Digital Law Center University of Geneva

AI Tech & Policy Talks

Property right approaches in the regulation of AI in the European Union

Prof. Dr. Thomas Margoni
Research Professor of Intellectual Property Law
Centre for IT & IP Law (CiTiP)
Faculty of Law, University of Leuven (KUL)

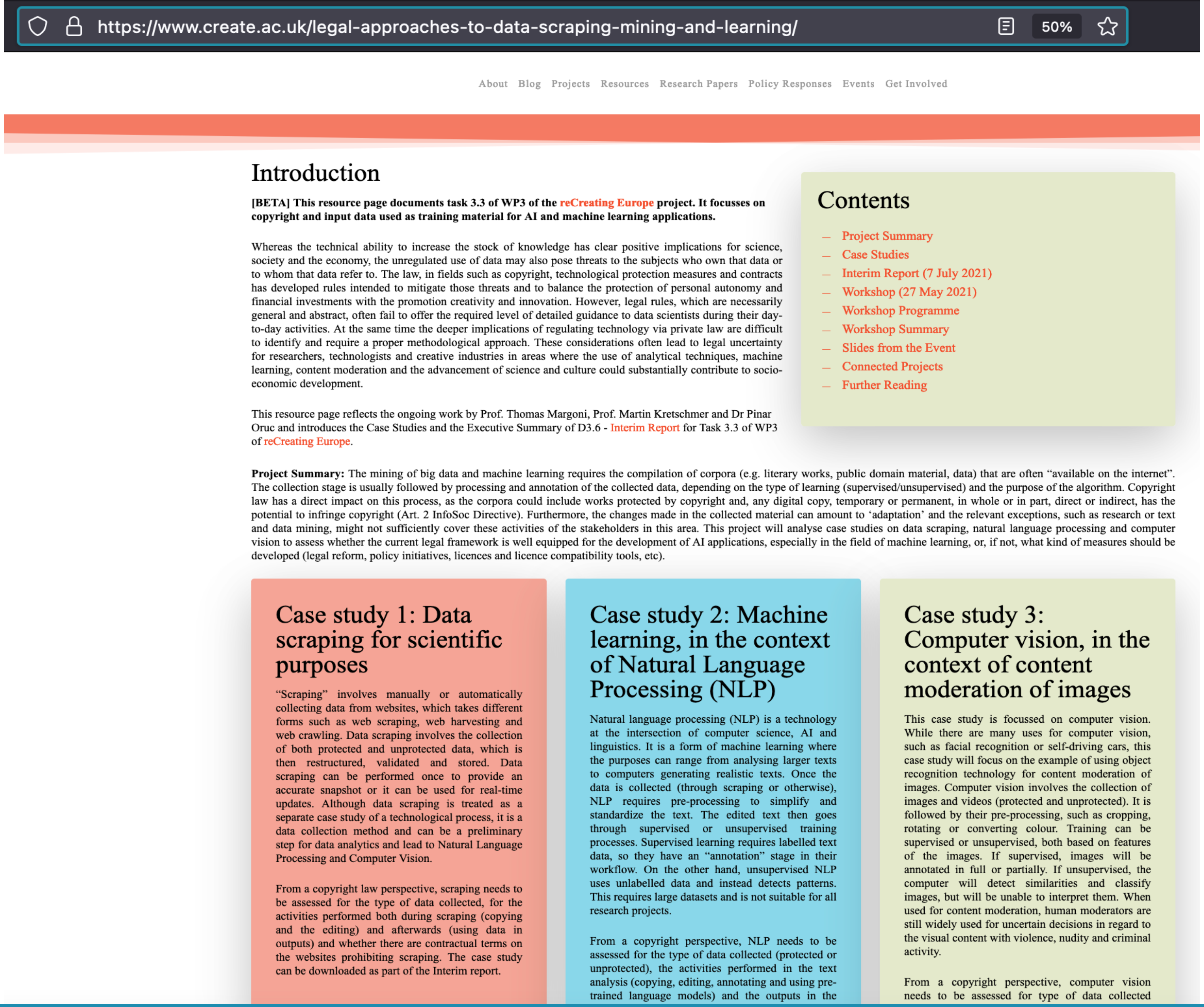
Background: Project and objectives

- Part of our work in ReCreating Europe project (www.recreating.eu, H2020, WP3, Task 3.3).
- Focus **on training data in AI applications** (machine learning and other data intensive approaches) from **a copyright perspective** (subject matter, selected rights and exceptions, harmonization and influence on/of technology).
- Dedicated web resource documenting work: https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/ and project website www.recreating.eu
- Team: Thomas Margoni, Martin Kretschmer, Pinar Oruc (until 2021)

“AI, Machine Learning and EU Copyright Law: A Socio-Legal Analysis of Ownership Issues in Training Data in the Context of Three Case Studies” (forthcoming)

1. Creation and development of 3 case studies in data analytics (by Oruc under scientific supervision of Kretschmer&Margoni):
   - Data scraping*
   - NLP
   - Computer vision
2. Inductive approach:
   - From the reality (represented in the cases) to the law and policy
   - To avoid the filters of pre-existing representations (policy, professional, business groups, lobbying, etc) shaping the discourse around TDM
   - Developed in consultation with scientific researchers and stake holders
3. Analysis is conducive and support findings of legal paper and conclusions

Case studies



https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/
50%
About Blog Projects Resources Research Papers Policy Responses Events Get Involved

Introduction

[BETA] This resource page documents task 3.3 of WP3 of the reCreating Europe project. It focusses on copyright and input data used as training material for AI and machine learning applications.

Whereas the technical ability to increase the stock of knowledge has clear positive implications for science, society and the economy, the unregulated use of data may also pose threats to the subjects who own that data or to whom that data refer to. The law, in fields such as copyright, technological protection measures and contracts has developed rules intended to mitigate those threats and to balance the protection of personal autonomy and financial investments with the promotion creativity and innovation. However, legal rules, which are necessarily general and abstract, often fail to offer the required level of detailed guidance to data scientists during their day-to-day activities. At the same time the deeper implications of regulating technology via private law are difficult to identify and require a proper methodological approach. These considerations often lead to legal uncertainty for researchers, technologists and creative industries in areas where the use of analytical techniques, machine learning, content moderation and the advancement of science and culture could substantially contribute to socio-economic development.

This resource page reflects the ongoing work by Prof. Thomas Margoni, Prof. Martin Kretschmer and Dr Pinar Oruc and introduces the Case Studies and the Executive Summary of D3.6 - Interim Report for Task 3.3 of WP3 of reCreating Europe.

Contents

- Project Summary
- Case Studies
- Interim Report (7 July 2021)
- Workshop (27 May 2021)
- Workshop Programme
- Workshop Summary
- Slides from the Event
- Connected Projects
- Further Reading

Project Summary: The mining of big data and machine learning requires the compilation of corpora (e.g. literary works, public domain material, data) that are often “available on the internet”. The collection stage is usually followed by processing and annotation of the collected data, depending on the type of learning (supervised/unsupervised) and the purpose of the algorithm. Copyright law has a direct impact on this process, as the corpora could include works protected by copyright and, any digital copy, temporary or permanent, in whole or in part, direct or indirect, has the potential to infringe copyright (Art. 2 InfoSoc Directive). Furthermore, the changes made in the collected material can amount to ‘adaptation’ and the relevant exceptions, such as research or text and data mining, might not sufficiently cover these activities of the stakeholders in this area. This project will analyse case studies on data scraping, natural language processing and computer vision to assess whether the current legal framework is well equipped for the development of AI applications, especially in the field of machine learning, or, if not, what kind of measures should be developed (legal reform, policy initiatives, licences and licence compatibility tools, etc).

Case study 1: Data scraping for scientific purposes

“Scraping” involves manually or automatically collecting data from websites, which takes different forms such as web scraping, web harvesting and web crawling. Data scraping involves the collection of both protected and unprotected data, which is then restructured, validated and stored. Data scraping can be performed once to provide an accurate snapshot or it can be used for real-time updates. Although data scraping is treated as a separate case study of a technological process, it is a data collection method and can be a preliminary step for data analytics and lead to Natural Language Processing and Computer Vision.

From a copyright law perspective, scraping needs to be assessed for the type of data collected, for the activities performed both during scraping (copying and the editing) and afterwards (using data in outputs) and whether there are contractual terms on the websites prohibiting scraping. The case study can be downloaded as part of the Interim report.

Case study 2: Machine learning, in the context of Natural Language Processing (NLP)

Natural language processing (NLP) is a technology at the intersection of computer science, AI and linguistics. It is a form of machine learning where the purposes can range from analysing larger texts to computers generating realistic texts. Once the data is collected (through scraping or otherwise), NLP requires pre-processing to simplify and standardize the text. The edited text then goes through supervised or unsupervised training processes. Supervised learning requires labelled text data, so they have an “annotation” stage in their workflow. On the other hand, unsupervised NLP uses unlabelled data and instead detects patterns. This requires large datasets and is not suitable for all research projects.

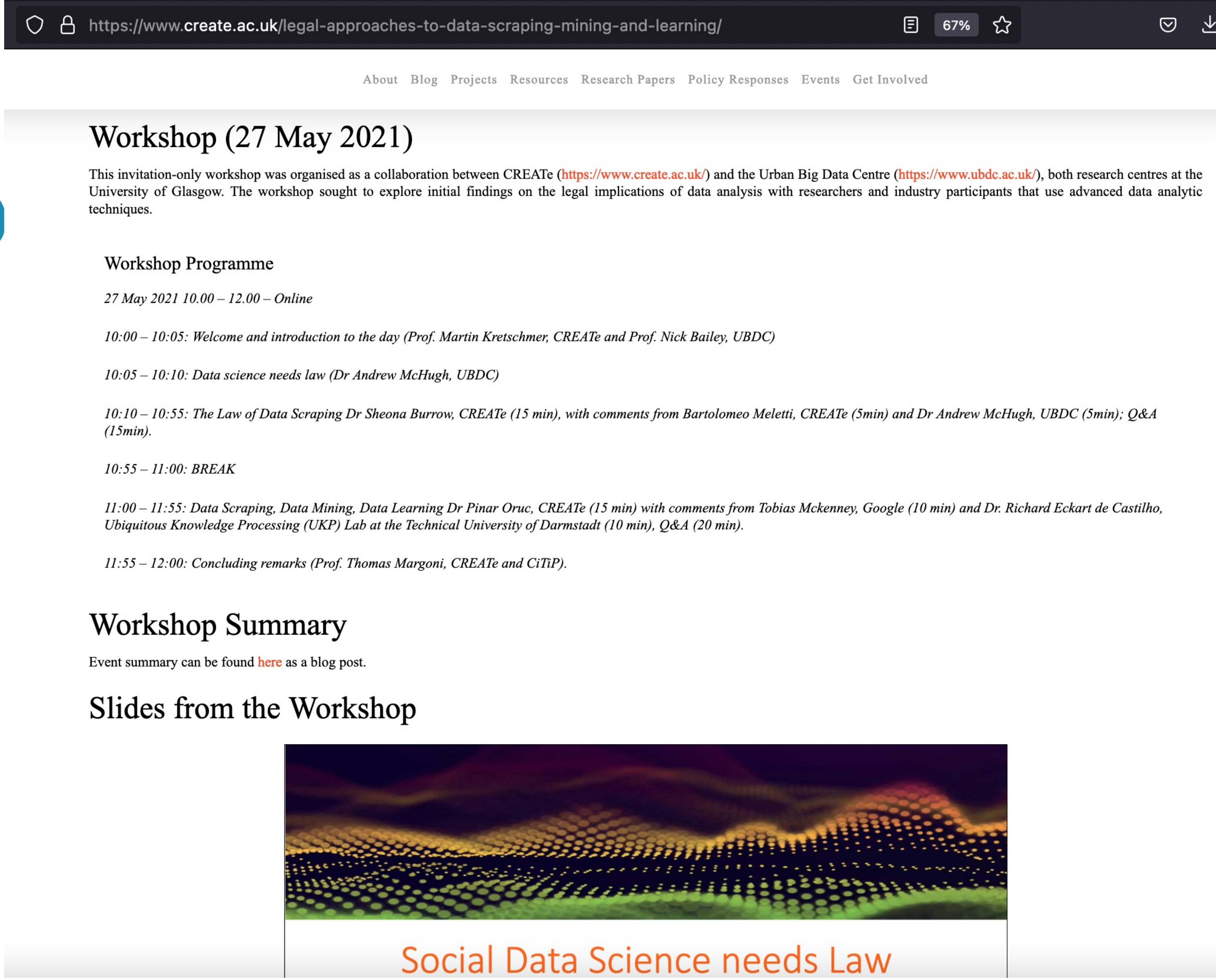
From a copyright perspective, NLP needs to be assessed for the type of data collected (protected or unprotected), the activities performed in the text analysis (copying, editing, annotating and using pre-trained language models) and the outputs in the

Case study 3: Computer vision, in the context of content moderation of images

This case study is focussed on computer vision. While there are many uses for computer vision, such as facial recognition or self-driving cars, this case study will focus on the example of using object recognition technology for content moderation of images. Computer vision involves the collection of images and videos (protected and unprotected). It is followed by their pre-processing, such as cropping, rotating or converting colour. Training can be supervised or unsupervised, both based on features of the images. If supervised, images will be annotated in full or partially. If unsupervised, the computer will detect similarities and classify images, but will be unable to interpret them. When used for content moderation, human moderators are still widely used for uncertain decisions in regard to the visual content with violence, nudity and criminal activity.

From a copyright perspective, computer vision needs to be assessed for type of data collected

Validation Workshop


https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/
67%
About
Blog
Projects
Resources
Research Papers
Policy Responses
Events
Get Involved

Workshop (27 May 2021)

This invitation-only workshop was organised as a collaboration between CREATe (https://www.create.ac.uk/) and the Urban Big Data Centre (https://www.ubdc.ac.uk/), both research centres at the University of Glasgow. The workshop sought to explore initial findings on the legal implications of data analysis with researchers and industry participants that use advanced data analytic techniques.

Workshop Programme

27 May 2021 10.00 – 12.00 – Online

10:00 – 10:05: Welcome and introduction to the day (Prof. Martin Kretschmer, CREATe and Prof. Nick Bailey, UBDC)

10:05 – 10:10: Data science needs law (Dr Andrew McHugh, UBDC)

10:10 – 10:55: The Law of Data Scraping Dr Sheona Burrow, CREATe (15 min), with comments from Bartolomeo Meletti, CREATe (5min) and Dr Andrew McHugh, UBDC (5min); Q&A (15min).

10:55 – 11:00: BREAK

11:00 – 11:55: Data Scraping, Data Mining, Data Learning Dr Pinar Oruc, CREATe (15 min) with comments from Tobias Mckenney, Google (10 min) and Dr. Richard Eckart de Castilho, Ubiquitous Knowledge Processing (UKP) Lab at the Technical University of Darmstadt (10 min), Q&A (20 min).

11:55 – 12:00: Concluding remarks (Prof. Thomas Margoni, CREATe and CiTiP).

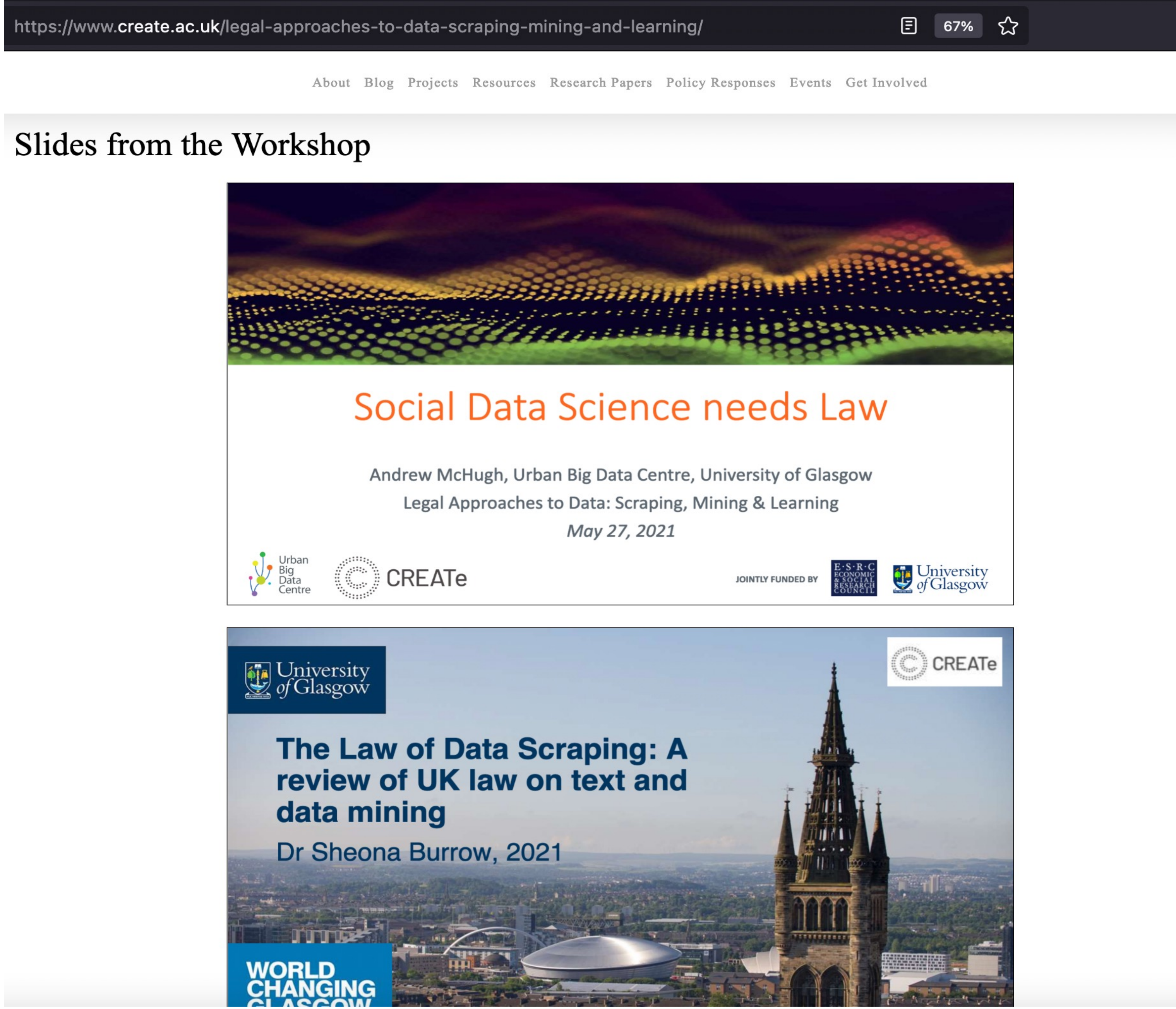
Workshop Summary

Event summary can be found here as a blog post.

Slides from the Workshop

Social Data Science needs Law

Examples



https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/

About Blog Projects Resources Research Papers Policy Responses Events Get Involved

Slides from the Workshop

Social Data Science needs Law

Andrew McHugh, Urban Big Data Centre, University of Glasgow
Legal Approaches to Data: Scraping, Mining & Learning
May 27, 2021

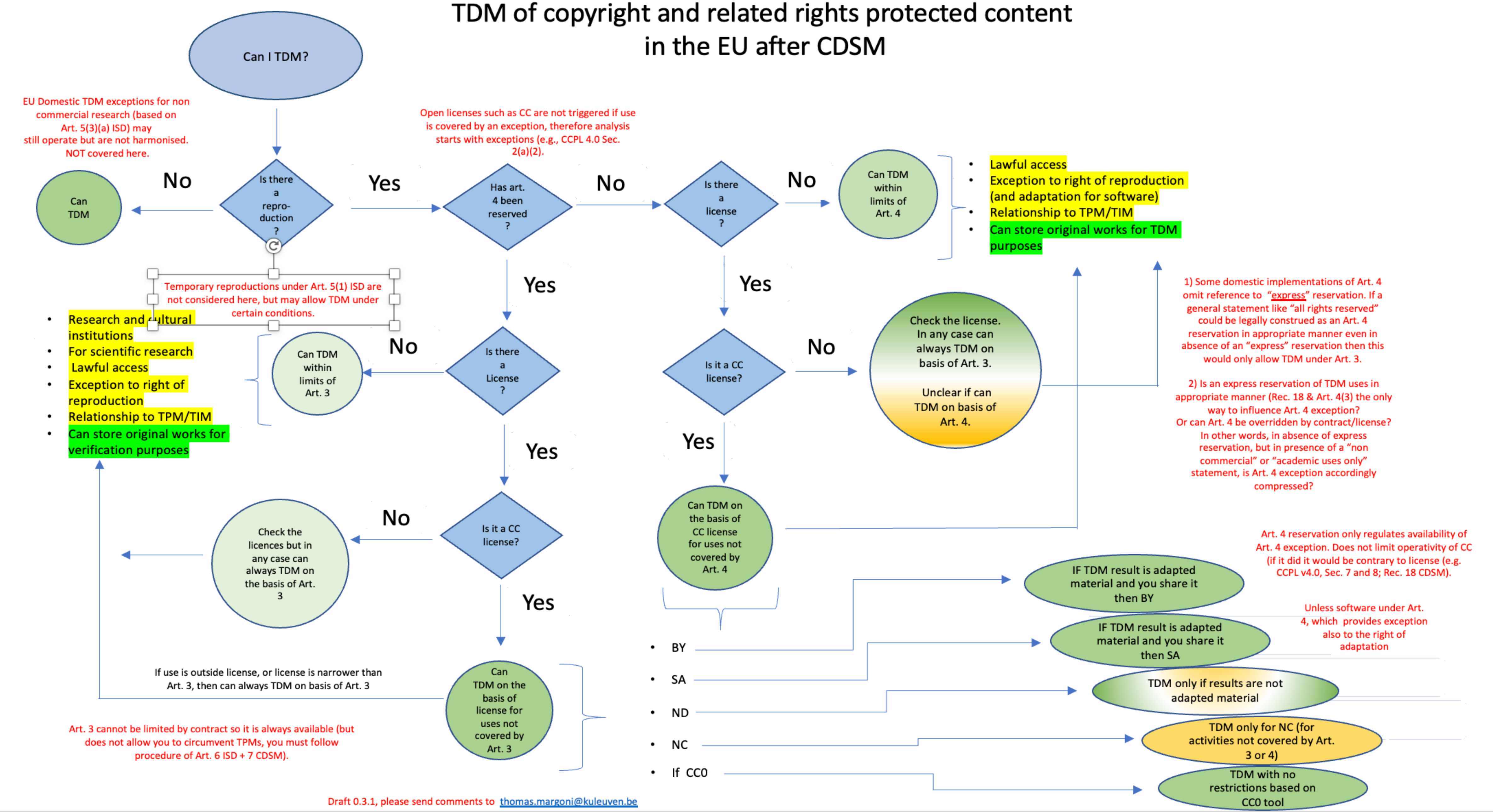
The Law of Data Scraping: A review of UK law on text and data mining

Dr Sheona Burrow, 2021

“A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology” (by Margoni&Kretschmer, GRUR)

1. Focus on right of reproduction (Art. 2 ISD), SGDR, exceptions (Arts. 3&4 CDSM, 5, esp. 5(1) ISD):
   - EU has low threshold to enjoy projection, high threshold to benefit from exemptions
2. Assess tensions in EU copyright law between “EU law” and “copyright law”:
   - Harmonisation EU law and regulation of new technologies not always integrated
3. Identifies a property-based approach to regulation of data (and thus AI):
   - EU AI development largely relies on 2,5 copyright exceptions. What costs/incentives does this create within and beyond single market (e.g. regulatory competition, UK, US, JP?)
4. Property-based in apparent tension with governance-based approaches:
   - The latter emerging in a more recent wave of legislation (e.g. DGA, DA, AIA, DSA/DMA, PSI/OD, Free Flow Reg., etc).

TDM of copyright and related rights protected content in the EU after CDSM



Can I TDM?

EU Domestic TDM exceptions for non commercial research (based on Art. 5(3)(a) ISD) may still operate but are not harmonised. NOT covered here.

Is there a repro-duction ?

No → Can TDM

Temporary reproductions under Art. 5(1) ISD are not considered here, but may allow TDM under certain conditions.

Yes → Has art. 4 been reserved ?

Open licenses such as CC are not triggered if use is covered by an exception, therefore analysis starts with exceptions (e.g., CCPL 4.0 Sec. 2(a)(2).

No → Is there a license ?

No → Can TDM within limits of Art. 4

- Lawful access
- Exception to right of reproduction (and adaptation for software)
- Relationship to TPM/TIM
- Can store original works for TDM purposes

Yes → Is it a CC license?

No → Check the license. In any case can always TDM on basis of Art. 3. Unclear if can TDM on basis of Art. 4.

Yes → Can TDM on the basis of CC license for uses not covered by Art 4

1) Some domestic implementations of Art. 4 omit reference to "express" reservation. If a general statement like "all rights reserved" could be legally construed as an Art. 4 reservation in appropriate manner even in absence of an "express" reservation then this would only allow TDM under Art. 3.

2) Is an express reservation of TDM uses in appropriate manner (Rec. 18 & Art. 4(3) the only way to influence Art. 4 exception? Or can Art. 4 be overridden by contract/license? In other words, in absence of express reservation, but in presence of a "non commercial" or "academic uses only" statement, is Art. 4 exception accordingly compressed?

Art. 4 reservation only regulates availability of Art. 4 exception. Does not limit operativity of CC (if it did it would be contrary to license (e.g. CCPL v4.0, Sec. 7 and 8; Rec. 18 CDSM).

- BY → IF TDM result is adapted material and you share it then BY
- SA → IF TDM result is adapted material and you share it then SA
- ND → TDM only if results are not adapted material
- NC → TDM only for NC (for activities not covered by Art. 3 or 4)
- If CC0 → TDM with no restrictions based on CC0 tool

Unless software under Art. 4, which provides exception also to the right of adaptation

Yes (Has art. 4 been reserved?) → Is there a License ?

No → Can TDM within limits of Art. 3

- Research and cultural institutions
- For scientific research
- Lawful access
- Exception to right of reproduction
- Relationship to TPM/TIM
- Can store original works for verification purposes

Yes → Is it a CC license?

No → Check the licences but in any case can always TDM on the basis of Art. 3

Yes → Can TDM on the basis of license for uses not covered by Art. 3

If use is outside license, or license is narrower than Art. 3, then can always TDM on basis of Art. 3

Art. 3 cannot be limited by contract so it is always available (but does not allow you to circumvent TPMs, you must follow procedure of Art. 6 ISD + 7 CDSM).

Draft 0.3.1, please send comments to thomas.margoni@kuleuven.be

General conclusions

- **Property-based approach to data** is problematic. AI applications based on machine learning and other data intensive approaches, i.e. where an algorithm needs to be trained on data, can only be developed based on a narrow (or wider but non imperative) copyright exception. Is this the intended function of copyright? To be the ultimate judge of whether, how and by whom technological development can happen and which direction should it take?
- **Property rights create issues of access** (authorization to use) and establish **conditions** (availability, price, purposes). Is the intended function of copyright to offer data holders control over data-based downstream markets such as AI development? What consequences may this frame lead to?
- **Access to data for AI in EU may be limited to** those:
  - Who are willing/can pay the price (**will EU AI be then more expansive/less competitive than US AI? Or Japan? CH? UK AI?**)
  - Train outside the EU in "cheaper" legal systems and use so trained AI in EU or import pretrained models: but what would be the impact in the EU to employ AI trained on a body of 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 hide the sources, leading to **opacity in the training process** (which would plausibly contrast with high-risk AI in AIA) – not a desirable mix of incentives for innovation.

General conclusions

- **Governance-based approaches** (such as PSI/OD; DGA; DA; DSA/DMA*) seem to offer an alternative model based on access, control, reusability and portability of non personal data (GDRPization of all data?)
- A topology of data:
  - High-value datasets (can be made available for free and easily re-usable across the entire EU). PSI/OD
  - PSB and certain public undertakings data, including research data (except certain categories, mainly third party IP/PeD); PSI/OD
  - For the latter excluded category, development of trustworthy data-sharing systems to facilitate voluntary and create incentives to share data; DGA
  - IoT data: specific rules on access and portability of co-created data and specific rules in relation to IP/PeD; Data Act
  - **Data spaces** as a mixed private-public regulatory framework for the development of a single market of data
- Property-based and governance-based approaches are developed in parallel, but may in fact lead to divergent solutions, which point towards the need to coordinate these two different approaches (perfect example is Art. 35 DA).

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 (Eds), **Data Act Blog Series**, 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