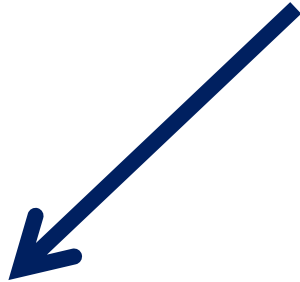


text mining



data mining

# What is TDM?



text mining

text analytics

text analysis

analysis of (large corpora  
of) unstructured text

using

linguistic, statistical, and  
machine learning  
techniques

# Text analysis

analysis of (large corpora of)  
unstructured text

using

linguistic, statistical, and  
machine learning  
techniques

methods

information  
retrieval

natural language  
processing

statistical  
techniques

goal

collecting corpora of  
relevant texts  
indexing

abstracting/summarizing

knowledge representation

automatic translation

speech recognition

proper name recognition

pattern/face / recognition

associative relations among  
terms, facts, events



# What is data?



**Database Directive** of 11 March 1996  
legal protection of databases

a **database** is by definition composed of “data”, but in reality and in accordance with the Database Directive, it can be composed of elements such as **texts, images, sounds, data**, etc. ...

“data” can therefore be considered a generic or general term and includes all types of contents such as text, images, video, etc. “

data mining

data analytics

data analysis

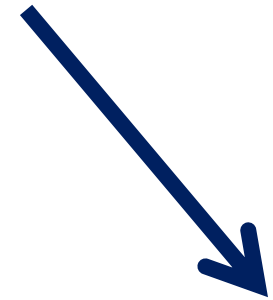
Source: [http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)  
Study on the legal framework of text and data mining (TDM) de Wolf & Partners

# What is TDM? (according de Wolf-study)

it is generally admitted that “**to mine**” means “to extract data from texts qua informational resources”, whereas data **analysis is encompassing much more than the mere extraction of data.**

data analysis

“The automated processing of **digital materials**, which may include **texts, data, sounds, images** or other elements, or a combination of these, in order **to uncover new knowledge** or insights.”



data mining

data analytics

Source: [http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)  
Study on the legal framework of text and data mining (TDM) de Wolf & Partners

# **The intellectual property right/ copyright challenge to TDM**

# What is the intellectual property right/copyright problem with TDM?

“Text and data and data analytics methods **extract** data from existing electronic information, to establish new facts and relationships, building new scientific findings from prior research. **These new methods involve copying of prior works as part of the process to extract data**” (Definition by UK IPO)

streaming is not sufficient

“1. Individual content is **extracted from outside sources (or sometimes created)** – we assume here that it falls under copyright protection or database protection. We will call this phase “Obtaining the sources”;

2. Content is, when necessary, **transformed to fit operational needs**;

3. Content is **loaded into a data set, repository or collection**;

Reproduction right

4. Data miners gain access to the data and the mining (**analysis**) **tools are applied to the data set**;

5. New knowledge is created as a result of the analysis (usually a **report can be drafted**). “

Source: [http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)

Study on the legal framework of text and data mining (TDM) de Wolf & Partners

# What is the intellectual property right/copyright problem with TDM? EU Database Directive

Article 5 a) of the Database Directive

“In respect of the expression of the database which is protectable by copyright, **the author of a database shall have the exclusive right to carry out or to authorize: (a) temporary or permanent reproduction by any means and in any form, in whole or in part**”

“author of a database shall have the exclusive right to carry out or to authorize:

- (a) temporary or permanent **reproduction** by any means and in any form, in whole or in part;
- (b) **translation, adaptation, arrangement** and any other alteration;
- (c) any form of **distribution to the public** of the database or of copies thereof (...);
- (d) any **communication, display or performance** to the public;
- (e) any reproduction, distribution, communication, display or performance to the public of the results of the acts referred to in (b).”

Source: [http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)

Study on the legal framework of text and data mining (TDM) de Wolf & Partners

# What is the intellectual property right/copyright problem with TDM? EU Database Directive

Data mining/analysis is based on extraction of data

**“the data analysis process will entail the extraction of all or a substantial part of the data held in the database** in order for such data to be processed for the purpose of “data analysis”; in this case, the authorization of the maker must be obtained. This is likely to be the case in many situations where databases are part of a corpus to be analyzed.”

Source: [http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)  
Study on the legal framework of text and data mining (TDM) de Wolf & Partners

# What is the intellectual property right/copyright problem with TDM? Permitted as part of a technological process?

**As an integral and essential part of a technological process**

*which could not function correctly and efficiently without reproducing otherwise copyright-protected material*

But there are so many conditions to be met “that this exception **will not provide much relief (or really rarely) for data analysis activities** “ (p. 50)

Source: [http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)  
Study on the legal framework of text and data mining (TDM) de Wolf & Partners

# How publishers try to cope with the TDM challenge



# How publishers try to cope with the licenses challenge - stm

Licensing initiative by stm – 2013  
12 key points for OA-compatible licences

Key point Nine

**Text and data mining user rights  
and obligations should be clarified**

Twelve Points to Make Open Access Licensing Work – stm May 2013

[http://www.stm-assoc.org/2013\\_05\\_29\\_Twelve\\_Points\\_to\\_Make\\_Open\\_Access\\_Licensing\\_Work.pdf](http://www.stm-assoc.org/2013_05_29_Twelve_Points_to_Make_Open_Access_Licensing_Work.pdf)

# How publishers try to cope with the licences challenge - EPC

European Publishers Council – From Vision to Reality June 2014

A new exception for text and data mining at EU level carries a huge risk from 'the law of unintended consequences'.

Legislation at this juncture would be highly premature especially as the publishing industry is moving fast to deliver solutions

**Any exception for text and data mining**, however carefully defined or limited, for example to non-commercial use, **could effectively destroy the very primary market** of the content that could be mined.

Even if an exception were to be limited to non-commercial research, it is impossible adequately to delineate the boundaries between research and other activities and between non-commercial and commercial.



Therefore the EPC is adamantly opposed to the introduction of a new exception in this field.




[https://www.sugarsync.com/pf/D96901\\_402827\\_8766618?directDownload=true](https://www.sugarsync.com/pf/D96901_402827_8766618?directDownload=true)

# European Publishers Council (EPC) – Support for CrossRef’s Application Programming Interface (API)

## CrossRef Text and Data Mining Services Simplify Researcher Access

29 May 2014, Boston, MA – CrossRef Text and Data Mining services, allowing publishers to provide information that will simplify access arrangements for researchers who desire to mine and analyze scholarly publisher sites, is now available to CrossRef Members. CrossRef, a not-for-profit association of worldwide scholarly publishers, made the announcement at the Society for Scholarly Publishing Annual Meeting here today.

 Publishers participating in CrossRef Text and Data Mining services may now deposit full-text links in the metadata for their DOIs, as well as license URIs by which researchers can determine whether they have permission to mine a particular content item. Through CrossRef’s Application Programming Interface (API), researchers will then be able to access the full-text, CrossRef DOI-identified content across participating publishers’ sites, regardless of their access models. 

For all publishers, whether using open access or subscription business models, CrossRef Text and Data Mining services easily direct  researchers to the appropriate location  the full text content and licenses for that content. In addition, publishers can add download rate limits  to the information they provide to minimize any impact of text and data mining activities on web site performance.

<http://www.crossref.org/01company/pr/news052914.html>

# How publishers try to cope with the licences challenge - Elsevier

## Elsevier updates text-mining policy to improve access for researchers – 1/2014

“we have adopted a **license-based approach** which both formalizes the right to mine into our academic agreements and delivers a flexible way in which researchers can gain access to our API through a self-service portal.”

**Academic subscribers:** Researchers can text mine subscribed content on ScienceDirect for non-commercial purposes, **via the ScienceDirect API's.**

**Text and data mining enabling clauses** for non-commercial purposes will be **included in all new ScienceDirect subscription agreements** and upon renewal for existing customers.

Librarians interested in **adding the TDM clause to their existing agreement** prior to renewal are able to **request a simple contract e-amendment** via their Elsevier Account Manager.

<http://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers>

<http://www.elsevier.com/about/policies/content-mining-policies>

# How publishers try to cope with the licences challenge - Elsevier

## Elsevier updates text-mining policy to improve access for researchers – 1/2014

“We request that all access to content for text mining purposes takes place through our API’s”

“When researchers have completed their text-mining project through the API, the **output can be used for non-commercial purposes** under a [CC BY-NC license](#).”

The output can **contain "snippets"** of up to 200 characters of the original text... Elsevier also requests that text-mining researchers include a **DOI link back to the original content** to ensure that authors receive credit and that future researchers have a reliable reference to the authoritative source of the underlying articles.”

<http://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers>

<http://www.elsevier.com/about/policies/content-mining-policies>

# How publishers try to cope with the licences challenge - summary

Publishers oppose a (new) TDM exception in copyright

Publishers want to keep control over TDM activities

Publishers want to control TDM by contractual licences agreements

<http://www.elsevier.com/connect/elsevier-updates-text-mining-policy-to-improve-access-for-researchers>

<http://www.elsevier.com/about/policies/content-mining-policies>

# UK national regulation TDM exception

# How national regulation can cope with the TDM challenge - UK

UK Intellectual Property Office

Exceptions to copyright: Research  
– 10/2014

What's changed?

“Text and data mining usually requires copying of the work to be analysed.

**Before the law was changed**, researchers using text and data mining in their research **risked infringing copyright** unless they had specific permission from the copyright owner.”



# How national regulation cope with the licenses challenge - UK – r2r = r2c=r2m

UK Intellectual Property Office  
Exceptions to copyright: Research

–  
What's changed?

If a researcher has the **right to read** a copyright document under the terms of the licensing agreement with the content provider, they must be permitted **to copy the work for the purpose of non-commercial text and data mining.**

r2r=r2c=r2m

Contract terms which have the effect of preventing this **will be unenforceable.**

# How does Elsevier's text mining policy work with new UK TDM law? (June 2014)

**researchers with lawful access to works published by Elsevier can copy these without asking, using tools we have provided** for this purpose, provided they are doing the copying to carry out **non-commercial text and data mining**.

For starters, **access via the API** provides full-text content of ScienceDirect in XML and plaintext formats, which researchers tell us they prefer to HTML for mining.

Under the UK legislation, publishers can use "reasonable measures to maintain the stability and security" of their networks, and so the requirement to use this **API is fully compatible with the copyright exception**.

we no longer request a project description as part of the API registration process, and we now allow TDM **output to be hosted in an institutional repository**. --- **third-party images and graphics ...cannot currently download automatically** via our API. We of course make this content **available to researchers on request ...**

<http://www.elsevier.com/connect/how-does-elseviers-text-mining-policy-work-with-new-uk-tdm-law>

# Realising the innovative potential of digital research methods: a call from the research community.

Open Letter to Michel Kolman, Senior VP Global Academic Relations, Elsevier (July 2014)

## 1. Protect the academic freedom of the researcher

Presenting researchers with no other option but **to register their details and agree to a click-through licence**, the terms of which can change at any time, in order to gain access to content for the purpose of TDM is unacceptable.

It places **undue liability on the researcher** and undermines the role of the institution as an intermediary tasked with protecting its researchers' privacy, as well as its ability to protect its intellectual capital.

we do not agree that any **publisher** should **mandate** how **researchers licence** the output of their **TDM research**.

**Research institutions** are already paying a **significant amount to Elsevier** and other publishers for electronic access to journals.

Given this significant revenue, we believe that **publisher infrastructure should be robust enough to cope with the added load of TDM**.

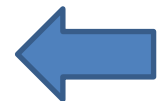
<http://libereurope.eu/wp-content/uploads/2014/07/Open-Letter-To-Elsevier1.pdf>



# Realising the innovative potential of digital research methods: a call from the research community.

Open Letter to Michel Kolman, Senior VP Global Academic Relations, Elsevier (July 2014)

- LIBER Europe
- ADBU
- CRISTIN
- CSUC
- EBLIDA
- ENCES
- FinELib Consortium
- Finnish Research Library Association (STKS)
- IFLA
- LATABA
- Library Association of Latvia
- LERU
- Open Knowledge Foundation Germany
- Portuguese Association of Librarians, Archivists and Documentalists
- REBIUN
- Research Libraries UK
- SPARC Europe
- Wellcome Trust



<http://libereurope.eu/wp-content/uploads/2014/07/Open-Letter-To-Elsevier1.pdf>

**How the EU  
Commission try to  
cope with the TDM  
challenge**

# How the EU Commission try to cope with the TDM challenge

Standardisation in the area of innovation and technological development, notably in the field of text and data mining

Directorate-General for Research and Innovation - Directorate B — Innovation Union and European Research Area Unit B.1 — Innovation Union Policy

Yet the **European digital economy has been slow in embracing the data revolution compared to the USA**

this results, at least in part, from the nature of Europe's laws with regard to **copyright, database protection** and, perhaps increasingly, **data privacy**.

[http://ec.europa.eu/research/innovation-union/pdf/TDM-report\\_from\\_the\\_expert\\_group-042014.pdf](http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf)

# EU action to provide the right framework conditions

## Regulatory issues

### 1. Personal data protection and consumer protection

The fundamental right to personal data protection applies to big data where it is personal: **data processing has to comply with all applicable data protection rules.**

**The Commission's reform package aims to build a single, modern, strong, consistent and comprehensive data protection framework for the EU.**

By strengthening individuals' trust and confidence in the digital environment and enhancing legal certainty, it will provide a regulatory environment essential for the development of innovative and sustainable data goods and services.

**EU-Commission: Towards a thriving data-driven economy - {SWD(2014) 214 final} -**  
<http://edz.bib.uni-mannheim.de/edz/pdf/swd/2014/swd-2014-0214-en.pdf>

# EU action to provide the right framework conditions

## Regulatory issues

### 2. Data-mining

The Commission is investigating ways in which **data-driven innovation based on data-mining, including text-mining**, might be enhanced, including in relation to the relevant copyright aspects.

The Commission takes note of Member States' initiatives that facilitate these activities by **implementing (or reviewing the implementation of) the exceptions available under the current copyright framework.**

**EU-Commission: Towards a thriving data-driven economy** - {SWD(2014) 214 final} - <http://edz.bib.uni-mannheim.de/edz/pdf/swd/2014/swd-2014-0214-en.pdf>



# EU action to provide the right framework conditions

## Regulatory issues

### **Fostering Open Data policies**

To facilitate the implementation of the EU open data policy and legal framework, the Commission is preparing guidelines on recommended standard licences, datasets and charging for the re-use of documents.

**EU-Commission: Towards a thriving data-driven economy - {SWD(2014) 214 final} -**  
<http://edz.bib.uni-mannheim.de/edz/pdf/swd/2014/swd-2014-0214-en.pdf>

# Conclusion

if ... a **new specific data mining exception is introduced** and considered **unwaivable**, then contract terms, license terms (included in DRMs information, in websites' terms and conditions or in negotiated license contracts) could no more impact the possibility to do TDM:

the licensor would not have the possibility, via contractual terms, to prohibit datamining – at least as long as the other conditions of the exception would be respected).

Study on the legal framework of text and data mining (TDM). De Wolf & Partners – commissioned by the European Commission 03/2014 -

[http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)

If this **recommendation is not followed**, and if the licensor remains free to prohibit data mining as he wishes, then obviously, contract law and license terms will remain **a significant obstacle to data mining**, since researchers will need to have the authorization of the licensor before they can start a TDM project.

Study on the legal framework of text and data mining (TDM). De Wolf & Partners – commissioned by the European Commission 03/2014 -

[http://ec.europa.eu/internal\\_market/copyright/docs/studies/1403\\_study2\\_en.pdf](http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf)



Attribution 3.0 Unported (CC BY 3.0)

## You are free to:

**Share** — copy and redistribute the material in any medium or format

**Adapt** — remix, transform, and build upon the material

for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

## Under the following terms:



**Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

**No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

