Café on Text and Data Mining Exceptions a Year After *Has the Pony become a Horse?*

CLARIN ERIC

08 November 2022

CLARIN

Organisers

This edition of the CLARIN Café is organized by

This edition of the CLARIN Café is organised by Paweł Kamocki, chair of the CLARIN Legal and Ethical Issues Committee (CLIC).

CLARIN hosts is
Antal van den Bosch (*CLARIN BoD*)

Technical support by

David Bordon

The event is recorded for further dissemination purposes.

Questions and comments? Put them in the chat box.

Schedule

14:00 - 14:15 Introduction and CLARIN 101 - Antal van den Bosch (CLARIN ERIC Board of Directors)

14.15 - 14.35 A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology

Thomas Margoni, KU Leuven

14.35 - 14.55 Tabula rasa: TDM exceptions in post-Brexit UK

Toby Bond, Bird & Bird

14.55 - 15.15 Imagine all the researchers crawling the Internet in peace. The HPLT project and the future of European language research

Jan Hajič, Charles University Prague

15.15 - 16.00 Discussion

Introducing CLARIN

https://www.youtube.com/watch?v=ut9wOIYWDfc

https://www.clarin.eu/content/clarin-in-a-nutshell

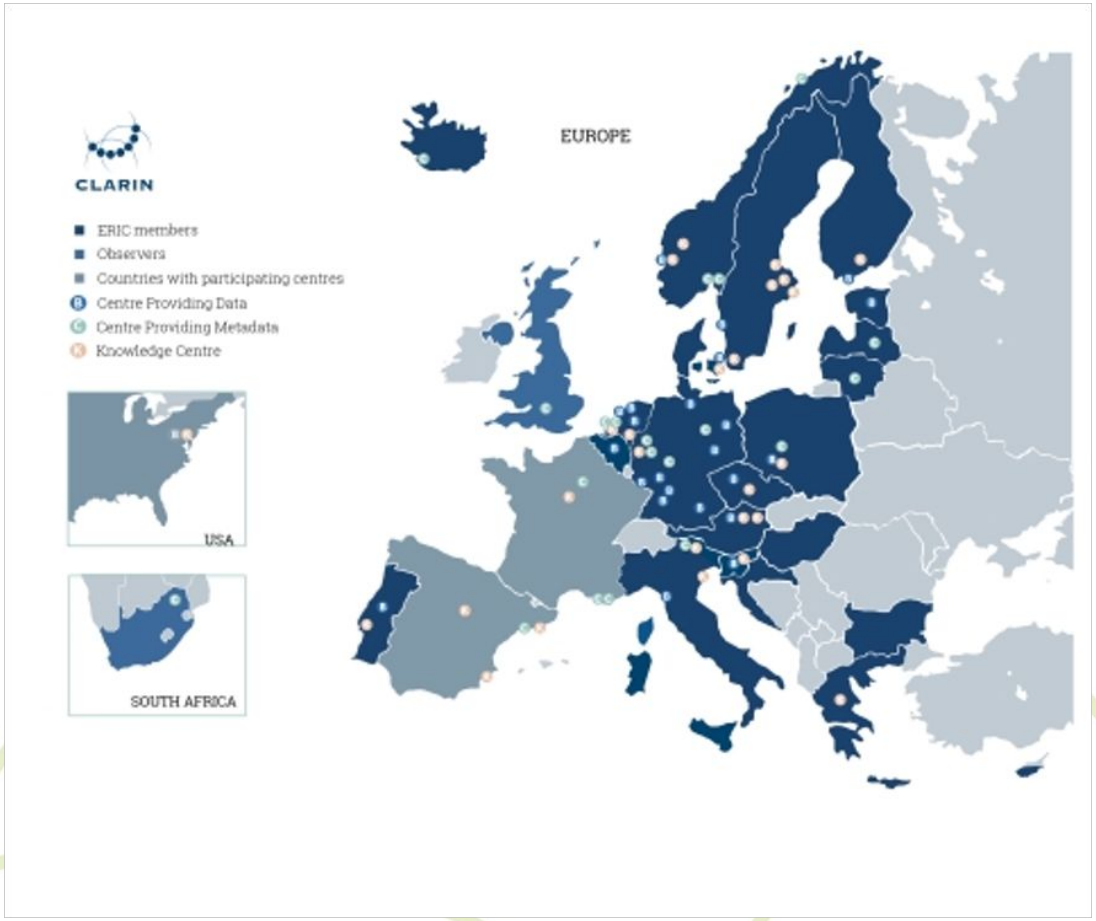
CLARIN

CLARIN …

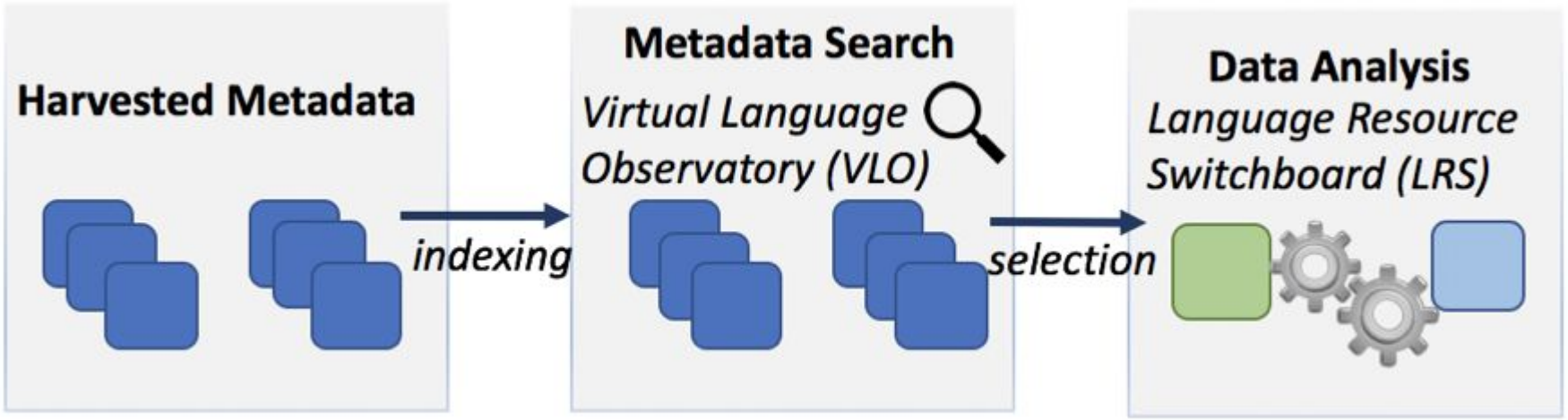
- is the *Common Language Resources and Technology Infrastructure*
- has the **ESFRI** ERIC status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing and training**
- is one of the European RIs in the SSH cluster (aka SCI)
- is an integral part of **the European Open Science Cloud**
  - See **clarin.eu/eosc**

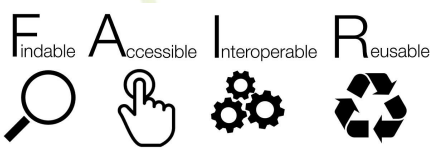
CLARIN today

- a distributed network of **70 centres**
- **22 members**: AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU,IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
- **2 observers**: UK, ZA
- **1 third party**



The Technical Infrastructure





clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu

The Knowledge Infrastructure

https://www.clarin.eu/content/knowledge-infrastructure

CLARIN – CLIC

Legal and Ethical Issues Committee

- https://www.clarin.eu/governance/legal-issues-committee

Legal Information Platform

- https://www.clarin.eu/content/legal-information-platform
- https://www.clarin.eu/content/bibliographyfurther-reading-legal-and-ethical-issues

Previous CLARIN cafés:

- 30 March 2021 - CLARIN Café on the Rights of Data Subjects in Language Resources
- 28 October 2021 - CLARIN Café on Text and Data Mining Exceptions in the Directive on Copyright in the Digital Single Market

The Café

A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology

Thomas Margoni

CiTiP, KU Leuven

CLARIN

CDSM: some key points

- Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market to make EU copyright fit for the digital age
- Introduces mandatory exceptions for e.g. text and data mining (Arts. 3 and 4)
- Clarifies Art. 5(1) Info Soc Directive 2001 (mandatory exception for certain temporary acts of reproduction) continues to apply to TDM as before
- Clarifies that broader national exceptions adopted on the basis of aquis remain available, e.g. national TDM exceptions based on e.g. Art. 5(3)a InfoSoc (non commercial research)

Arts. 3&4

Definition: “any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations” (Art. 2 CDSM);

Scope: exception to the right of reproduction (3&4);

Beneficiaries: research&cultural organisations for research purposes (Art. 3 CDSM), anyone for any purpose but can be opted-out (Art. 4 CDSM).

Type of access: lawful access (3&4)

Relationship to contracts: Cannot be limited by contract (3); can be opted out if express reserve in appropriate manner (4)

Relationship to technology: Can be limited by technological measures (integrity measures and TPM although in different ways)

Storage: Different wording but both 3&4 allow for retention of stored copies for verification (3) and TDM (4)

Observations

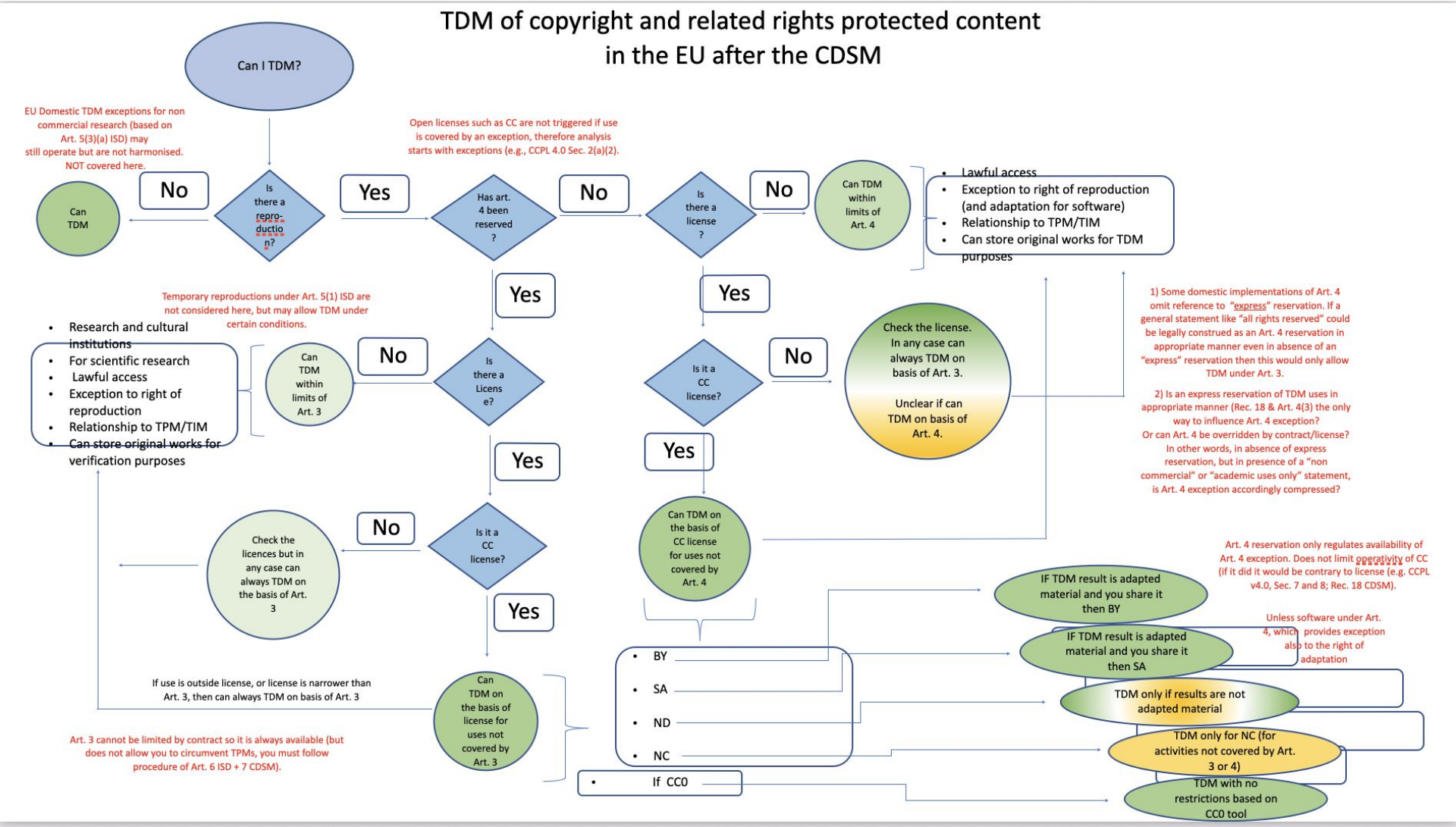
EU Acquis: quantitative low level of originality; broad right of reproduction; SGDR; limited exceptions and limitations, no transformative/free uses as such, all fundamental rights limitations to C must be found in Art. 5, etc.

Claim: **Reuse of non personal data is much more "costly" in EU than outside the EU**

Definition is very broad, does not only cover TDM but virtually all data analytic techniques, e.g., modern AI

Claim: **most of EU AI largely rely on 2,5 copyright exceptions**

Picture here



Final observations

Property-based approach to data is problematic. AI applications can only be developed based on a narrow (or wider but non imperative) copyright exception in the EU.

- Is the function of copyright to be the ultimate judge of whether, how and by whom technological development can happen and in which direction?
- If we read CDSM IA probably not, TDM focus on needs of research organisations to perform research and needs of publishers to retain their licensing business model
- Industrial and innovation policy assessment was not central

Property rights as a right to say no (authorization to use) and establish **conditions** (availability, price, purposes).

- Is the function of copyright to offer data holders control over downstream markets such as AI development?
- If yes, how would this policy (in)decision shape AI markets?

Hypothetical (?) scenarios

- PD material

  EU AI as an AI based on an average 70 years old body of knowledge? Or only on information contained in Wikipedia? **Second class AI?**

- Who are willing/can pay the price

  will EU AI be then more expensive/less competitive than US AI? Or Japan? CH? UK AI?)

- Train outside the EU in "cheaper" legal systems or import pretrained models in EU

  What would be the impact of AI trained on data embedding a system of knowledge, values and rules belonging to a different tradition? E.g.: See Art. 17, **would we import in the EU a US based concept of "parody" via close-to-mandatory filtering obligations?**

- Or train in the EU anyway and don't tell (difficult to reverse engineer models)

  Would this lead to **opacity in the training process** (which would plausibly contrast with high-risk AI in AIA) – not a desirable mix of incentives for innovation.

- Or you already possess a vast databases to use (e.g. internet platforms who normally acquire contractual permission to use user uploaded content to improve their own services)

Additional resources

Margoni T., Kretschmer M**., A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology**, GRUR INT, Volume 71, Issue 8, August 2022, Pages 685–701, https://doi.org/10.1093/grurint/ikac054

Ducuing, Margoni, Schirru (Eds.), **CiTiP's White Paper on the Data Act Proposal** https://www.law.kuleuven.be/citip/blog/category/data-act-series/

Margoni, Quintais, Schwemer, **Algorithmic propagation: do property rights in data increase bias in content moderation?**, http://copyrightblog.kluweriplaw.com/2022/06/08/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-i/

Report on AI Data Inputs and accompanying background material: https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/

AI, Machine Learning and EU Copyright Law: A Socio-Legal Analysis of Ownership Issues in Training Data in the Context of Three Case Studies, interim report https://zenodo.org/record/5069507

Tabula rasa: TDM exceptions in post-Brexit UK

Toby Bond

Bird & Bird

The past…

InfoSoc Directive - Article 5

1. Temporary acts of reproduction referred to in Article 2, which are transient or incidental and an integral and essential part of a technological process and whose sole purpose is to enable:

(a) a transmission in a network between third parties by an intermediary, or

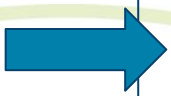
(b) a lawful use

of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2.

---

3. Member States may provide for exceptions or limitations to the rights provided for in Articles 2 and 3 in the following cases:

(a) **use for the sole purpose of** illustration for teaching or **scientific research**, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved;



June 2014 (following adoption of recommendation from 2011 Hargreaves Review)

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).

…

(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**.

The present…

lawful access

- Access is lawful where researchers have the legal right to access a copyright work to read it.
- Examples could include paying for a subscription to a journal or database or material published under open licences including Creative Commons and Open Government Licences.

The present…

Contractual restrictions on TDM

- Contract terms which have the effect of preventing use of the exception become unenforceable.
- This applies to licence terms in place before entry into force of the exception in 2014.

Technical restrictions on access

- Publishers may wish to apply technological measures on networks for a number of purposes such as to ensure security or stability.
- Examples of possible measures could be to impose a reasonable limit on download speeds or to control the number of times a user can access a network in a given period.
- These measures should not stop or unreasonably restrict any researcher's ability to benefit from the exception.

The present…

Commercial vs non-commercial

- Contract research for an outside company is unlikely to falls within the definition of non-commercial.
- University departments funded by a company can still perform non-commercial research if researchers can choose their own research topics and are free to publish my work without interference from the company.

Can the results of non-commercial research be used for commercial purposes?

- If results are simply facts they are not covered by copyright.
- There are no restrictions on how or where outputs of text and data mining can be published, including journals published for profit by academic publishers and under licences that permit commercial research, such as CC-BY. Other commercialisation of the research outputs is not restricted either.
- It is important to be scrupulous in assessing whether the original purpose of carrying out the text and data mining analysis is solely non-commercial; if it isn't, then researchers are very likely to be infringing copyright.

The future…

DSM transposition date fell after the end of the Brexit transition period…

Article 3

Text and data mining for the purposes of scientific research

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

2. Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of research, including the verification of research results.

3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.

4. Member States shall encourage rightholders and research and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligations referred to in paragraphs 2 and 3 respectively.

Article

Exception or limitation for text and data mining

1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

4. This Article shall not affect the application of Article 3 of this Directive.

October 2021 - UKIPO consultation on TDM exceptions

Title: Consultation stage impact assessment on Artificial Intelligence and Intellectual Property.
IA No:

RPC Reference No: RPC-BEIS-IPO-5101(1)
Lead department or agency: Intellectual Property Office (an executive agency of the Department for Business, Energy and Industrial Strategy).
Other departments or agencies:

Impact Assessment (IA)

Date: 29 October 2021
Stage: Consultation
Source of intervention: Domestic
Type of measure: Primary legislation
Contact for enquiries: AICallForViews@ipo.gov.uk

Summary: Intervention and Options

Cost of Preferred (or more likely) Option (in 2019 prices)

Total Net Present Social Value	Business Net Present Value	Net cost to business per year	Business Impact Target Status
£m	£m	£m	Qualifying provision

What is the problem under consideration? Why is government action or intervention necessary?

The government wants the UK to be the best place in the world for research and innovation, and at the forefront of the artificial intelligence and data revolution. The new National AI Strategy will secure the UK's position amongst the global AI superpowers. Venture Capital investment in UK firms increased significantly over the last five years and UK firms are the main beneficiaries in Europe. However, AI uptake remains low relative to other European countries. IP is one of the levers available to government to increase returns on investments for inventors and creators and thereby incentivise investment in AI to invent and create. This consultation considers whether the current IP regime strikes the appropriate balance to encourage the development of AI and its use across the UK economy.

What are the policy objectives of the action or intervention and the intended effects?

The government's objective is to incentivise investment in AI development and to promote the use of AI for public benefit, whilst enabling competitive markets, consumer choice and fair access to IP-protected goods for the benefit of society.

What policy options have been considered, including any alternatives to regulation?

At consultation stage there are no preferred options.

We are consulting on 3 areas of potential change for the Intellectual Property regime.

***Section A: Computer Generated Works (CGW) options*:** Option 0- no change, Option 1 - remove CGW protection, Option 2- replace current provision with an alternative with a reduced scope/duration.

Section B: Text and Data Mining (TDM) options: Option 0- no legal change but possible guidance, Option 1 - adopt a licence-based model, Option 2 - extend existing exception to cover commercial research, Option 3 - adopt an exception for any use with a rights holder opt out, Option 4 - adopt an exception for any use with no possibility for rights holder opt out.

***Section C: Patent options*:** option 0- no legal change, option 1- expand definition of "inventor", option 2- recognise AI as inventor in patent applications, option 3- protect AI devised invention through a new type of protection.

The future…

Options Considered

- **Option 0**: Stick with only the current exception under section 29A CDPA, perhaps with updated guidance on the definition of non-commercial research.
- **Option 1**: Improve the licensing environment for the use of works and databases for TDM through educational material, model licences or codes of practices which would assist parties to conclude TDM licences. The Impact Assessment contains an intriguing but passing reference to a legislative backstop for codes of conduct, which suggests that under this option the Government could consider legislating in future if voluntary codes of conducts are not followed. It also mentions the potential use of the extended collective licensing framework.
- **Option 2**: Extend section 29A CDPA to cover commercial scientific research and database rights. This would provide a slightly broader exception than the EU's Article 3, as it would be defined solely based on the purpose (scientific research) and the beneficiaries would not be limited to research organisations and cultural heritage institutions.
- **Option 3**: Adopt a TDM exception to copyright and database rights permitting both commercial or non-commercial TDM but with the ability of rights holders to opt-out. This would effectively be modelled on the EU's Article 4 exception.
- **Option 4**: Option 3, but without the option for rights holders to opt out of the exception.

The future…

And the winner was…

- **Option 4**: A TDM exception to copyright and database rights permitting both commercial or non-commercial TDM but without the option for rights holders to opt out of the exception.

  Rights holders will no longer be able to charge for UK licences for TDM and will not be able to contract or opt-out of the exception. The new provision may also affect those who have built partial business models around data licensing. However, rights holders will still have safeguards to protect their content. The main safeguard will be the requirement for lawful access. That is, rights holders can choose the platform where they make their works available, including charging for access via subscription or single charge. They will also be able to take measures to ensure the integrity and security of their systems.

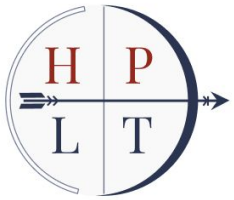
  The Government's ambition is to make the UK a global centre for AI innovation. The new exception will ensure the UK's copyright laws are among the most innovation-friendly in the world. All users of data mining technology will benefit, with rights holders having safeguards to protect their content.

- **Timing** of proposed legislation currently unknown.

The future…

The wider perspective..

- Exceptions (and the prohibition on contractual opt-out) will only apply where data is protected by copyright or database right.
  - Impact of Brexit on UK protection of databases made by EU based makers.
  - Retained EU Law (Revocation and Reform) Bill?
- Are data providers in a better position if their data is not protected by copyright or database rights?
- Interaction with other potential legal rights which are potentially engaged when web scraping, e.g. Computer Misuse offences and trespass to chattels?
- If the exceptions only cover the TDM process, are there circumstances where use of the results of TDM could also give rise to a copyright or database rights infringement, e.g. generative networks?
  - GitHub Copilot Litigation filed 3 November 2022 - alleging violation of OSS attribution requirements
- Can running TDM on virtual machines in jurisdictions with broader exceptions circumvent the lack of local exceptions?



Imagine all the researchers crawling the Internet in peace. The HPLT project and the future of European language research

Jan Hajič

Charles University Prague

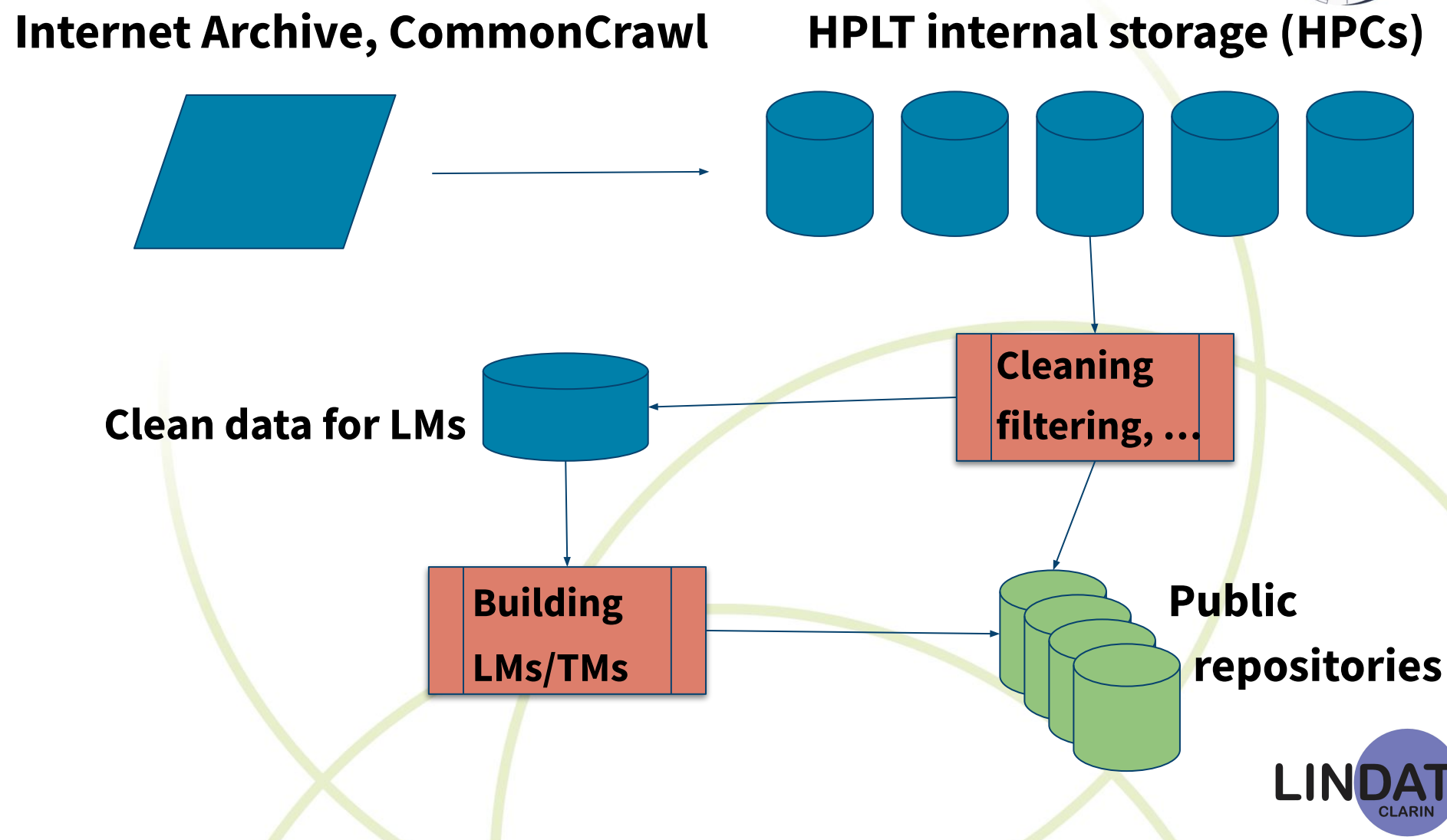
The HPLT project

- HPLT = High-Performance Language Technologies
  - Univ. of Edinburgh, Charles Univ., Oslo, Helsinki, Turku, Prompsit, HPC: Sigma2, CESNET
- Horizon Europe DATA project
  - 2022-2025 (Started Sept. 1, 2022)
- Goal
  - Get large amounts of (textual) data in 30 languages
    - Internet Archive (located in the U.S.)
    - CommonCrawl
    - … (other large sources)
  - Create large language and translation models
  - Make all of it available for the research community
    - Open, free of charge

HPLT: current status

- Agreement with Internet Archive
  - In the works
  - Technical details
  - Legal conditions
    - Can copy and store data for HPLT project partners
      - Agreement is with University of Oslo
    - Can process data and create derivative works
      - Derivative works can be shared freely
    - Payment (IA is Nonprofit)
- Other source (CommonCrawl)
  - No legal issues, data is available
- Total amount
  - 12 PB of raw data

HPLT: schema


Internet Archive, CommonCrawl
HPLT internal storage (HPCs)
Clean data for LMs
Cleaning
filtering, …
Building
LMs/TMs
Public
repositories

HPLT: Flow of Data

- Download from the Internet Archive
  - Store at CESNET (IT4Innovations, Czech eInfra), Sigma2
- Clean the data (filtering, language ID, deduplication, parallel data discovery, formatting for later use, metadata creation)
  - at CPU clusters in CZ and NO (FI?)
  - decreases size to about 10% of original
- Train large language models
  - Evaluate for a number of standard LM applications
- Train large translation models
  - Add to the OpusMT collection, evaluate on MT
- Publish data in repositories for long-term preservation
  - replicability, further processing, …
  - Huggingface, ELG, LINDAT, OpusMT, …

HPLT: Any legal issues?

- Internet Archive
  - contains data from all over the world
  - legal grounds: Fair Use (U.S.)
  - what happens when copied to Europe? Is the licence signed with IA enough to copy, store, process it here?
- Derivative works
  - clean and filtered data is derivative work (or…?)
    - original metadata is filtered out, layout is deleted, pictures/media are deleted, texts are preserved (almost all)
    - what about (chain) re-sharing? I.e., can we give it CC-type license?
    - Personal data - is pseudonymization enough?
- Resulting models
  - not derivative works: how to “licence” them?

Getting involved in CLARIN

Join our NewsFlash

https://www.clarin.eu/content/newsflash

Check out our events

https://www.clarin.eu/events

Open calls

https://www.clarin.eu/content/funding-opportunities

Next events

CLARIN café on the “Ravensbruck project” in December, 2nd

Stay tuned: https://www.clarin.eu/content/clarin-cafe

Share your **#clarincafe** impressions with **@CLARINERIC**