ReCreating Europe

Rethinking digital copyright law for a culturally diverse, accessible, creative Europe

Grant Agreement No. 870626

Deliverable Title	**D3.6 Interim study on the state of harmonisation of the rights of reproduction and adaptation and connected exceptions**
Deliverable Lead:	UGLA-CREATe
Partner(s) involved:	N/A
Related Work Package:	WP3 - Authors and performers
Related Task/Subtask:	T3.3 AI, machine learning and EU copyright law: ownership issues in training data
Main Author(s):	Martin Kretschmer (UGLA-CREATe), Thomas Margoni (UGLA-CREATe), Pinar Oruc (UGLA-CREATe)
Other Author(s):	N/A
Dissemination Level:	Public
Due Delivery Date:	30.06.2021
Actual Delivery:	29.06.2021
Project ID	870626
Instrument:	H2020-SC6-GOVERNANCE-2019
Start Date of Project:	01.01.2020
Duration:	36 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Version history table			
Version	**Date**	**Modification reason**	**Modifier(s)**
v.01	14.06.2021	First draft	Thomas Margoni, Martin Kretschmer
v.02	22.06.2021	Second draft	Pinar Oruc
v.03	26.06.2021	Accommodated suggested actions requested from the peer-reviewers; Final version released	Thomas Margoni

Legal Disclaimer

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 870626

Table of Contents

List of Figures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Abbreviation list

AI	Artificial Intelligence
API	Application Programming Interface
CDSM	Copyright in Digital Single Market (Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market
CJEU	Court of Justice of the European Union
CNN	Convolutional neural network
DMA	Digital Markets Act (Proposal for a Regulation on contestable and fair markets in the digital sector Brussels, 15.12.2020 COM(2020) 842 final)
GAN	Generative adversarial network
ISD	Information Society Directive (Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society)
ML	Machine learning
PSI	Public Sector Information
TDM	Text and Data Mining
TPM	Technological Protection Measures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Executive summary

There is global attention on new data analytic methods. Machine learning (essentially pattern recognition dressed as Artificial Intelligence or AI) is seen as a critical technology. Data scraping, the acquiring and structuring of information from online sources, is a typical first step for many advanced data analytic methods.

The technologies of scraping, mining and learning are often conflated, as are the legal regimes under which they are regulated. One regulatory lever under one legal regime will not deliver policy aims, such as innovation, personal dignity, Open Science, or the currently popular 'data sovereignty'. The legal issues involved in the governance of data range from proprietary approaches (copyright, database rights) to privacy and data protection.

In addition, there are a wide range of public law instruments, for example relating to public sector data governance[1] or the right to non-discrimination.[2] Competition law again (which may be both privately and publicly enforceable) increasingly prescribes conduct in relation to data, such as in merger or acquisition cases, or in transparency provisions (Art. 17 CDSM; and centrally in the proposed DMA and AI Regulation).

The scope of our enquiry in this report is within private law, specifically on the attempt to assert quasi-proprietary control of information and data, or vice versa limit such attempts, for example by exempting desired activities via copyright exceptions, such as the exception for text and data mining in Arts. 3 and 4 CDSM.

The copyright regime offers a template with a centuries old tradition of exclusive rights, supplemented in the EU since 1996 by a sui generis database right.[3] While data or information are not subject matter within copyright law, almost all materials used to construct so-called corpora for new data analytic methods are protected by copyright law: scientific papers, images, videos, and so on.

The research design we adopt for this interim report is a reverse inductive strategy. We focus on case studies of three technological processes to explore in detail possible descriptions that would allow legal analysis, and an assessment of the need for a harmonisation of rights and connected exceptions under copyright law.

The case studies were selected in consultation with stakeholders, reflecting a need by scientific researchers and technology companies for a better legal understanding of what they do. They are designed to reflect a range of techniques and processes that underpin advanced data analytics, responding to the EU policy objective of supporting innovation in this field.

The three case studies are:

(1) Data scraping for scientific purposes
(2) Machine learning, in the context of Natural Language Processing (NLP)
(3) Computer vision, in the context of content moderation of images

In parallel, we offer a thorough analysis of the policy rationale and legal context for the introduction of the two exceptions for text and data mining in the CDSM Directive (Art. 3 Text and data mining for the purposes of scientific research; Art. 4 Exception or limitation for text and data mining) which includes an analysis of how the right of reproduction (Art. 2 ISD) and its limitations (mainly Art. 5(1) ISD) interface with the overall

[1] Directive 2019/1024/EU of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information
[2] Charter of Fundamental Rights of the European Union Art.21, as reflected in the Proposal for a Regulation laying down harmonised rules on artificial intelligence
[3] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

regulatory framework of data analytics. This part of the Interim report is written as a self-contained scientific paper and is appended to this Report as Annex A.

> *We have created a resource page that will be regularly updated with project results (workshops and outputs): https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/*

1. Methodology

Legal research on data analytic techniques typically starts with an identification of relevant legal regimes and proceeds to doctrinal analysis of the scope of certain concepts and rules. The analysis is then evaluated against practical implications, often using particular factual constructions (scenarios) to illuminate potential effects of interpretations or interventions.

There are dangers to this legal approach to policy making. The analysis often lags technological developments. Scenarios may be filtered via professional representations or trade bodies that were constituted in a different context, perpetuating past discussion. In a wider sense, policy making may be anecdotally driven, by examples that surface through lobbying processes.

The research design adopted for this project aims to reverse this direction of travel. We adopt an inductive approach, attempting to get close to the "real world" of data analytics. Through a detailed empirical description of a selection of cases (in a social science sense) we seek to explore legal issues that are implicated.

The selection of sites for case analysis poses its own generalisability challenge. In case study research, we need to reflect on why selected empirical settings are more or less reflective of the phenomenon under investigation, i.e. rapidly evolving data analytic technologies.[4] In consultation with scientific researchers and technology companies, we identified three case studies that together reflect a range of techniques and processes that underpin advanced data analytics. The selection takes account of the EU policy objective to support innovation in this field.[5]

The three selected cases are:

(1) Data scraping for scientific purposes;
(2) Machine learning, in the context of Natural Language Processing (NLP);
(3) Computer vision, in the context of content moderation of images.

In researching the cases in a legal context, there is a further tension between an unstructured approach that offers rich descriptions inductively from multiple sources (such as public documents, observations, conferences, or interviews) and the need to capture the empirical world in a form recognisable for subsequent legal analysis. In Law, this challenge of "fact-finding" is discussed under the concept of evidence.[6] In legal disputes, there is an assumption that a representation of facts can be settled (typically in first instance cases). It is then the application of rules to the facts that can be the subject of appeals. The case studies presented in this interim report offer such a possible description of facts that will allow fuller legal analysis and policy recommendations in the final report due M28.

[4] For a classic account of the selection problem in case study research, see Seawright J. and Gerring J. (2008) Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options, Political Research Quarterly 61(2):294-308. doi:10.1177/1065912907313077

[5] Cf. Recital 8, CDSM Directive.

[6] Ho, Hock Lai (2015) The Legal Concept of Evidence, Stanford Encyclopedia of Philosophy, available at <https://plato.stanford.edu/entries/evidence-legal/> (accessed 1 June 2021)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

We prepare the ground for this second stage of the project with a self-contained paper that addresses the legal context of the introduction of the exceptions for text and data mining in the CDSM Directive (Art. 3 Text and data mining for the purposes of scientific research; Art. 4 Exception or limitation for text and data mining). What was the policy problem the interventions sought to solve? Which were the legal hurdles to the development of certain data analytic technologies in the EU that needed to be addressed? What are unresolved issues and shortcomings of the legal approach chosen in the CDSM Directive?

We highlight:

- that the definition of text and data mining may be too broad, making the entire field of data-driven AI development dependent on exceptions;
- that the scope of the exceptions is limited to the right of reproduction;
- that the limitation of the Art. 3 to certain beneficiaries remains problematic;
- that the requirement of lawful access is difficult to operationalize;
- that rightholders *de facto* can override the exceptions by technological interventions.

2. Cases for the study of copyright and *training data* in selected AI environments

These case studies have been prepared by Dr Pinar Oruc, under the supervision of Profs. Martin Kretschmer and Thomas Margoni.

2.1 Introduction to the case studies and delimitation of the area of enquiry

Copyright law has a direct impact on the processes of data scraping, mining and learning.[7] So called "corpora", i.e. collection of information needed for "training" purposes could include works protected by copyright, other related subject matter, or simple facts and data. When copyright or a related right are present, any digital copy, temporary or permanent, in whole or in part, direct or indirect, has the potential to infringe that right, in particular the economic right of reproduction. Furthermore, the changes made in the collected material can amount to an 'adaptation' within the scope of the exclusive right. The relevant exceptions, such as for research or text and data mining, might not sufficiently cover the activities of the researchers and firms in this area. This report presents three case studies to provide an in-depth exploration of the complex technological processes involved in some of the most popular AI applications. The results of the case analysis will be functional to a proper legal classification and assessment of the relevant regulatory framework.

Three different case settings were selected: web scraping, natural language processing, computer vision. The case studies rely on publicly available sources (e.g. published scholarly analysis, official information issued by companies for instance on their websites and policy documents) and expert feedback.[8] Within the cases, the unit of analysis for comparison is the technological process.

2.1.1 Terminology

This part introduces the technological distinctions that will be employed in the report.
The first distinction relates to data collection methods. "Scraping" involves manually or automatically collecting data from websites. There are different kinds of such data collection, such as web scraping, web harvesting and web crawling, which will be addressed specifically under Case Study 1.

[7] For an analysis of the link between EU copyright law and machine learning form an input data point of view, see Margoni, Thomas (2018), Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?, AIDA (Annali Italiani del Diritto d'Autore della Cultura e dello Spettacolo); 2018; Vol. XXVII; iss. 1; pp. 281 – 304 available at: https://zenodo.org/record/2001763.
[8] As part of the validation process for the case studies, a workshop was held at the University of Glasgow on 27 May 2021. It is documented here:<https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/>

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

It is useful to clarify that data scraping and data mining are not the same, but the terms are sometimes used interchangeably in the literature. Data scraping is the collection/extraction of the necessary information to build a set of data. Data mining, which does not necessarily include data collection, is the analysis of these datasets and sometimes this analysis requires machine learning to reach more complex purposes. Case Study 1 uses the term 'scraping' but will also touch on the stages for mining and outputs.

A second distinction relates to the type of machine learning, addressed in Case Studies 2 and 3. Supervised learning uses training data labelled by humans. In unsupervised learning, the algorithm uses unlabelled data and detects similarities and patterns. In reinforcement learning, the algorithm relies on trial and error to reach the maximum reward for an activity.

There are also more specific technological developments that might be relevant across the case studies. Deep learning is a form of machine learning (ML), where multiple artificial neural networks[9] carry and interpret complex raw data. With more layers, it becomes more likely to solve complex problems but it also means less clarity on why the AI system decides one way over the other, which reduces the accountability. Although neural networks have been proposed as early as 1943, the research on neural networks and the use of deep neural networks have increased in recent years with the availability of cheaper computational power and resources.[10]

Generative adversarial networks (GAN) have two deep learning networks (one generator and one discriminator) and they learn by competing with each other.[11] GANs can be supervised and unsupervised.
There is also "transfer learning", which is not a ML technique but a way to design a research methodology where there is not enough training data. A pre-trained model for a similar task is taken and adopted into the project at hand.[12] It is used for both NLP and Computer Vision.

How to categorise the stages in the process? When categorising these activities for the purpose of our legal analysis, it becomes apparent that the stages are not the same for data analysis and ML training.

For scraping, our legal analysis would require focusing on the data collection and data processing stages. There is no annotation or training, but there are outputs based on data analysis. (3 stages)
For machine learning projects, researchers usually start with defining the problem, choosing the data sources and algorithms, and the trained model will then be released (deployment stage).[13] For the stages in between, both Natural Language Processing (NLP) and Computer Vision stages are to be categorised similarly: data collection (which can also be achieved through scraping), data processing, training (different process if it is supervised or unsupervised) and then the output stage – depending on what the algorithm is for, such as language understanding or audiovisual content moderation. (4 stages)

Accordingly, data scraping as such may be likewise seen as a form of data collection that then leads to different possibilities of data analytics, including those identified in Case Studies 2 and 3, therefore as a

[9] "The network is a connected framework of many functions (neurons) working together to process multiple data inputs. The network is generally organized in successive layers of functions, each layer using the output of the previous one as an input." WIPO Technology Trends 2019 — Artificial Intelligence, WIPO, 2019, <https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf> 146; See also OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, Adopted on 22/05/2019.
[10] Seifert et al, Visualizations of Deep Neural Networks in Computer Vision: A Survey' in Tania Cerquitelli, Daniele Quercia and Frank Pasquale. *Transparent Data Mining in Big and Small Data* (Springer 2017) 123.
[11] Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity* (MIT Press 2019) Chapter 10.
[12] Niklas Donge, 'What is Transfer Learning? Exploring the Popular Deep Learning Approach' (2020) https://builtin.com/data-science/transfer-learning , Orhan G Yalcin, '4 Pre-Trained CNN Models to Use for Computer Vision with Transfer Learning' (2020) https://towardsdatascience.com/4-pre-trained-cnn-models-to-use-for-computer-vision-with-transfer-learning-885cb1b2dfc.
[13] Richmond Alake, '10 Stages Of A Machine Learning Project In 2020 (And Where You Fit)' (2020) https://towardsdatascience.com/10-stages-of-a-machine-learning-project-in-2020-and-where-you-fit-cb73ad4726cb.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

preliminary step for Natural Language Processing and Object Recognition. However, given its relevance and complexity we decided to offer it as a self-standing case, seeking feedback during expert consultations.

2.2. Data Scraping

Scraping involves manually or automatically collecting data from websites. Screen scraping involves scraping the data that is displayed on users' screens. Web-scraping or web-harvesting is collecting all underlying data from a website, including website scripts. Web-crawling is "accessing web content and indexing it via hyperlinks; thus, only the URL but no specific information is extracted".[14]

There are multiple ways of categorising scraping tasks: (1) accessing the web pages, (2) finding specified data elements, (3) extracting them, (4) transforming them and (5) saving these as a structured data set.[15] Alternatively, they can also be further divided: '(1) information identification, (2) choice of strategy, (3) data retrieval, (4) information extraction, (5) data preparation, (6) data validation, (7) debugging and maintenance, (8) generalisation'.[16]

For our purposes, we will merge some stages that are similar for legal analysis: (1) data collection, (2) data processing and, (3) data analysis and outputs.

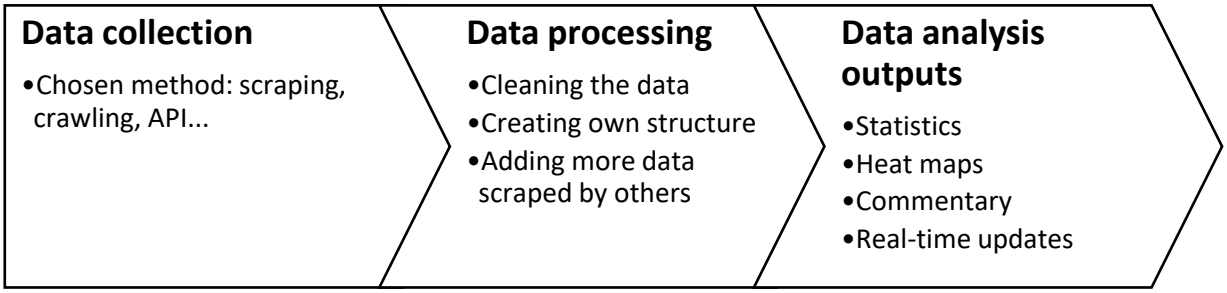

Figure 1: Data scraping workflow

This part will rely on the example of a property website (such as AirBnB) being scraped for research on short/long-term letting market effects.

2.2.1. Collection stage

Most property listing websites such as AirBnB do not create a new page for every listing. Instead, a template exists, and it is automatically filled with data for that specific property as entered by the users/property hosts. The data available on the website includes property descriptions, user reviews, photographs of the property (only saved as hyperlinks), location, longitude and latitude of the property, neighbourhood ID, available dates, maximum and minimum price, place type and number of guests, user scores.

If the collection purpose is unknown, it is often useful and possible to collect all available information. At this stage, no distinction may be made whether particular data is created by AirBnB or uploaded by the property hosts.

[14] Fiona Campbell, 'Data scraping - what are the privacy implications?' (2019) Privacy & Data Protection 20(1), Frank Jennings and John Yates, 'Scrapping over data: are the data scrapers' days numbered?' (2009) JIPLP 4(2) 120, Judith Hillen, 'Web scraping for food price research' (2019) British Food Journal 121(2) 3350.

[15] Geoff Boeing and Paul Waddell, 'New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings' (2017) Journal of Planning Education and Research 37(1) 457, 459.

[16] Tasks are identified as '(1) information identification, (2) choice of strategy, (3) data retrieval, (4) information extraction, (5) data preparation, (6) data validation, (7) debugging and maintenance, (8) generalisation' Simon Munzert, Christian Rubba, Peter Meißner and Dominic Nyhuis. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (John Wiley & Sons 2015).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

As introduced before, screen scraping is limited to what is available to the visitors, web-harvesting targets and collects all data, and web-crawling follows and indexes all links (those can be visited and scraped). Since scraping relies on how data is displayed, even the small changes in the display of the website can disrupt the collection stage.[17]

Another method of making the process faster and more efficient is using the application programming interface (API) scraping. The stages of using an API for data collection can be summarised as follows: (i) finding the API and becoming familiar, (ii) registering for API use and retrieving keys, (iii) calling the API to collect data and then (iv) processing the data.[18] API scraping does not mean access to previously inaccessible data, but it speeds up the process by circumventing the rendering stage.

While it makes it easier for scrapers, APIs require substantial resources for hosts to develop and maintain'.[19] In fact, some websites do not make their API openly available to stop the competitors from scraping data from them in order to ensure their competitive edge remains.[20] Although not directly competitors, any researchers interested in this data and unable to collect it via the API, then have to come up with their own strategies and/or rely on different scraping strategies.

In the example of AirBnB, their API is not openly available to the general public, but may be requested by developers, certain groups of users, such as hosts wanting to use their own interface to add multiple listings at once[21] or external partners such as travel companies and Groupon.[22]

In addition to the concerns about loss of control over the data and its devaluation reported by website operators, they have to make sure that the scraping does not cause system overload.[23] Examples in this sense are blocks of excessive requests from the same IP range to ensure stability of the servers. Since data hosts can detect unusually high or repetitive tasks from the same IP address (even easier to detect if it comes from the same user account, if the scraping is performed after the login page), the scrapers usually use proxies to distribute their requests to avoid exceeding this threshold and being blocked.[24]

Additionally, as common practice many websites have terms and conditions that restrict the collection and analysis of their data. Under the AirBnB Terms of Service (both for European Users and non-European Users), there are terms that limit the ways and purposes of using the platform. Under 12.1 of their Terms of Service, the following is not allowed: "scraping, hacking, reverse engineering, compromising or impairing the platform, using bots, crawlers, scrapers or other automated means, attempts to circumvent any security or technological measure, taking any action that could damage or adversely affect the performance or proper functioning of the platform". Furthermore, the Content cannot be used without the permission of Content owner and can only be used as necessary to enable to use of the website as a Guest or Host.[25]

---

[17] Jeffrey Hirschey 'Symbiotic Relationships: Pragmatic Acceptance of Data Scraping' (2014) Berkeley Technology Law Journal, Vol. 29, 906

[18] Simon Munzert, Christian Rubba, Peter Meißner and Dominic Nyhuis. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (John Wiley & Sons 2015) 277

[19] Brett Massimino, 'Accessing Online Data: Web-Crawling and Information-Scraping Techniques to Automate the Assembly of Research Data' (2016) Journal of Business Logistics 37(1) 34.

[20] Janet Williams, 'Web Scraping better than a Data API' https://www.promptcloud.com/blog/web-scraping-better-alternative-to-api/

[21] https://www.airbnb.co.uk/partner

[22] Ingrid Lunden, 'Airbnb eyes expansion with affiliate program for sites with 1M+ users, new API' (2017) https://techcrunch.com/2017/10/16/airbnb-eyes-expansion-with-affiliate-program-for-sites-with-1m-users-new-api/

[23] Frank Jennings and John Yates, 'Scrapping over data: are the data scrapers' days numbered?' (2009) JIPLP 4(2) 120

[24] Jeffrey Hirschey 'Symbiotic Relationships: Pragmatic Acceptance of Data Scraping' (2014) Berkeley Technology Law Journal, Vol. 29, 918; Manthan Koolwal, '10 Tips to avoid getting Blocked while Scraping Websites' (2020) https://www.codementor.io/@scrapingdog/10-tips-to-avoid-getting-blocked-while-scraping-websites-16papipe62

[25] https://www.airbnb.co.uk/help/article/2908/terms-of-service#EU12

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

2.2.2. Processing stage

After the targeted data is collected, it is then structured in a manner that is more suitable for the upcoming data analysis. Researchers typically store this raw data in a way that is structured by them, that is more in line with their research purposes and internal structure. As the computational power and storage costs are constantly getting more effective, it is suggested that scrapers are now able to scrape more data and can choose to be less conservative.[26] But that also means holding more data to be filtered and cleaned.

As the property information in this example are added by the users, it can be messy and the researchers might have to go through substantial wrangling and validation to make the data usable.[27] It requires identifying and removing duplicate listings (by relying on things as the Property ID, location and the size of the property) or identifying other mistakes such as typos in the rental price.[28] As part of validation, the researchers have to ensure that the new data is reliable and usable for their purposes. Depending on the purpose of each research output, the necessary data is then pulled from these databases.

It is also possible to enrich the scraped data with the data from other sources. For example, there are websites and analytics companies based in the United States that collect and aggregate AirBnB data, such as AirDNA and SmartHost, to guide the hosts and nearby businesses. These are not prevented to do so by AirBnB so far. There are also US sources that provide scraped data together with own analysis, such as Tom Slee (tomslee.net) and InsideAirBnB (insideairbnb.com).[29] Researchers, in both inside and outside United States, often rely on such scraped datasets, commentary and research outputs by such third parties.

2.2.3 Analysis and outputs stage

The collected data can be one-off and shows the exact situation at a certain time of it can allow real-time updates (such as price comparison websites).[30] It is up to the researcher to choose which data will be collected and analysed, to solve the problem at hand.

The results of the analysis are then shared in formats chosen by the researcher (such as reports, journal articles, heat maps or blog posts). The extent of the data used in these outputs is determined case by case. These outputs are not a replacement of the website; however they can convince policymakers about changes that indirectly affect websites like AirBnB.

Restructured datasets based on the scraped data may or may not be shared with other researchers. Parties might contact AirBnB for permission.

There is a growing body of academic literature based on AirBnB. A wide range of issues are addressed, such as the extent to which neighbourhoods are vulnerable to the switch from long-term letting to short-term letting.

[26] Judith Hillen, 'Web scraping for food price research' (2019) British Food Journal 121(2) 3350, 3354, Zachary Gold and Mark Latonero, 'Robots Welcome: Ethical and Legal Considerations for Web Crawling and Scraping' (2018) 13 Wash J L Tech & Arts 275, 281.
[27] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques* (2012) The Morgan Kaufmann Series in Data Management System, 592.
[28] Geoff Boeing and Paul Waddell, 'New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings' (2017) Journal of Planning Education and Research 37(1) 457, 460.
[29] "Other companies include, but are not limited to, Beyond Pricing (beyondpricing.com), SmartHost (smarthost.co.uk), Everbooked (www.everbooked.com) and PriceLabs (www.pricelabs.co)". Teresa Scassa, 'Ownership and control over publicly accessible platform data' (2019) Online Information Review 43(6) 986, 991.
[30] "Once the script is written, it is up to the user whether it should run and extract prices and other data monthly, weekly, daily or even at a higher frequency" Judith Hillen, 'Web scraping for food price research' (2019) British Food Journal 121(2) 3350, 3353.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

2.3 Natural Language Processing

Natural language processing (NLP) is in 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. The applications of NLP include information extraction, machine translation, natural language generation and sentiment analysis.[31]

NLP can be supervised or unsupervised. Supervised learning requires labelled/tagged text data, so they have an "annotation" stage in their workflow. On the other hand, unsupervised NLP uses unlabelled data and instead detects patterns, but it requires large datasets to achieve that and is therefore not suitable for all research projects. If some labels are from humans and others are not, then it will be classified as semi-supervised machine learning – which is useful for projects holding small annotated datasets together with large amount of raw data found online.[32]

NLP research focuses on achieving and improving various tasks.[33] Some tasks have direct applications, such as translation or summarisation. Other tasks such as segmentation or named entity recognition are used to inform other tasks and turn the texts into machine-readable data.

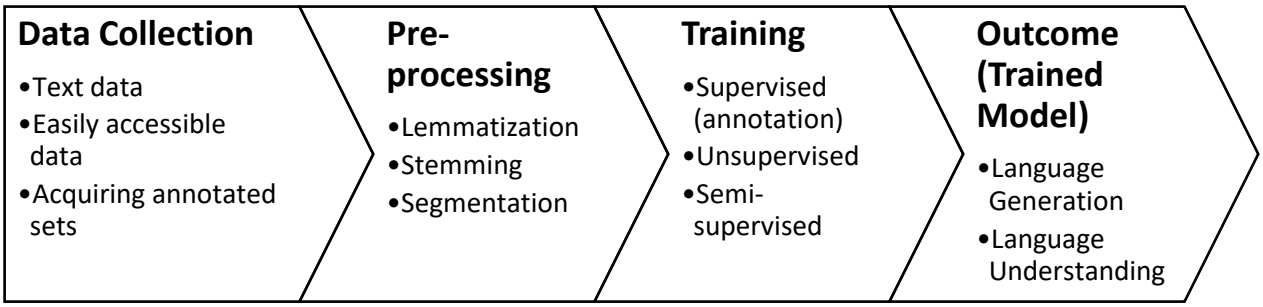

Figure 2: NLP workflow

2.3.1 Collection Stage

The first step for Natural Language Processing is the compilation of the necessary data. The data can come from anywhere, ranging from user comments to ancient philosopher corpus.
The data collection stage is similar to the scraping process described above: the necessary data is identified in line with the research purpose and then it will be targeted with the appropriate data collection methods (scraping or otherwise).
It is also possible to find freely available datasets online, such as the books from Project Gutenberg or the Spoken Wikipedia Corpora – depending on the task at hand.[34] The NLP researcher can also choose to focus on licensed corpora[35] or scholarly literature held in databases that they have access to.[36]
While it can be possible to build the models in a way to reduce access to data and keep it as temporary as possible, it would both be impractical and very straining on the resources of third parties.

[31] WIPO Technology Trends 2019 — Artificial Intelligence, WIPO, 2019, https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf
[32] Ben Dickson, 'What is semi-supervised machine learning?' (2021) https://bdtechtalks.com/2021/01/04/semi-supervised-machine-learning/
[33] For an overview of different tasks in NLP: https://paperswithcode.com/area/natural-language-processing
[34] ODSC - Open Data Science, '20 Open Datasets for Natural Language Processing' (2019) https://medium.com/@ODSC/20-open-datasets-for-natural-language-processing-538fbfaf8e38 ; Jason Brownlee, 'Datasets for Natural Language Processing' (2020) https://machinelearningmastery.com/datasets-natural-language-processing
[35] Eckart de Castilho et al, 'A Legal Perspective on Training Models for Natural Language Processing' (2018)
[36] Przybyła et al, 'Text mining resources for the life sciences' (2016) Database, Volume 2016, 2016, baw145

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

2.3.2. Pre-processing

The data then goes through pre-processing. This part involves different tasks to understand the texts.[37] The collected material goes through some changes at this stage, which will be important for the legal analysis later.

First, formats such as PDF of MS Word need to be converted into text for the NLP tasks that follow.[38]

Tokenization is when text is separated into smaller units in a way that can be read by the machine. This smaller unit can be word pieces or characters.[39] Parts of speech (POS) tagging is when words are tagged as noun, verb, or prepositions.

Normalization is when a more normalized version of the text is created by removing variations that are not important for the final research target. Through normalization, the text becomes more standard and is easier for the machines to "read". It includes tasks such as lemmatization, stemming or spelling correction, which all change the text.[40] Stemming removes the end of the word, while lemmatization changes the word into its base or dictionary form.[41] Such tasks are sometimes performed by an algorithm, but humans can be consulted as well, at least while developing these methods or applying it to a new application domains.

2.3.3. Training

As mentioned earlier, the stage after pre-processing then differ according to the type of the learning.

(a) Supervised:
If the project relies on supervised learning, then pre-processed data is annotated by humans and the human input then helps the development of AI. The data that was previously unreadable to the machine becomes something usable through the annotation stage.
During the annotation process, it is possible to both add the annotations to the original text or create a separate file for annotations.[42] The former has the advantage of keeping both the text and annotations in a single file – such as XML file – so the NLP algorithms have access to both.

(b) Unsupervised:
If unsupervised, then learning requires no human input once the data is collected. There is no annotation stage. The project could involve multiple tasks that support each other by creating annotations, but as long as the NLP rely only on pre-trained models and the final task does not involve humans, it would still count as unsupervised training.
Although unsupervised learning is possible and is a growing field in NLP, it is also not widely accessible to smaller groups due to the need for computer power and large amount of data. Companies that have such resources, such as Google or OpenAI, use it to create pre-trained models. As long as they make these models available, smaller projects can then obtain these pre-trained models and use it on their datasets.

---

[37] Mirantha Jayathilaka, '25 NLP tasks at a glance' (2020) https://medium.com/@miranthaj/25-nlp-tasks-at-a-glance-52e3fdff32e2; '.Natural Language Processing (NLP) Guide – What Is NLP & How Does it Work?' https://monkeylearn.com/natural-language-processing/

[38] Bruce H Cottman, 'Converting PDF and Gutenberg Document Formats into Text: Natural Language Processing in Production' (2020) https://towardsdatascience.com/natural-language-processing-in-production-converting-pdf-and-gutenberg-document-formats-into-text-9e7cd3046b33

[39] 'Tokenization' <https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html>

[40] Dan Jurafsky and James H. Martin, *Speech and Language Processing* (3rd edn, 2020) https://web.stanford.edu/~jurafsky/slp3/; Tiago Duque, 'Text Normalization' (2020) https://towardsdatascience.com/text-normalization-7ecc8e084e31

[41] 'Stemming and lemmatization' https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[42] Przybyła et al, 'Text mining resources for the life sciences' (2016) Database, Volume 2016, 2016, baw145.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Use of embeddings and language models

Pre-trained embeddings and models mentioned here are trained on a large corpus in an unsupervised manner (by Google, OpenAI and similar companies with such resources), then fine-tuned in a supervised manner.[43] These are then made available for other users, so that they can use them to support their other supervised and semi-supervised learning projects. This means that as long as these pre-trained versions are available, other researchers can skip some stages or reap the benefits of the collection and pre-processing done by other companies. But this also creates a monopoly over language modelling.[44]

The paragraphs below will explain where embeddings and models sit within the developments of NLP. It is useful to take such developments into consideration for our legal analysis, as the approaches determine the amount and type of data that is used and the parties' involvement.

- In earlier NLP projects, 'bag of words' approach assigns a unique token to words, so a text can be displayed in numbers. While transforming words to numerical representations (vectors), the basic method is to count how many times a word occurs in a text, without paying attention to the order of the words. Since this approach would determine words like "the" or "is" as the most common and therefore the most important, the weights of the words need a separate adjustment (TF-IDF encoding).[45] N-grams extracts a consecutive n-number of words from the text to analyse.[46] These methods are still used, but are now supported by the others below.
- Word embeddings (2013 onwards): Embedding models mean giving vectors that show the connection between words. This allows the machines to understand which words go together, which helps in tasks like prediction or translation. There are word embedding models like *word2vec* (by Google) and *GloVe* (by Stanford).

The researchers then have the option of either (i) relying on in pre-trained word embeddings (based on the training done by their developers) such as *word2vec* trained on Google News corpus[47] or (ii) train the embeddings themselves to make sure that it assigns numerical values based on their specific dataset/research topic - so that it can be used on later NLP tasks with greater accuracy.

Since the first option is trained on generic texts, they are not overly helpful for using on very specialist texts, for example legal documents.[48] This means that researchers of specific topics still might prefer to train their own word-embedding models with their own training data.

The fact that pre-trained embeddings rely on easily found text material also leads to bias problems. For example, it was determined that *word2vec* carries the same the sexist biases present in the news corpora it

[43] Edward Ma, 'Combining supervised learning and unsupervised learning to improve word vectors: Introduction to Generative Pre-Training' (2019) 'https://towardsdatascience.com/combining-supervised-learning-and-unsupervised-learning-to-improve-word-vectors-d4dea84ec36b ; OpenAI, ' Improving Language Understanding with Unsupervised Learning' (2018) https://openai.com/blog/language-unsupervised/

[44] Taylor Soper, ''OpenAI should be renamed ClosedAI': Reaction to Microsoft's exclusive license of OpenAI's GPT-3' https://www.geekwire.com/2020/openai-renamed-closedai-reaction-microsofts-exclusive-license-openais-gpt-3/

[45] Rostyslav Neskorozhenyi, 'Word embeddings in 2020. Review with code examples' (2020) https://towardsdatascience.com/word-embeddings-in-2020-review-with-code-examples-11eb39a1ee6d , Antonio Lopardo, Word2Vec to Transformers' <https://towardsdatascience.com/word2vec-to-transformers-caf5a3daa08a>

[46] Timothy Tan, 'Evolution of Language Models: N-Grams, Word Embeddings, Attention & Transformers' (2020) https://towardsdatascience.com/evolution-of-language-models-n-grams-word-embeddings-attention-transformers-a688151825d2

[47] https://code.google.com/archive/p/word2vec/

[48] Ilias Chalkidis and Dimitrios Kampas 'Deep learning in law: early adaptation and legal word embeddings trained on large corpora' (2019) Artificial Intelligence and Law 27, 171, 174.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

was trained on.[49] But since the researchers can only view the trained *word2vec*, and not the news corpus it was trained on, it is also hard to pinpoint the reasons of this bias or make it less biased.[50]

- Language models (2018 onwards): Most recent ones rely on deep learning. They also excel in analysing the whole document, but here the 'vectors' are dynamic and adapt to the context. This means that transformers will be better at understanding the difference when same word is used in different context.[51]

These rely on deep neural networks, which are better at detecting and predicting 'complicated linguistic structures along with their long-distance relationships, as humans do'.[52] Another difference of transformers is that they can process words "in parallel", instead of "sequentially one by one" like the former methods, which makes it faster at going through large amounts of data.[53]

Transformer models found online are also trained on unlabelled data, for example Google's BERT trained on Wikipedia and Brown Corpus.[54] They can then be tweaked for the task/corpus at hand in other projects. One of the drawbacks is they do not exist for all languages, so not all researchers will have the same advantage. Additionally, the pre-trained versions might still require some fine-tuning. So, they might not be sufficient on their own, but they can make the other smaller projects easier.

2.3.4. Trained Model

The final stage is the creation of the trained model (a permanent file). Once the researchers have a trained model, they can use it on previously unseen datasets or use it to inform and support other larger tasks. What the trained model achieves depends on what task it was trained for. As mentioned above, some tasks have direct applications, while the others mainly help other NLP tasks.

Algorithms developed for Natural Language Understanding aim to determine the meaning of a sentence. Through syntactic and semantic analysis, AI applications manage to "read" the text. Document classification, sentiment analysis or named entity recognition are some of the examples such "understanding" tasks. Algorithms that "write" or "speak" are for Natural Language Generation.[55] For example, machine translations or chat bots answering questions achieve both understanding and generation through the multiple NLP tasks.

As a final note for the trained model, both for NLP and Computer Vision, it is not possible to remove some of the data after the model is trained. So, if a small part of the data needs to be removed (due to copyright or another reason, following an injunction), then the whole model needs to be retrained from the beginning.

[49] Tommaso Buonocore, 'Man is to Doctor as Woman is to Nurse: The Gender Bias of Word Embeddings' (2019) https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17

[50] Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2017) 93 Wash. L. Rev. 579, 582-583.

[51] Lavanya Gupta, 'Differences between Word2Vec and BERT' (2020) https://medium.com/swlh/differences-between-word2vec-and-bert-c08a3326b5d1

[52] Ilias Chalkidis and Dimitrios Kampas 'Deep learning in law: early adaptation and legal word embeddings trained on large corpora' (2019) Artificial Intelligence and Law 27, 171.

[53] Rostyslav Neskorozhenyi, 'Word embeddings in 2020. Review with code examples' https://towardsdatascience.com/word-embeddings-in-2020-review-with-code-examples-11eb39a1ee6d

[54] Sejuti Das, 'Top 8 Pre-Trained NLP Models Developers Must Know' (2020) https://analyticsindiamag.com/top-8-pre-trained-nlp-models-developers-must-know/

[55] Eda Kavlakoglu, 'NLP vs. NLU vs. NLG: the differences between three natural language processing concepts' https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

2.4. Computer Vision

The third case study will focus on computer vision. The developments in this field have been largely driven by industry uses, such as facial recognition or self-driving cars.[56] The discussion here will rely on the example of using object recognition technology for content moderation.

In supervised learning, AI is trained with annotated datasets and also gets human feedback when it wrongly classifies something based on the features. In unsupervised learning, AI learns by looking at the different images and recognising the similarities, like the way humans do by observation.[57]

As mentioned earlier, the use of deep neural networks has developed together with the increase in computing power. Although they can be used in language processing (as illustrated earlier), their earliest application was in the field of computer vision.[58] An example of using deep learning is the use of generative adversarial networks (GAN) in creating art. In this unsupervised form of learning, the generator continuously tests the discriminator with a realistic works, that are not very different from what is currently perceived as art. In addition to requiring large datasets of images of paintings,[59] such practices lead to questions about the copyright status of the AI-created works, that is outside the scope of this paper.

Although computer vision tasks vary widely, the process also starts with the collection of input data, followed by the processing of the data (which are different from NLP pre-processing tasks), followed by the training and leading to the outputs (which could range from a simple yes/no classification decision to a detailed, AI-generated response).

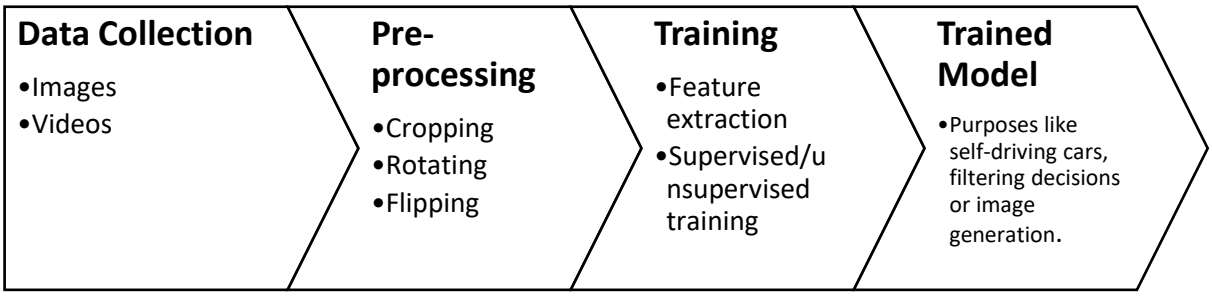

Figure 3: Computer vision workflow

2.4.1. Data Collection

The images or videos can come from various sources, such as phone cameras or medical devices. When training a computer vision model, it is important to use a dataset that is similar to the data it will be used for.[60]
For common objects, there are open datasets of labelled images online.[61] As one of the earlier projects of computer vision, ImageNet was launched in 2007 and holds over 14 million images labelled by participants.[62]

[56]Taylor Arnold, Lauren Tilton and Annie Berke, 'Visual Style in Two Network Era Sitcoms' (2019) Journal of Cultural Analytics.
[57] Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity* (MIT Press 2019) Chapter 10.
[58] Seifert et al, Visualizations of Deep Neural Networks in Computer Vision: A Survey' in Tania Cerquitelli, Daniele Quercia and Frank Pasquale. *Transparent Data Mining in Big and Small Data* (Springer 2017) 125.
[59] 75753 paintings were used to train the Generative Adversarial Network in the project where creative adversarial networks were proposed for the first time: Elgammal et al, 'CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms' (2017). See also Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity* (MIT Press 2019) 113-118.
[60] Lila Mullany, 'Introduction to Computer Vision Model Training' (2020) < https://towardsdatascience.com/introduction-to-computer-vision-model-training-c8d22a9af22b>
[61] Meiryum Ali, '20 Free Image Datasets for Computer Vision' https://lionbridge.ai/datasets/20-best-image-datasets-for-computer-vision/
[62] http://www.image-net.org/about

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

But easily accessible datasets are not sufficient for very specific research problems[63] and does not give any competitive edge if everyone trains their AI systems with the same images.[64]

Another option is using own image data or even a digitally generated dataset (synthetic data).[65] If the collected data is too small, it can be augmented (see below).

2.4.2. Pre-Processing

Once the data is collected, the images or videos go through pre-processing tasks, which are relevant for the legal analysis.

One of the tasks in pre-processing is the resizing of the image, so that all images in the dataset are the same size. Converting colour images to grayscale reduces the computation complexity, for research problems where the colour does not matter.[66]

Another task is noise reduction where the background features are smoothed and removed, so that the machine can focus on a single feature.[67]
One way to increase the dataset and preparing the AI application for recognising the same objects in different environments is data augmentation. This can be achieved by rotating, scaling, cropping or flipping the image.[68] While augmentation follows similar steps as above, it is only applied to the training data sets and not to the test sets.[69]

2.4.3. Training stage

Similar to NLP, Computer Vision also has supervised, semi-supervised and unsupervised training options. Supervised and semi-supervised requires annotated datasets. In unsupervised learning, computer vision is able to recognise common features in images (cluster analysis), so it works without annotations.[70]

Annotation is performed by assigning a label to the selected part of the image, or a single label for the entire image.[71] Feature extraction can be included under this stage – or alternatively be seen as a separate stage in the computer vision process. A feature is defined as "a measurable piece of data in your image that is unique to that specific object…a distinct color or a specific shape such as a line, edge, or image segment".[72] The features can be extracted manually or automatically. The training then occurs based on the extracted features.

[63] Appen, How to Create Training Data for Computer Vision Use Cases (2019) https://appen.com/blog/how-to-create-training-data-for-computer-vision-use-cases/
[64] https://www.dynam.ai/computer-vision-projects-management-part-1/
[65] Lila Mullany, 'Introduction to Computer Vision Model Training' (2020) < https://towardsdatascience.com/introduction-to-computer-vision-model-training-c8d22a9af22b>
[66] Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020).
[67] Sharath Kumar and Manjula Hosurmath, 'Multiclass image classification of yoga postures using Watson Studio and Deep Learning as a Service' (2019) 'https://developer.ibm.com/technologies/artificial-intelligence/tutorials/image-preprocessing-for-computer-vision-usecases/
[68] Appen, How to Create Training Data for Computer Vision Use Cases (2019) https://appen.com/blog/how-to-create-training-data-for-computer-vision-use-cases/; Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020)
[69] Joseph Nelson, 'Why Image Preprocessing and Augmentation Matter' (2020) https://blog.roboflow.com/why-preprocess-augment/
[70] Appen, How to Create Training Data for Computer Vision Use Cases (2019) https://appen.com/blog/how-to-create-training-data-for-computer-vision-use-cases/
[71] Lila Mullany, 'Introduction to Computer Vision Model Training' (2020) https://towardsdatascience.com/introduction-to-computer-vision-model-training-c8d22a9af22b
[72] Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Some steps here can be merged due to the technological developments in deep learning. Convolutional neural networks (CNN) are used for image classification and recognition problems.[73] Prior to CNNs, the standard ML training process (for videos) included (i) extracting the features, (ii) combining the features into a fixed-sized video level description and (iii) a classifier is trained on 'bag-of-words' level descriptions - CNNs combine all these stages.[74]

CNNs have layers of "small computational units that process visual information hierarchically in a feed-forward manner", so each layer works as an image filter and extracts a feature from the image and the image becomes increasingly more explicit along this hierarchy.[75] The process is a slightly different for videos. When used for a video, AI technology has to detect key images which are the most relevant images in the video and eliminate redundant or blurry images. Doing so simplifies the analysis work afterwards.[76] CNNs can be used both supervised and unsupervised, and although widely used for image classification, they can also be used for text classification.[77]

2.4.4. Models for Content Moderation

Trained models can be used in tasks such as image classification (used for example in medical diagnosis or reading traffic signs), object detection and localisation, generating images, face recognition and image recommendation.[78] Some tasks of computer vision are more suitable for unsupervised methods (such as image classification), while others might require more human input.
When using AI for content moderation, it is also possible to combine computer and human moderation: for example, when determining if the user generated content is harmful; the AI application can flag some as "uncertain", which then goes to human moderators whose decisions can be fed back as training data for the AI to learn how to address similar images or videos.[79] Trained on datasets for recognising things like nudity, violence or drugs, there are various companies that are using AI technology for content moderation.[80]

As a final note, in the example of using computer vision for content moderation, AI is only one of the methods. There are also methods called hashing and fingerprinting. Hashing works by generating unique identifiers for files and then comparing it with reference databases for detecting things like terrorist content or viruses.[81] Fingerprinting is similar to hashing, but the unique identifier is not based on the file, but for the characteristics of its content.[82] While it is easier to match content found online to previously flagged content, training AI to make decisions on new content is more difficult. Furthermore, the reasoning for AI decisions is more obscure.[83]

---

[73] Andrej Karpathy et al, 'Large-scale Video Classification with Convolutional Neural Networks' (2014) IEEE Conference on Computer Vision and Pattern Recognition
[74] Andrej Karpathy et al, 'Large-scale Video Classification with Convolutional Neural Networks' (2014) IEEE Conference on Computer Vision and Pattern Recognition; Cambridge Consultants, Use of AI in Online Content Moderation 2019 Ofcom Report, 51-52
[75] Leon S Gatys, Alexander S Ecker and Matthias Bethge, 'A Neural Algorithm of Artistic Style' (2015)
[76] 'Mission Report: Towards more effectiveness of copyright law on online content sharing platforms: overview of content recognition tools and possible ways forward' (English version) Joint Report by CSPLA, CNC and HADOPI (January 2020).
[77] Joris Guérin, Olivier Gibaru, Stéphane Thiery, and Eric Nyiri, 'CNN features are also great at unsupervised classification' (2018) 8th International Conference on Computer Science, Engineering and Applications.
[78] Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020).
[79] European Parliament Study, 'The impact of algorithms for online content filtering or moderation' Policy Department for Citizens' Rights and Constitutional Affairs (2020) 23.
[80] Examples include Clarifi, Amazon Rekognition, Valossa and Sightengine. EUIPO Automated Content Recognition – Discussion paper Phase 1 (2020) https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/documents/reports/2020_Automated_Content_Recognition/2020_Automated_Content_Recognition_Discussion_Paper_Full_EN.pdf.
[81] EUIPO Automated Content Recognition – Discussion paper Phase 1 (2020) 7.
[82] EUIPO Automated Content Recognition – Discussion paper Phase 1 (2020) 15.
[83] 'Mission Report: Towards more effectiveness of copyright law on online content sharing platforms: overview of content recognition tools and possible ways forward' (English version) Joint Report by CSPLA, CNC and HADOPI (January 2020).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

At this stage, AI technology is mainly used for improving and making fingerprinting faster, but is not sufficient on its own for copyright content moderation.[84]

3. Legal analysis

The case studies seek to capture legally pertinent stages of the identified technological processes. Accordingly, they represent the underlying factual set of assumptions on which we will base our legal analysis.

At the outset, it is important to note how the legal framework in this specific field is in a phase of transition, with new exceptions entering into force with the adoption and transposition of the CDSM Directive. In particular, Arts. 3 and 4 significantly reshape the *acquis* applicable to text and data mining but also more generally to the broader field of data analytics. These two articles introduce two mandatory exceptions that will exempt respectively acts of reproduction for the purpose of text and data mining made by research organisations and cultural heritage institutions for the purpose of scientific research (Art. 3) or by anyone for any purposes but with the possibility of "contract-out" (Art. 4).

It should be noted, however, that while the deadline to implement the CDSM provision was 7 June 2021, a delay in the transposition by a considerable number of MS should be accounted for. Therefore, the legal analysis presented in this paper will focus on the EU provisions, while the domestic implementations will be considered in the final report when more information will be available.

Furthermore, in this section we also focus on the underlying legal framework in which the TDM exceptions are set, i.e. the right of reproduction contained in Art. 2 ISD. In doing so, proper consideration is given to Art. 5(1) ISD which creates a limited but key exception for certain acts of temporary reproduction which has been tasked by the EUCJ with the crucial role of enabling technological development. Consequently, the legal analysis concentrates on the relationship between the aforementioned legal provisions (Arts. 2, 5(1) ISD; 3 & 4 CDSM) and the technological processes identified in the case studies (e.g. data acquisition, data (pre-)processing, and data analysis).

The analysis is structured as follows:

- Identification of the relevant *acquis* and the systematic classification of the new TDM exceptions;
- The right of reproduction;
- The exception for temporary reproductions (Art. 5(1)) as interpreted by the CJEU;
- Temporary and permanent reproductions in TDM and data analytics;
- The nature of data analytics as a copyright relevant act
    - International and EU law considerations;
- Detailed analysis of the provisions of Art. 3&4
    - Definitions
    - Rights
    - Contractual and technological overridability
    - Lawful access to original sources
    - Storage of copies for verifiability;
- Final interim considerations on the role of the TDM exceptions in data analytics and Artificial Intelligence systems.

[84] EUIPO Automated Content Recognition – Discussion paper Phase 1. Further work on this case study will be coordinated with WP6 which is focusing on similar technological elements, albeit from a different research angle, as also pointed out by one QE during the revision phase.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Annexes

Annex A – Legal analysis of rights and exceptions for input data in machine learning environments:
Thomas Margoni and Martin Kretschmer (2021) 'A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology' (full paper)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626