OECDpublishing

INTELLECTUAL PROPERTY ISSUES IN ARTIFICIAL INTELLIGENCE TRAINED ON SCRAPED DATA

OECD ARTIFICIAL INTELLIGENCE PAPERS

February 2025 **No. 33**

GPAI

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This report examines recent developments at the intersection of artificial intelligence (AI) and intellectual property rights, with a particular focus on data scraping practices. It provides an overview of the role of data scraping in AI training, current legal frameworks and stakeholder perspectives, as well as preliminary considerations and potential policy approaches to help guide policymakers in navigating these issues and facilitate a greater understanding of data scraping.

This report discussed by the OECD Working Party on AI Governance (AIGO) at its November 2023 and June 2024 meetings. The Global Partnership on Artificial Intelligence (GPAI) discussed this work during its November 2024 Plenary.

The report was written by Professor Lee Tiedrich (Distinguished Faculty Fellow in Law and Responsible Technology at Duke University), Karine Perset, and Sara Fialho Esposito under the supervision of Audrey Plonk, Deputy Director of the OECD Science, Technology and Innovation Directorate.

This paper was approved and declassified by written procedure by the Global Partnership on Artificial Intelligence (GPAI) on 30 January 2025 and prepared for publication by the OECD Secretariat.

Note to Delegations:

This document is also available on O.N.E Members & Partners under the reference code:

DSTI/DPC/GPAI(2024)2/FINAL

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Cover image: © Kjpargeter/Shutterstock.com

Acknowledgements

The paper “Intellectual property issues in artificial intelligence trained on scraped data” was presented and discussed at the AI Governance Working Party (AIGO) and the Global Partnership on Artificial Intelligence (GPAI) meetings. As such, it greatly benefited from the feedback and suggestions of numerous country delegates and experts. The authors would like to extend their sincere gratitude to the Delegations of Brazil, Colombia, France, Germany, Israel, Japan, Mexico, New Zealand, Singapore, Switzerland, the Republic of Türkiye, the United Kingdom, and the United States for their invaluable insights. In particular, they gratefully acknowledge contributions from Franklin Rodrigues Hoyer (Brazil), David Turnbull (United States) Susan Allen (United States), Jesse Dunietz (National Institute of Standards and Technology, United States), Michael Shapiro (United States), Ulrike Moetzel (Germany), Larissa Lim (Singapore), Melisa Tekeli Oltulu (Turkey), Angie Lizzethe Contreras Sanabria (Colombia), Yohann Ralle and Michaël Reffay (France), Ziv Katzir and Eden Israely (Israel), Sarah Box (New Zealand), Julian Frohnecke (Germany) and Joseph Phillips (United Kingdom).

This paper also benefited from discussions at GPAI workshops on data scraping and intellectual property rights, supported by the French GPAI Expert Support Centre (Inria) in January, March, and May 2024. These workshops were conducted under Chatham House rules and brought together a diverse group of scholars and practitioners and other stakeholders, we are thankful to all participants for their valuable contributions. The authors are especially grateful to Aurélie Simard, Yann Dietrich, Kaitlyn Bove, Josef Drexl and Daria Kim for their leadership in helping organise these events and for their valuable insights. Appreciation is also extended to Duke University in hosting two workshops at its Washington, D.C., office and the Max Planck Institute for Innovation and Competition for hosting an event in Munich, Germany. The authors further thank Michele Woods and Ulrike Till from the World Intellectual Property Organization (WIPO) for their interventions and support during the AIGO Committee meeting and the workshops. Additionally, Celine Caira and Rashad Abelson provided substantial inputs for the drafting of this report.

The Secretariat would also like to thank various stakeholder groups at the OECD for their valuable input, particularly Nicole Primmer, Barry O'Brien, and Maylis Berviller (Business at OECD – BIAC); Pam Dixon (Civil Society Information Society Advisory – CSISAC); and Sarah Jameson (Trade Union Advisory Committee – TUAC).

Finally, the team acknowledges the helpful input and reviews by OECD team members: John Tarver, for his editorial assistance, and Andreia Furtado, for her support with publishing and formatting. Their engagement significantly enhanced the quality of this report.

Table of contents

FIGURES

TABLES

Abstract

Recent technological advances in artificial intelligence (AI), especially the rise of generative AI, have raised questions regarding the intellectual property (IP) landscape. As the demand for AI training data surges, certain data collection methods give rise to concerns about the protection of IP and other rights. This report provides an overview of key issues at the intersection of AI and some IP rights. It aims to facilitate a greater understanding of data scraping — a primary method for obtaining AI training data needed to develop many large language models. It analyses data scraping techniques, identifies key stakeholders, and worldwide legal and regulatory responses. Finally, it offers preliminary considerations and potential policy approaches to help guide policymakers in navigating these issues, ensuring that AI's innovative potential is unleashed while protecting IP and other rights.

Résumé

Les récentes avancées technologiques en matière d'IA, en particulier l'essor de l'IA générative, ont soulevé des questions concernant le paysage de la propriété intellectuelle. Alors que la demande de données d'entraînement pour l'IA ne cesse de croître, certaines méthodes de collecte de données suscitent des préoccupations quant à la protection des droits de propriété intellectuelle et d'autres droits. Ce rapport propose un aperçu des enjeux majeurs à l'intersection de l'IA et de certains droits de propriété intellectuelle. Il vise à faciliter une meilleure compréhension du *scraping* de données (extraction de données), une méthode clé pour obtenir les données d'entraînement nécessaires au développement de nombreux grands modèles de langage. Il analyse les différentes techniques de *scraping*, identifie les principales parties prenantes, ainsi que les réponses juridiques et réglementaires à l'échelle mondiale. Enfin, ce rapport présente des pistes de réflexion et des propositions de politiques publiques pour favoriser l'innovation technologique tout en protégeant les droits de propriété intellectuelle.

Executive summary

Recent advancements in AI, particularly the emergence of generative AI, have introduced complex challenges in the intellectual property (IP) landscape. The development, testing, and validation of AI models rely heavily on access to large datasets, driving a surge in demand for training data. A widely used method for collecting such data is “data scraping” which, in this report, refers to the automated extraction of information from third-party websites, databases, or social media platforms. Data scraping directly affects creators and owners of IP-protected works, especially when conducted without consent or payment to rights holders. Scraping activities can implicate several types of IP and similar rights, including copyright, database rights, trademarks, trade secrets, publicity, and moral rights.

The legal landscape surrounding IP data scraping is complex and rapidly evolving. Existing IP laws, many predating modern AI practices, differ across jurisdictions, complicating their application. Data scraping frequently involves content protected by IP rights, raising questions about infringement, the applicability of exceptions such as fair use or text and data mining (TDM) provisions, and adherence to contractual terms and conditions. Scraping copyrighted materials raises questions about whether the collection or use of the scraped data constitutes copyright infringement. Litigation in this area is increasing globally, with prominent cases emerging in the United States, European Union, and beyond. Additionally, concerns about AI-generated outputs—particularly those that mimic an individual's style, voice, or likeness without authorisation— have prompted varied legal responses aimed at protecting rights and preventing misuse.

Data scraping is now a widespread practice, but it encompasses various methods and currently lacks a universally accepted definition. The term data scraping is often conflated with “data mining,” which refers to computational processes for identifying patterns, trends, and correlations, as well as with techniques such as “web crawling.” This report highlights the inconsistencies in definitions and proposes a broad working definition for data scraping. Components of data scraping include data collection, data pre-processing, and data usage. The report analyses different scraping techniques and emphasises the need for common/standard terminology and clearer distinctions between these methods.

Different actors in the data scraping ecosystem raise various types of legal issues. Some also use data scraping to support research and other endeavours, suggesting the need for policy tools tailored to different use cases. The data scraping ecosystem encompasses research institutions and academia, AI data aggregators, as well as technology companies and platform operators. Research institutions and academia frequently employ data scraping to gather data for academic and scientific purposes. AI data aggregators are reported to make scraped data available to third parties, often without clear licensing terms or clear disclosure of data provenance, raising IP and other legal concerns. Technology companies and platform operators are sources of scraped data and regular data scrapers themselves.

A “data scraping code of conduct,” standard contract terms, standard technical tools and initiatives for building awareness could help chart a responsible path for data scraping in an internationally coordinated manner. This would be particularly effective if it is developed with input from a broad and diverse set of stakeholders, including rightsholders, researchers, AI developers, civil society, and policymakers.

- **A voluntary "data scraping code of conduct"** could establish broadly applicable provisions while providing specific guidelines for different actors in the AI ecosystem. These provisions could address the distinct roles of AI data aggregators and users of scraped data. To promote consistency, the code could include standard terminology, ensuring a shared understanding of data scraping activities among stakeholders. Additionally, it could include mechanisms for monitoring adherence, such as a registry system, and offer recommendations for transparency and documentation practices. Finally, the code could include **standard contract terms.**
- **Standard technical tools** could help protect IP rights and enable rights holders to manage access to their data with greater ease. These tools could include data access control mechanisms, automated contract monitoring, and direct payment systems. Such standardised tools could streamline compliance for organisations while simplifying protection of rights holders' across multiple platforms.
- **Standard contract terms** could address legal and operational issues associated with data scraping. These terms could serve as optional starting points while allowing organisations to negotiate their specific conditions. The development of these terms would benefit from collaboration among multiple stakeholders and could be tailored for different use cases, ranging from non-profit research to commercial applications.
- **Raising awareness** of data scraping and its legal implications could empower stakeholders with information on how to protect and manage their rights. This includes helping rights holders to understand their protections, educating AI system users about responsible usage, and ensuring that all participants in the AI data ecosystem understand their roles and responsibilities.

Introduction

AI, including generative AI, is raising complex issues in the intellectual property landscape

Intellectual property (IP) laws have historically incentivised innovation by protecting creators, leading to the development of valuable products, technologies and other creations (collectively "creations") that greatly benefit society. While these laws may vary to some extent across jurisdictions, they share common principles by granting rights holders certain rights over their creations.[1] These legal frameworks help incentivise innovation and foster creativity by protecting the assets developed by individuals and businesses.

To balance the protection of IP owners' rights with broader societal interests, IP laws generally include limited, context-specific exceptions that allow third parties to use IP-protected works without needing permission from the rights holders. For example, certain copyright laws allow the use of copyright protected works for non-commercial or other limited purposes if specific conditions are met. Some exceptions require payment, while others do not. Additionally, IP protections generally expire after a specified duration , allowing IP to enter the public domain.

Although IP laws may differ across national legal systems, international treaties have significantly contributed to the harmonisation of IP rights and principles globally. These include the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS Agreement), administered by the World Trade Organisation (WTO), and key conventions administered by the World Intellectual Property Organisation (WIPO) such as the Berne and Paris Conventions and the WIPO "Internet Treaties". This harmonisation ensures a baseline of commonality, particularly in areas such as copyright and trademark protections as outlined in Section 3.[2]

In contrast to copyright and trademark laws, trade secret laws typically protect against the third-party unauthorised use or disclosure of confidential information, instead of granting developers exclusive rights. IP protection extends to various materials including documents, software, photographs, and graphic works across any medium or format. Some jurisdictions recognise *sui generis* (unique) rights to protect specific types of materials under IP law. For example, the European Union (EU) provides *sui generis* database rights (European Union, 1996). These specialised protections may help address the evolving nature of intellectual property across different mediums and formats.

IP protections are considered to have contributed significantly to global innovation, including AI elements such as copyright-protected software (available via open-source or other types of licenses). Major breakthroughs have emerged through open scientific publications and open-source initiatives, enabling widespread adoption and driving major advances in language models. For instance, Google's development of word vectors (Word2Vec) was made publicly available, revolutionising natural language processing and accelerating the development of advanced text analysis techniques (Mikolov, Chen, Corrado, & Dean, 2013). This diverse ecosystem, encompassing both proprietary and open innovation, underscores the importance of maintaining strong, effective, and predictable IP frameworks to support the continued advancement of AI innovations and other creations.

Although this paper focuses on IP considerations related to data scraping as a mechanism to obtain data for training AI models, it is important to acknowledge that AI and IP considerations arise throughout the entire AI system lifecycle. For instance, when generative AI systems produce new content ("AI-generated outputs") in response to a prompt, questions emerge about whether these outputs should qualify for IP protection. Many jurisdictions currently require human involvement for copyright protection, but questions remain about the level of human involvement needed to obtain such protections and who the rights holders would be (Mammen, et al., 2024).

Complex liability questions can arise when AI-generated outputs allegedly infringe third party IP rights or cause other harm. These issues are particularly challenging as AI systems operate within value chains that involve multiple parties, from upstream suppliers to downstream users. Additionally, cross-border AI operations raise jurisdictional questions, as discussed in Section 3. While this report addresses concerns related to data collection and the potential for AI outputs to mimic input data, it does not explore broader liability issues or questions about IP eligibility for AI-generated content. In particular, this report does not discuss issues raised by the creation or co-creation of artificially generated patented inventions.

This report provides an overview of key issues in AI and IP rights in the context of data scraping. It facilitates a greater understanding of data scraping, the techniques and actors involved, as well as how legal and regulatory regimes have responded around the world. It concludes by providing preliminary considerations and potential policy approaches to help guide policymakers in the path ahead, to ensure AI's innovative potential is unleashed while IP and other rights are protected. In doing so, it aligns with the OECD AI Principles [OECD/LEGAL/0449], which advocate for the development and use of AI that is both innovative and trustworthy while respecting and addressing risks related to IP and other rights. Additionally, the report contributes to discussions on the OECD Recommendation on Enhancing Access and Sharing of Data (EASD Recommendation) [OECD/LEGAL/0463], aiming to maximise the benefits of data access and sharing while fostering trustworthiness and safeguarding individual and organisational rights, including IP rights.

While data scraping raises significant concerns regarding privacy, data protection and related issues, this report focuses on its implications for IP. Privacy and data protection concerns are being explored through complementary work at the OECD and beyond, including by the OECD.AI Expert Group on AI, Data, and Privacy. This expert group, a joint initiative of the OECD Working Party on Data Governance and Privacy (WPDGP) and the OECD Working Party on AI Governance (WPAIGO), is conducting analysis at the intersection of AI and privacy, particularly when AI training data includes personal data. Their previous and ongoing work complements this analysis by examining various methods of collecting and processing AI training data. This can help ensure that privacy considerations are thoroughly integrated with broader data governance frameworks for a more comprehensive approach (OECD, 2024). These efforts are in line with broader international initiatives, such as the joint statement on data scraping and data protection published by the UK Information Commissioner's Office (ICO) in August 2023. Signed by twelve data protection authorities from around the world, the statement highlights the need for global efforts to address privacy risks associated with data scraping (ICO, 2023).

"Data scraping'' used to compile data to train AI systems poses significant challenges

Recent AI advances, including generative AI, pose several new policy challenges to current IP frameworks. Governments are taking note globally and beginning to reflect and act on how such challenges should be addressed. For example, during Japan's G7 Presidency in 2023, G7 leaders identified infringement of IP rights as one of the major risks stemming from generative AI (OECD, 2023).

IP policy challenges can arise in the early stages of the AI system lifecycle, including during the data collection and processing phase, when data is aggregated to train, fine-tune, test, or validate AI models and systems, such as generative AI systems (OECD, 2023). Developing high-performance generative AI systems and other AI systems based on machine learning often requires access to vast amounts of data for training (AI training data) and to improve their accuracy and performance (Clark & Perrault, 2022). AI training data can include personal information, facts, creative content, software, audio files, video, digital images, and just about any other digital content. AI input data can be collected in various ways, including through data scraping, where AI system developers extract information from third party websites or social media platforms without any coordination with the entity hosting the data. While data scraping for AI systems and similar practices are not new, its use has significantly increased with the rapid growth of generative AI (Metz, Kang, Frenkel, Thompson, & Grant, 2024).

Data scraping occurs across jurisdictions and presents many pressing, and potentially competing, privacy, IP and other policy issues. On the one hand, if undertaken responsibly, AI data scraping can provide access to diverse and vast amounts of data which is essential to advance AI research and innovation consistent with the OECD AI Principles [OECD/LEGAL/0449].[3] Increased access to reliable and diverse legally sourced AI training data can also reduce potential risks of bias and other harms and help close digital divides by enabling the development of localised AI tools for historically underserved communities (Chason, 2024; Lee & Lai, 2022; Chen, Wu, & Wang, 2023; Hall, Vassilev, Greene, Perine, & Patrick, 2022).

On the other hand, without appropriate guardrails, data scraping for training AI models or systems may violate IP, privacy, and other rights, threaten safety, and/or contribute to other harms. From an IP perspective, risk assessments typically focus on whether data scraping violates IP rights such as copyright and trademark, trade secrets or other rights that may exist in the AI training data. However, additional IP-related issues include whether data scraping practices have unlawfully removed so-called rights-management information embedded in the AI training data and/or circumvented such technical protection measures, or breached contract terms pertaining to IP-protected data.

1 Understanding AI data scraping

Developers of generative AI systems (that produce content as an output) typically rely on two main sources of data to train their models: targeted sourced data and large volumes of scraped data (Baack, 2024). Targeted sources are characterised by their smaller size, well-defined parameters, and selection based on specific quality metrics (including accuracy, completeness, consistency, reliability, validity, timeliness). These sources are chosen to help AI systems imitate diverse styles and expressions. A critical advantage of targeted sources is that their provenance can be clearly identified, allowing developers to verify the quality and reliability of the data. From an IP perspective, this can also help enable developers to confirm compliance with IP right and legal requirements.

Within targeted sourced data, various licensing frameworks exist for payment-free use. 'Gratis' licenses permit free usage while potentially limiting modifications or redistribution rights, whereas 'libre' licenses offer broader freedoms for modification, sharing, and redistributions, subject to specific license terms. Notable examples include Wikipedia snapshots and scientific texts from arXiv. The ecosystem also encompasses "public domain" materials, (i.e., works that are not protected by IP rights or similar restrictions), such as those available through Project Gutenberg (Project Gutenberg, 2024). However, many targeted datasets contain IP-protected content that typically require payment for use and are governed by specific terms that require compliance with IP laws and additional contractual obligations. Such licensing arrangements are viewed as providing legal certainty that helps both developers and rights holders manage risks.

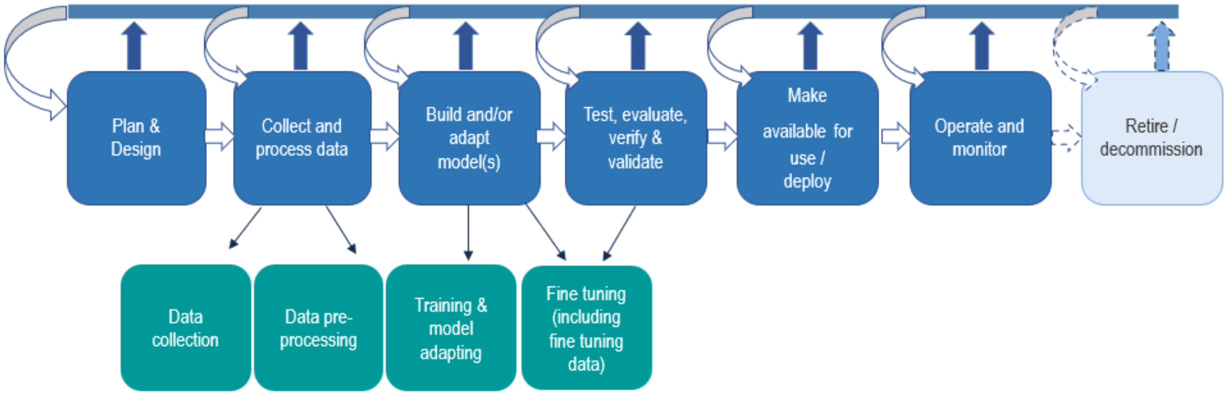
Given the size constraints of targeted datasets, developers frequently supplement their training data with large-scale scraped data (Baack, 2024). This scraped data, sourced from books, websites, blogs, forums, databases, and social media platforms, has proven crucial in advancing general-purpose AI models by providing the extensive and contextually rich corpus necessary for effective model training (Soldaini, et al., 2024). For instance, the BLOOM language model, developed by the BigScience research initiative, used data from Common Crawl, an extensive source of web-scraped multilingual data, to enhance language diversity and include low-resource languages (Teven, et al., 2022). However, scraped data presents significant challenges. For instance, it typically contains more irrelevant information or errors than targeted sources, requiring advanced filtering techniques for quality control. It also raises complex IP compliance issues, especially when the provenance of the scraped data is unclear, making it difficult to verify whether appropriate permissions have been secured.

While data scraping has become commonplace, it currently lacks a broadly accepted standard definition and can encompass a range of techniques and activities, as discussed later in this section. Data scraping is commonly referred to as "web scraping", as it primarily involves extracting data from websites. However, it can also include other methods of data extraction, such as screen scraping and web crawling. "Web crawling," though also not well defined, generally refers to the automated indexing or other aggregation of information from websites through the systematic browsing of web pages, often for purposes such as search engine indexing (S. Gillis, Definition: web crawler, 2024). For purposes of this report, indexing is considered a form of aggregating, collecting and/ or gathering data.

"Data mining" generally refers to computational processes used to identify patterns and correlations in large datasets. There is some issue about whether and how the terms "data mining"' and "data scraping'' relate to legal definitions. The legal term "text and data mining" (TDM) is commonly used by many countries in copyright laws to create limited exceptions to copyright protection. Unlike technical definitions, legal definitions may vary from jurisdiction to jurisdiction. Legal definitions may be open to interpretation by judicial or regulatory authorities.

Data scraping can involve an array of different activities, including: (1) data collection, (2) data pre-processing, (3) using the data for model training, model improvement, and/or (4) fine-tuning based on testing, evaluation, verification and validation (see Figure 1) (OECD, 2024).

Figure 1. The AI model development lifecycle


Plan & Design
Collect and process data
Build and/or adapt model(s)
Test, evaluate, verify & validate
Make available for use / deploy
Operate and monitor
Retire / decommission
Data collection
Data pre-processing
Training & model adapting
Fine tuning (including fine tuning data)

Source: OECD illustration based on AI system lifecycle (OECD, 2024),

This paper uses the term “data scraping” to refer broadly to activities related to the collection and pre-processing and training of data extracted from various sources, using techniques such as web scraping, web crawling, and screen scraping. It also considers the use of this data within the broad context of the AI system lifecycle, including subsequent stages, such as when data is leveraged to identify patterns, extract features, optimise models, and make predictions.

Having a clear understanding and overview of these activities and techniques is important for addressing policy challenges related to data scraping. Policymakers may wish to explore the development of standard definitions to provide stakeholders with a common nomenclature and taxonomy. In developing these definitions, it would be helpful to consider technical, legal and/or other terms that may already exist in different jurisdictions. One approach could be to develop a broader, overarching definition of data scraping, with other definitions differentiating between specific techniques and/or activities that fall under this broader concept, while noting their potential copyright implications.

Box 1. Key concepts: defining terminology around how AI training data is collected to develop AI systems

Access to training data is crucial for building and/or adapting AI model(s); testing, evaluating, verifying and validating (known as TEVV) AI systems (Figure 1). The demand for AI training data has increased dramatically over the past few years, particularly with the rise of generative AI systems. Today, AI training data is sourced in a variety of ways, including by procuring curated datasets, engaging in data-sharing agreements, collecting user data, using stored data, and automatically collecting *i.e.,* “scraping” publicly accessible data from the Internet. Although consensus is needed , the following describes key terms used in this report. A preliminary working definition of data scraping is also proposed below:

- **AI training data**: AI training data refers to datasets used for AI system development, including for training, testing, evaluation, verification and validation. They include both structured and unstructured data, such as numerical, text, image, or audio data, collected through human or machine means. AI training data is leveraged during the development of the AI system ("in the lab") to build AI systems e.g. leveraging data to establish patterns and predictive behaviours (OECD, 2022). AI training data is generally complemented with other types of input and data used for specified purposes in the AI system lifecycle, notably input data used at runtime (such as a user query).
- **Data scraping:** The term does not yet have a widely accepted definition. This report proposes a broad working definition of data scraping, which can be complemented with other definitions focused on specific techniques and/or activities used in data scraping. Thus, in this report, data scraping refers to the automated extraction of AI training data from the web, online databases and from other sources using automated software tools or scripts. While data scraping may use a variety of techniques, it is generally characterised by the following features:
    - **Automation**: Data scraping typically involves the use of software tools or scripts designed to quickly and efficiently harvest or otherwise aggregate data with minimal human intervention**.**
    - **Scalability:** Data scraping is often used to collect or make accessible large amounts of data that would be impractical to aggregate manually. In addition, the tools and techniques employed can be scaled up to extract data from numerous sources simultaneously.
    - **Lack of coordination:** Data scraping is often done without coordination between the data scraper and the entity hosting the data.
- **Generative AI**: A class of AI models that emulate the structure and characteristics of AI training data in order to generate derived synthetic content (US Executive Office of Science and Technology Policy, 2023). This can include images, videos, audio, text, and other digital content.
- **Data mining:** While data scraping focuses on the extraction of data, data mining refers to the automated techniques used to analyse the data. Data mining is dedicated to analysing insights from datasets, including datasets obtained through scraping and/or other data extraction methods. ISO has defined data mining: "computational process that extracts patterns by analysing quantitative data from different perspectives and dimensions, categorizing them, and summarizing potential relationships and impacts" (ISO, 2022).

Text and data mining (TDM): In EU law, TDM is defined as "[a]ny 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 of Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market) (EUR-Lex, 2019). Other jurisdictions may define TDM differently in their legal frameworks.

Source: Indicated in-text.
Note: This is an inexhaustive and preliminary list of terms selected for the analysis in this report.

Data scraping components include data collection, data pre-processing, and data storage

Developing a clear definition of data scraping poses challenges due to the variety of techniques it encompasses. These techniques range from web scraping and web crawling to screen scraping, each with distinct technical approaches and applications. To ensure a common understanding among stakeholders, it is important to define and differentiate these methods clearly. These definitions could also provide clarity to facilitate compliance with IP laws and enforcement of such laws, when needed. This sub-section provides a high-level overview of the main activities involved in data scraping, including data collection, data pre-processing, and data storage.

Data collection via data scraping

A variety of techniques are used to collect scraped data, including web scraping, web crawling, screen scraping and other methods. These techniques often raise important IP considerations that vary depending on the nature of the data being collected and the manner of collection. Detailed IP implications related to data scraping are discussed in Section 3 of this report.

Web scraping and web crawling

Web scraping and web crawling represent closely related but distinct methods of extracting information from websites.

Web scraping typically involves extracting data from third-party websites or social media platforms, using specialised software like crawlers, bots and scrapers. The process typically entails sending a request to a website, parsing the website's HTML content, and extracting the desired data from such website.

Web scraping practices vary in scope, from collecting entire webpages to targeting specific amounts of third-party website data, depending on the intended purpose (National Library of Medicine, 2024). Websites often employ contractual safeguards in their Terms of Service and/or technical safeguards such as the Robot Exclusion Protocol (robots.txt). This protocol allows website owners to provide machine-readable instructions that request web crawlers to not collect data. Although many web scraping practices comply with these safeguards, others may bypass or ignore these protections. Certain AI-powered systems, for example, may selectively disregard these restrictions when prompted by specific user requests, raising concerns about the effectiveness of current measures intended to limit data scraping (Mehrotra & Courts, 2024).

Web crawling refers to the automated navigation and indexing of web pages via hyperlinks. This technique uses web crawlers, also known as spiders or bots, to systematically collect data from web pages, databases or other digital sources, often by crawling through the content and retrieving specific information (Shayne Longpre, 2024). This technique is commonly used by search engines to index website content for search functionality. As discussed in the Section 3, some AI data aggregators rely on links, rather than downloading and storing large datasets. For instance, the AI data aggregator LAION (Large-scale Artificial Intelligence Open Network) primarily provides access to datasets through links to externally hosted data sources (Guadamuz, 2023).

Screen scraping

In contrast to web scraping, screen scraping involves the automated extraction of data that is visually displayed on a screen, such as text, images and videos, PDFs and other documents formats. Instead of interacting directly with the website's underlying text or HTML code, screen scraping captures and processes visual output that is rendered as it would be on a user's device. This can involve various methods, ranging from simple image capture and text recognition to more complex processes simulating user interactions with a graphical interface (S. Gillis, Screen Scraping, 2023).

This technique is particularly prevalent in the consumer services sector, such as for open banking initiatives and lending applications. However, many jurisdictions are transitioning towards more secure Application Programming Interface (API)-based systems. (The Australian Government the Treasury, 2023). APIs allow users to request data downloads within predefined operational and legal parameters, as seen in websites such as Wikipedia. (The Australian Government the Treasury, 2023). API usage is usually governed by contractual agreements. When data is obtained using an API in compliance with the applicable contract, it does not fall under the category of data scraping, which typically occurs without direct coordination between the data scraper and the entity hosting the data.

Other techniques

The range of AI data scraping techniques is expanding due to the significant and urgent demand for larger quantities of data to train advanced AI systems. A group of researchers predicts that AI developers may run out of high-quality language data between 2023 and 2026 (Villalobos, et al., 2022). This has led to innovative approaches, such as using speech recognition tools to transcribe publicly accessible videos to generate additional AI training data (Metz, Kang, Frenkel, Thompson, & Grant, 2024)

Data pre-processing and storage

Once the desired scraped data is collected or otherwise aggregated, it is often processed and stored. For many kinds of AI systems, raw data must be selected and transformed to be usable for AI model training. Consequently, scraped data is typically pre-processed to convert it into usable formats suitable for AI training. Organisations typically rely on a variety of methods to pre-process scraped data, depending on the type of data that is collected (images, text, videos, contact data, etc.).

Scraped data generally also needs to be stored. Scraped data is typically stored in structured databases. Meta, for instance, used to rely on different storage systems for specific tasks, until it consolidated everything into a much larger, unified storage system capable of holding massive amounts of data, reaching an exabyte scale (equivalent to 1 billion gigabytes) (Engineering at Meta, 2021).

Scraped data usage

Finally, as reflected in Figure 1, organisations may use scraped data in different ways, such for training, fine-tuning, validation and/or testing of AI models and systems. The following section further describes some uses of scraped data.

2 The AI data scraping ecosystem

The AI data scraping ecosystem includes a diverse range of commercial, non-commercial, and government entities that engage in and/or benefit from data scraping. The key actors in the ecosystem can be categorised into distinct groups, each with specific roles and challenges.

Research institutions and academia often use data scraping to collect data for academic and scientific purposes

Research institutions and universities play significant roles in the AI data scraping ecosystem. These entities often rely on data scraping techniques to collect large volumes of data for academic and scientific purposes. Researchers use this data to conduct studies, develop new AI models, and refine analytical methods, thereby advancing scientific knowledge and contributing to the development of cutting-edge technologies. For instance, scraped data has been used to enhance sustainability analysis and improve climate modelling (Guttridge-Hewitt, 2023).

While their research objectives are typically legitimate, these institutions and researchers may face complex legal challenges such as navigating copyright and data privacy regulations when scraping data. Some jurisdictions provide special exceptions for research purposes or fair use doctrine, though navigating these legal frameworks can be complex, especially when research involves international datasets subject to multiple legal standards.

AI data aggregators collect and make scraped data available to third parties

The data scraping ecosystem includes entities that aggregate and make scraped data available to third parties. AI data aggregators may provide data on an open-source and/or fee-based basis. Some AI data aggregators, such as Common Crawl, LAION and EleutherAI are non-profits that make scraped data available without charge on their websites (Common Crawl, 2024) (LAION, 2024) (EleutherAI, 2024). Common Crawl, founded in 2007, was originally developed to crawl the web and provide accessible data for researchers and smaller businesses, well before the recent rise of generative AI. AI data aggregators also may scrape and/or use data for their own purposes.

Common Crawl offers a vast, freely accessible repository of web scraped data that is extensively used for pre-training LLMs (Baack, 2024). This data has a significant role in the development of LLMs as demonstrated by its use in training GPT-3, where over 80% of its tokens originated from Common Crawl. (B. Brown, et al., 2020). EleutherAI processes and refines portions of this web scraped data to produce specialised datasets such as the Pile-CC, optimised for training LLMs. These efforts attempt to reduce undesirable content and tailor the data for more specific AI training purposes (EleutherAI, 2024). Similarly, LAION uses data sourced from Common Crawl to create specialised datasets like LAION-400M. These datasets are specifically structured for AI applications such as image and text recognition models, further showcasing the value of processed scraped data in AI model training (LAION, 2024).

Open-source datasets like MNIST, ImageNet and Open Images are publicly accessible and provided under terms of an open license (Macgence, 2024). Open-source data aggregators like Hugging Face or GitHub list or provide links to open-source third party datasets on their websites for free downloads (Hugging Face, 2024). Off-the-shelf datasets are available for purchase from some commercial providers. For many AI data aggregators though, it remains unclear whether or under which license the datasets were acquired. The Data Provenance Initiative audited more than 1800 widely used datasets. It found that 70 percent of the datasets lacked information about licensed uses or were labeled as more permissive than the author's intended license. (Data Provenance, 2023). Overall, there is a lack of transparency and documentation about the origin of the data scraped (Tiku, 2023). Unsurprisingly, this situation has led to certain transparency requirements for general-purpose AI systems, for example in the EU AI Act, as discussed further below in Section 4. Transparency could also be enhanced through various policy tools, as discussed in Section 5 on Preliminary considerations and potential policy .

Disputes have arisen involving several AI data aggregator sites. Investigative journalism has reported that a popular training dataset 'Books3' contains more than 170,000 pirated books and has been used to train Meta's Llama, Bloomberg's Bloomberg GPT, EleutherAI's GPT-J and likely other generative AI models (Reisner, 2023). Similarly, Google's C4 dataset allegedly contains data scraped from 15 million websites and has been used to train Meta's Llama, Google's T5, and likely other LLMs. A scientific analysis of the C4 dataset indicates that much of its content originates from "journalism, entertainment, software development, medicine and content creation" websites (Schaul, Chen , & Tiku , 2024). Additional IP concerns have also emerged regarding the scraping of substantial amounts of data scraped from websites known for or otherwise associated with piracy and counterfeits (Schaul, Chen , & Tiku , 2024).

Technology companies and platform operators are both sources of scraped data and regular data scrapers themselves

Technology companies and platform operators, such as social media platforms, search engines, and e-commerce sites, are both sources of data for scraping and active participants in the data scraping ecosystem. These platforms are frequent targets for scrapers because they host vast amounts of user-generated content and other valuable data, which are often sought after for developing AI models. Platform operators often implement anti-scraping technologies (such as CAPTCHA and Internet Protocol address blocking) and terms of service restrictions, to regulate or prevent unauthorised access to their data. These measures aim to protect the platforms' data assets while also addressing IP and privacy concerns and safeguarding user rights.

In addition, many of these companies engage in data scraping activities themselves, collecting data to enhance their products and services. Data scraping underpins several key business models for these platforms, including for search engines, account or website aggregation, price comparison tools, and targeted advertising (Fei, 2024). For instance, LinkedIn has acknowledged using scraped data to improve its own services, reflecting the dual role that tech companies often play in the data scraping ecosystem (Wiggers, 2024).

Data scraping affects creators and owners of IP-protected works directly

Content creators, including writers, photographers, journalists, and artists, are seen as the most directly affected groups in the AI data scraping ecosystem. Their works are often scraped and used in datasets for training AI systems without their knowledge or consent, even though in some jurisdictions such acts may be deemed to be acts of copyright infringement. Some material may also be protected by trademarks. However, it is often difficult for creators to determine whether their work has been included in training datasets, because AI developers often operate with limited transparency. A report by Stanford University's Centre for Research on Foundation Models found that most developers of advanced AI systems are opaque about the origins and legality of their data, with only one out of fourteen developers disclosing details about data creators, copyright status, and data licenses (Bommassani, et al., 2024). This lack of transparency poses challenges for downstream developers and/or deployers, who are unable to verify compliance with licensing and/or laws, exacerbating issues related to data provenance. It also makes it challenging for creators and rights holders to track how their IP is used and, where applicable, authorise its use and assess whether their rights have been infringed. While in some jurisdictions, AI developers may be permitted to use copyright-protected works under specific legal frameworks, in others, rights holders may have legal avenues to seek redress and, where applicable, pursue stronger protections and/or adequate compensation for the use of their protected (or copyrighted) works.

The report also found that transparency regarding data access has significantly worsened dropping from 20% in the October 2023 version to just 7% in the May 2024 version. This steep decline may reflect legal uncertainties surrounding copyright and increasing concerns among developers about the risks of disclosing the datasets used for building AI models, particularly when those datasets may include copyrighted or illegal content.

To address these concerns, there are growing calls for policy measures that would require AI developers and AI data aggregators to disclose the sources of their training data. Implementing such transparency measures requires assessing the technical feasibility of identifying individual data sources in training datasets, associated costs to AI developers, as well as economic incentives and innovation outcomes for both rights holders and AI developers.

Other approaches under consideration include mechanisms for rights holders to provide explicit consent or to opt-out of data use, as well as systems to manage rights through licensing agreements and collective rights management organisations.

3 The legal landscape for data scraping and growing litigation

Since scraped data may include material protected by IP rights, including personal data and other data, it potentially presents various legal issues. Many existing IP laws were not enacted with current data scraping in mind, raising questions across jurisdictions on how such laws should be applied in this context. Further complicating matters, existing laws that may apply to data scraping typically vary across jurisdictions.

Data scraping can implicate several types of IP and similar rights, including copyright, database rights, trademarks, trade secrets, publicity and moral rights

Data scraping can implicate several types of IP and similar rights, depending on the nature of the data being scraped and how it is used.[4] Below are some of the primary IP rights that may be impacted:

- **Copyright**: Copyright protects original works of authorship, such as texts, images, music, and videos. When data scraping involves extracting copyrighted material without the rights holder's permission, this may constitute infringement. In some jurisdictions, copyright exceptions such as fair use or TDM may allow certain uses of copyrighted material without authorisation, but the scope of these exceptions varies widely. Copyright issues also arise with the removal of rights management information (RMI) (metadata) and circumvention of technological protection measures (TPMs) designed to prevent unauthorised access or copying of copyrighted works.
- **Database rights**: Some jurisdictions, for example in the EU, offer sui generis database rights, which protect the investment made in compiling databases (European Union, 1996). If data scraping involves extracting substantial parts of a database protected by these rights, it could violate the rights holder’s legal protections. These rights are distinct from copyright and cover the effort involved in the organisation and arrangement of data, even if the individual elements are not copyrighted.
- **Trademarks**: Scraping data that involves trademarked logos, names, or branding may lead to trademark infringement, especially if the scraped content is used in a commercial context that could imply endorsement or association by the trademark holder. This can result in consumer confusion or dilute the distinctiveness of the trademark (WIPO). Trademark laws vary across jurisdictions but generally protect the use of protected marks in commerce.
- **Trade secrets**: Trade secrets refer to confidential information that provides a business with a competitive advantage, such as proprietary algorithms, training datasets, and AI-related techniques used in the development or deployment of AI systems (WIPO, 2020). Trade secret protections often depend on “reasonable steps taken by the rightful holder of the information to keep it secret” (WIPO, 2024) Data scraping that involves accessing or disclosing such information without authorisation may constitute a violation of trade secret protections.
- **Publicity and likeness rights**: If scraped data includes personal likenesses, names, or voices of individuals, especially public figures, it may raise publicity rights issues. These rights, which protect

against the unauthorised commercial, and in some cases non-commercial, use of elements of a person's identity, can vary significantly depending on jurisdiction but are increasingly relevant in the context of AI-generated outputs that mimic real individuals.

- **Moral rights**: Moral rights allow individuals to claim authorship of their work and to object to modifications that could harm their honour or reputation. Recognised in many jurisdictions and harmonised under Article 6bis of the Berne Convention, moral rights protect the personal and reputational interests of creators (WIPO, 2024). Even when copyright ownership is transferred, moral rights may help protect against certain uses of a work that might harm the creator's reputation, such as modifications or distortions of their work in AI-generated content.

Data scraping of copyrighted materials raises questions as to whether the collection or use of the scraped data constitutes copyright infringement

The WTO Agreement on Trade-Related Aspects of Intellectual Property Rights (known as the 'TRIPS Agreement') and WIPO-administered treaties set forth certain international obligations with respect to IP, including copyright. Provisions relevant to copyright and data scraping are found in the TRIPS Agreement, the WIPO Copyright Treaty and the WIPO Performances and Phonograms Treaty (collectively known as the WIPO Internet Treaties), which were designed to update and supplement existing international treaties on copyright and related rights (the Berne and Rome Conventions). These provisions include:

1. The **prohibition against formalities**, meaning copyright protection is automatic without requiring actions like registration, an issue which some stakeholders raise in the context of "opt outs";
2. The **three-step test**, which sets forth when exceptions and limitations to copyright may apply. Under the three-step test, exceptions must:
   a. Be confined to special cases,
   b. Not conflict with the normal exploitations of the work, and
   c. Not unreasonably prejudice the legitimate interest of the rights holders
3. Protections for **rights management information (RMI)**, such as embedded metadata
4. **Protections to prevent the circumvention of technological measures (TMs),** such as passwords; and
5. Separate protections for **compilations of data** or other material.

A significant amount of scraped data may include copyrighted materials, raising questions as to whether the collection or use of the scraped data constitutes copyright infringement. Answering this question can require factual and legal analysis of various sub-questions, such as:

- Did the data scraping involve a reproduction or other activity that would implicate the exclusive rights under copyright and constitute a direct or indirect copyright infringement?
- Did the data scraping delete rights management information (RMI), such as metadata that tracks ownership and usage rights, or circumvent technical controls intended to prevent such scraping?
- Is there an applicable legal exception or defense that would permit an otherwise infringing activity ("copyright exceptions")?
- In which jurisdiction did the scraping activity take place, which jurisdictional laws apply, and in which jurisdiction or jurisdictions would it be deemed infringing of copyright, if at all?

The third-party use of copyright-protected material for use as training data may be legally permissible in certain circumstances, depending on the jurisdiction and specific facts of each case. For example, depending upon the circumstances, this may be allowed under different legal frameworks: the 'fair use' exception in the United States, the 'text and data mining' exception in the EU, the Copyright Act in Japan, the fair use provision in the Israeli copyright law and both fair use and text and data mining provisions in Singapore's Copyright Act. (US Government Publishing Office, 2010) (European Parliament, 2018) (Ministry of Justice, Japan, 2024) (Israeli Ministry of Justice, 2022) (Intellectual Property Office of Singapore, 2022).

However, the approach to answering these questions often varies across jurisdictions. For example, some jurisdictions, like the United States, provide a general statutory exception to copyright infringement (e.g. a fair use doctrine) that specifies various factors to be considered and often is interpreted by the courts. In jurisdictions with a fair use approach, the third question may involve, among other things, consideration of the AI-generated outputs to assess whether such outputs constitute "transformative uses" of the scraped data. In contrast, other jurisdictions such as the EU have different exceptions such as the EU's TDM exceptions to copyright protection. In these jurisdictions, the answer to this third question may turn on, among other things, the copyright holder affirmatively indicated a reservation of rights under the exception (opt-out mechanism). Meanwhile, other jurisdictions have a fair dealing or other different approaches.

Many jurisdictions have put in place explicit exceptions for TDM of copyrighted material, allowing automated data analysis to generate insights. However, the legal landscape varies widely, particularly in terms of the ability of copyright holders to opt-out. The scope and application of this carve-out differ among countries, leading to confusion and inconsistent legal interpretations across jurisdictions.

For example, in Japan, text and data mining (TDM) for both commercial and non-commercial purposes is authorised under the Copyright Act, provided the content is not used for enjoyment purposes. However, contractual terms or technological protection measures may override this exception. In contrast, in the EU, TDM for research purposes is permitted without the possibility for rights holders to opt out, whereas for other TDM activities, rights holders may reserve their rights by opting out, typically through contracts or machine-readable means.

Several jurisdictions do not have TDM exceptions and instead rely on concepts of fair use, fair dealing, or other measures to create limited exceptions or defenses against copyright infringement. For a summary of copyright exceptions in selected jurisdictions, see **Annex I.** These discrepancies between different legal regimes, including when dealing with data scraping for research or scientific purposes, create challenges in forming a coordinated approach to data scraping that extends across jurisdictions.

Data scraping litigation for alleged copyright infringement is growing rapidly around the world

Data scraping has been the subject of significant litigation and government enforcement for years, with high-profile cases involving various tech companies and AI developers (Neuburger , 2022) (Clarke, 2023). As the use of LLMs has rapidly expanded, new legal disputes have emerged, focusing on privacy, IP, and other issues. Investigations have been launched by multiple jurisdictions, including the United States and the EU, and task forces have been formed to address the challenges posed by AI-driven data scraping. For example, the European Data Protection Board (EDPB) has set up a ChatGPT task force and published its first report on the group's activities and findings (EDPB, 2024). Other data scraping litigation, arising principally in the United States, concentrates on IP, breach of contract, and tort claims.

In the United States, lawsuits have been filed by a range of stakeholders, including authors, news organisations, music publishers, and image providers (The Authors Guild, 2023) (Levi, Epstein, & Feirman, 2024). These cases generally center on whether the scraping and use of data to train AI models constitutes copyright infringement, breach of contract, or other violations of law. Some litigation also raises claims about AI-generated outputs that closely resemble the input data.

Litigation has also emerged outside the United States, with ongoing cases in Europe and other regions addressing the global impact of data scraping for AI purposes (Vincent, 2023) (WIPO, 2024). This suggests that the legal issues around data scraping will likely remain complex and evolving across multiple jurisdictions. While litigation may ultimately lead to desired outcomes factoring in interests of all affected stakeholders, it can be expensive, lengthy, and create uncertainty. Many aggrieved parties may also lack the resources to pursue litigation. This is another reason to consider alternate policy approaches, such as those outlined in Section 4.

Jurisdictional complexities in data scraping

Another legal issue raised by data scraping is the difficulty in determining the appropriate jurisdiction for resolving infringement claims. Data scraping activities often involve multiple jurisdictions—instructions to initiate scraping may be issued in one country, while the data subject to the scraping are located in another, and the data may ultimately be used, stored or processed in yet another jurisdiction. Additionally, the AI model trained on the scraped data may be used across a broader set of jurisdictions. This raises complex questions of private international law, particularly in determining where the infringement, if any, occurred and what laws should be applied. To address these challenges, it is important to consider the jurisdiction in which the scraping activity took place, as well as the jurisdictions where copyright infringement claims might be pursued.

The EU AI Act introduces requirements that have an international implications for data scraping practices by requiring providers of general-purpose AI models to comply with EU copyright laws, even if the model is trained outside the EU, as long as the output is used within the EU market (European Union, 2024). This regulation means that AI models trained outside the EU must comply with both the copyright laws of the jurisdiction where the training occurs and those of the EU if their outputs are used within the EU. This provision underscores the need for internationally coordinated approaches to copyright, particularly as AI systems increasingly operate across borders.

Contractual approaches play an increasing role in managing access to data and reducing or resolving IP rights disputes…

In navigating the jurisdictional complexities of data scraping, contractual agreements play an increasingly central role in managing access to data and reducing or resolving potential disputes around IP rights. The relationship between contractual agreements and IP rights plays a pivotal role in the legal landscape of AI data scraping. Contracts, particularly terms of service and end user license agreements, often help govern how data can be scraped or used as between the parties to the agreement, including specific provisions on permitted uses, attribution requirements, and liability allocations. However, the enforceability and interaction with IP laws can vary significantly across jurisdictions.

In many jurisdictions, contracts may be used to modify, as between the parties to the agreement, how data scraping might otherwise be permitted under applicable law. However, in some jurisdictions, such contractual terms may not always be enforceable. For example, in the EU, rights holders may impose contractual TDM limitations for commercial application. However, they are expressly prohibited from restricting TDM for scientific research purposes when conducted by research organisations and cultural heritage institutions.

Contractual terms may have different levels of enforceability depending on their context and the applicable laws. In some jurisdictions, terms in standard website agreements that users must accept to access content (such as terms of use) may be treated differently from individually negotiated agreements between parties. Questions often arise about whether standard form agreements (also known as ‘adhesion contracts’ where one party must accept the entire contract as presented) can restrict exceptions to copyright law, such as fair use or other permitted uses. For example, in Israel, an advisory opinion of the Ministry of Justice from 2016 indicated that such non-negotiable contracts may not be used to prevent users from claiming fair use defenses. Different considerations may apply to individually negotiated agreements.

Similarly, contracts may not always override certain moral rights, which are particularly strong in civil law jurisdictions such as France and Germany. Moral rights, such as the right of attribution or the right to prevent derogatory treatment of a work, typically remain with the original creator even if copyright has been transferred or licensed under contract. These rights can create additional layers of complexity in cases where scraped data is used to train AI models that generate outputs closely resembling the original creator's work. For example, an AI model trained on an artist's works might generate similar artistic styles, potentially raising moral rights concerns about attribution and integrity of the original works, even if the training data was obtained through valid contractual arrangements.

In this evolving landscape, stakeholders must carefully navigate the intersection of contract law and IP rights to ensure their contract terms are permissible across different applicable legal frameworks. Moving forward, standard contract terms and internationally coordinated policy approaches, as discussed in Section 5, may offer greater certainty for rights holders, data scrapers, and data users, while protecting the rights of all parties involved. Such coordination could address key issues including permitted uses, attribution requirements, territorial restrictions, and liability allocation, while protecting the rights of all parties involved.

… although policy responses – from codes of conduct to documentation and transparency requirements – vary significantly

Given growing data scraping disputes, policymakers are exploring a variety of policy solutions, including multilateral codes of conduct as discussed in Section 5. Initiatives are also occurring on a jurisdiction-specific basis, some of which are summarised in Table 1.

In response to these growing complexities, several jurisdictions have begun to address the challenges posed by data scraping in relation to IP rights. Table 1 below provides a summary of AI-specific IP initiatives in select jurisdictions to illustrate the varying approaches being adopted globally.

Table 1. Summary of AI-specific IP initiatives in select jurisdictions

Country	State of AI-specific initiatives	Details about initiatives
Canada	Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems (2023)	The voluntary code of conduct includes transparency provisions on training data (Government of Canada, 2023).Developers and managers of advanced generative systems commit to working to achieve the following outcomes: "[…]Transparency – Sufficient information is published to allow consumers to make informed decisions and for experts to evaluate whether risks have been adequately addressed."
European Union	EU AI Act has been enacted and implementation activities are underway; key copyright provisions are regulated by the Digital Single Market Directive (DSMD) (2024)	The EU AI Act introduces compliance obligations for providers of general-purpose AI models, requiring them to implement policies that align with EU copyright laws, including the use of state-of-the-art technologies to identify copyrighted works and comply with the TDM opt-out mechanism established in the DSMD. Providers of general-purpose AI models must also publicly disclose detailed summaries about the datasets used for training. To support implementation, the EU AI Office is developing a General-Purpose AI Code of Practice that will provide guidance on transparency and copyright-related obligations. (European Union, 2024)].
Israel	Ministerial Guidance (2023)	Ministry of Justice declared that the fair use exception is generally applicable to data scraping for AI except if the AI model's exclusive purpose is to imitate a single artist (State of Israel, Ministry of Justice, 2022).
Japan	New report (2024)	The interim report of the Study Group on Intellectual Property Rights in the Era of Artificial Intelligence states the need for a combination of laws, technology, and contracts to promote the advancement of AI technology and protection of intellectual property (Study Group on Intellectual Property Rights in the Era of AI, 2024)
Singapore	Copyright Act amended in 2021 to introduce a purpose-based exception for computational data analysis. The exception was most recently	In force since 21 November 2021, Singapore's computational data analysis exception expressly recognises the use of copyrighted materials for machine learning (Intellectual Property Office of Singapore, 2022). The most recent public consultation sought feedback on temporary exceptions to the prohibition against circumvention of technological

	considered in a public consultation that was conducted by the Ministry of Law and the Intellectual Property Office of Singapore from 22 April to 2 June 2024	measures for certain non-infringing uses of copyrighted materials (Ministry of Law and the Intellectual Property Office of Singapore , 2024) The computational data analysis exception was one of the non-infringing uses that was consulted on. Following the consultation, the Singapore Government decided against introducing a temporary exception to allow circumvention of technological measures for computational data analysis purposes, citing the need to ensure that, even with the computational data analysis exception in force, rights owners can continue to rely on such measures to control access to content used for machine learning purposes and obtain remuneration for such uses through charging for lawful access to the content.
United Kingdom	Plans to establish Voluntary Code of Practice (not in place)	The Working Group did not reach an agreement on an effective voluntary code of conduct as of February 2024. (Department for Science, Innovation & Technology, 2024). The United Kingdom government launched a consultation on copyright and artificial intelligence in December 2024 to explore solutions that support both creative industries and AI development (The Intellectual Property Office, 2024)
United States	Policy initiatives from US Copyright Office, US Patent and Trademark Office, EO Presidential Executive Order	United States Copyright Office requested public comment on AI and copyright (Copyright Office, 2023). The presidential AI Executive Order directs U.S. officials to develop an AI Action Plan aimed at promoting human flourishing, economic competitiveness, and national security (U.S. Presidential Executive Order, 2025). In addition, federal agencies are required to purchase AI systems that comply with IP laws.

Note: This table is for illustrative purposes and is not exhaustive.
Source: Sources indicated in-text.

In addition, style, likeness and publicity rights claims are emerging over AI-generated outputs

As large language models (LLMs) generate outputs influenced by their training data, questions have emerged regarding potential IP implications. The 2024 AI Index Report notes that popular LLMs have been found to generate content that mirrors the material on which they were trained, including copyrighted works such as passages from books or articles (Ray & Clark, 2024). Such instances raise concerns about whether such outputs could infringe upon existing copyrights or qualify as derivative works, particularly when substantial similarities to protected materials are observed (J. Gervais, 2022) (Henderson, et al., 2023).

Additionally, as AI-generated outputs can sometimes resemble the unique styles, voices, or likenesses of artists, writers, and other individuals, raising broader ethical and economic concerns. Many artists have voiced objections to AI models generating content in a "style" similar to their work, which may not directly copy the original but could nonetheless affect their livelihoods (Wakelee-Lynch, 2023) (Leffer, 2024) (Willman, 2023). The unauthorised use of individuals' voices and likenesses by virtual assistants has also raised privacy and consent issues, particularly in cases where AI-generated outputs closely resemble public figures (Mickle, 2024). Such cases illustrate the broader challenge of protecting creative and personal identity in the era of AI.

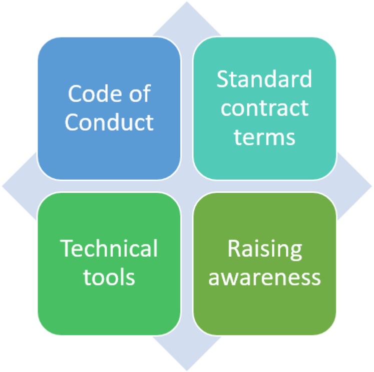
Protection for likeness, style, and publicity vary across jurisdictions and often fall under diverse legal frameworks such as trademark, copyright, rights of publicity, or privacy law. In the U.S., the scope and application of these rights are complex and jurisdiction-specific, and they are not always clearly categorised as traditional IP issues (U.S. Copyright Office, 2024). In the European Union, issues related to likeness and publicity are generally governed by moral rights and personality rights, which vary across EU member states. Under Article 6bis of the Berne Convention, moral rights allow creators to claim ownership of their work and to object to any modifications of their work that may harm their honour or reputation (Drexl, et al., 2021). However, the degree to which these rights are recognised and enforced differs significantly across EU member states (Hutukka, 2023). While this report primarily addresses IP considerations related to data scraping, these broader questions around AI-generated outputs highlight the growing challenges in IP law and beyond as AI technologies continue to develop.

4 Preliminary considerations and potential policy approaches

In the wake of AI advances, there are several areas where IP and other related laws concerning data scraping may need to be reviewed and potentially updated across various jurisdictions. Achieving this goal, however, may take time, particularly as policymakers strive to balance complex and potentially competing considerations across various legal fields. As this process continues, policymakers could consider developing flexible and voluntary measures that can accommodate diverse legal and regulatory approaches across jurisdictions. These measures, as highlighted in the EASD Recommendation (OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data, 2021), include:

- Preparing and adopting a cross-border "data scraping code of conduct";
- Supporting the development of a shared understanding of terms used in AI data scraping that can be referenced in the data scraping code or contract terms;
- Supporting the development of standardised and widely accessible technical tools that can be referenced in the "data scraping code of conduct"; and standard contract terms; and,
- Implementing initiatives to raise awareness among stakeholders about their rights and responsibilities.

Figure 2. Potential policy approaches to address IP related challenges in AI data scraping



Note: This figure is for illustrative purposes and is not exhaustive.
Source: OECD.AI

These potential approaches could be taken into consideration by policymakers when considering changes to applicable laws. These approaches could also potentially provide policymakers with flexibility to holistically address IP, privacy, and other issues presented by data scraping in an internationally coordinated manner while accounting for relevant limitations and exceptions and other legal and policy differences across jurisdictions. Such policy tools could be designed to adapt and evolve over time.

The process of developing these policy tools could help policymakers address core issues by working towards recognising important terminology, creating standardised definitions and measures that support innovation with the effective protection of IP rights, in accordance with the OECD AI Principles and other relevant multilateral efforts, such as the G7 Hiroshima AI Process (Government of Japan, 2024).

A voluntary code of conduct to help address issues posed by data scraping

Policymakers are increasingly turning to codes of conduct and other forms of voluntary commitments by business to address challenges such as data scraping. For example, the G7 adopted a voluntary code of conduct that underscores the need for "appropriate data input measures and protections for personal data and intellectual property (G7, 2023)". A voluntary "data scraping code of conduct" could build upon and align with the G7 code of conduct by setting forth some specific measures and protections referenced in the G7 code. This section examines potential elements that may be included in such a code of conduct, should it be pursued in relevant fora.

The OECD Guidelines for Multinational Enterprises on Responsible Business Conduct (RBC Guidelines) [OECD/LEGAL/0144] represent a voluntary, government backed framework for expectations on business behaviour covering all areas of business ethics (e.g., human and labour rights, environmental impacts, and consumer protection). The RBC Guidelines also note the expectation by governments that companies put in place safeguards to prevent and mitigate adverse impacts linked to granting licenses for the use of intellectual property rights or when otherwise voluntarily transferring technology (OECD, 2018). More specifically, the EASD Recommendation [OECD/LEGAL/0463], which provides a holistic data governance approach for data access and sharing, including for AI, calls on Adherences to "promote, where appropriate, self- or co-regulation mechanisms" (OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data, 2021). These mechanisms can include "voluntary guidance, codes of conduct and templates for data access and sharing agreements – that provide legal flexibility while ensuring that all relevant stakeholders have certainty as to applicable laws and regulations". The European Commission is also working on developing a General-Purpose AI Code of Practice to help implement the requirements of the EU AI Act, ahead of formal standards (European Commission, 2024).

A voluntary "data scraping code of conduct" could be drafted in a way that allows parties to adopt them without having to make any admissions or take positions about the types of activities that may be infringing or otherwise violate applicable law. This approach could encourage greater adoption and has been used successfully in other contexts. In 2007, several stakeholders adopted Principles for User Generated Content Services that made clear that adopters of such principles were not making any legal admissions (UGC Principles, 2007). This same approach could be used in a "data scraping code of conduct". The International Federation of Reproduction Rights Organisations has a code of conduct for Reproduction Rights Organisations (RROs) and guidelines for relationships between RROs (IFFRO, 2024) (IFFRO, 2024). The WIPO also has a Toolkit for Collective Management Organisations that includes a code of conduct (WIPO, 2021).

A "data scraping code of conduct" could assist policymakers in several ways. With a degree of flexibility and swiftness, governments could endorse the code and create a registry of organisations that have adopted the code. This registry would help governments monitor adherence to the code. Failures to comply with commitments to abide by the "data scraping code of conduct" could be voluntarily reported in the registry, the OECD's AI Incidents Monitor, or in a similar database (OECD.AI, 2024). To the extent desired, some jurisdictions may elect to treat non-compliance with voluntary commitments as a legal violation of applicable law or may choose to make portions of the code mandatory.

The code could also include provisions addressing circumstances when a person or entity violates the code, including remediation measures, or when an entity chooses to stop complying with the code. For example, efforts are underway to develop machine learning techniques aimed at removing AI training data that does not comply with applicable laws or policies (Achille, Kearns, Klingenberg, & Soatto, 2023). The code could address the use of techniques for helping to ensure compliance with IP rights or for detecting and mitigating, in cases where data does not meet legal or policy standards.

A voluntary “data scraping code of conduct” could also outline terms for AI system end-user agreements to prevent circumvention of safeguards and controls. These provisions could help direct AI system end-users to avoid certain prompts or other conduct likely to lead to IP infringement or other harmful outcomes.

It could include broadly applicable provisions and provisions directed to specific actors…

A “data scraping code of conduct” could potentially address an array of behaviors and actors. Portions of the code could apply to everyone. For example, the code could establish standard definitions, similar to those included in the OECD Framework for the Classification of AI Systems (OECD, 2022). This would help address uncertainties created by the current lack of standard definitions for “data scraping” and the various activities it potentially encompasses, as described in Section 1 above.

Additionally, the code could include broadly applicable data scraping principles, modeled after the OECD AI Principles. These principles would encourage responsible innovation by safeguarding privacy, IP, and other rights. The code could also discourage the development and use of websites with pirated data. Furthermore, it could provide a high-level framework for addressing payments in appropriate circumstances. This framework could be developed, taking into consideration existing collective management mechanisms and other IP payment frameworks. Similarly, the code could also establish a registry and processes for reporting code violations, as mentioned above.

To expand upon this framework, the code could promote the development of standard contract terms, technical tools, and awareness raising initiatives, each of which are discussed below. Standardising these practices could help address concerns from policymakers, especially regarding changes to terms of service and privacy policies that allow the collection of more AI training data without individuals truly understanding these practices or the changes (Federal Trade Commission, 2024).

In addition to broadly applicable terms, a voluntary “data scraping code of conduct’’ could have sections directed to specific activities, such as (a) collection, pre-processing and storage activities associated with data scraping, (b) the aggregation of scraped data, and (c) the use of scraped data in connection with AI systems. Organisations could agree to comply with the code’s general provisions as well as all relevant provisions related to their activities.

…as well as specific provisions for collection, pre-processing and storage activities in connection with AI data scraping

A voluntary “data scraping code of conduct’’ could include provisions specifically directed to the collection, pre-processing and storage of data through scraping. For instance, it could contain commitments to comply with relevant contracts and laws, to preserve copyright management and similar information embedded in scraped data, and to avoid circumventing technical safeguards.

It could also include clauses about documentation and transparency. For instance, adhering organisations could agree to maintain and/or disclose certain information about their training data.

Transparency around data scraping practices is currently limited. The European Union has addressed this issue for general-purpose AI models through the EU AI Act, which requires providers to "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model" (European Union, 2024). To support this, the EU AI Office will develop a template for making such disclosures. The General-Purpose AI (GPAI) Code of Practice complements this requirement as it will detail rules for GPAI model providers on how to comply with transparency and copyright-related obligations under the EU AI Act (European Commission, 2024). While different jurisdictions may take varying approaches to transparency, from formal requirements to voluntary guidelines or other flexible frameworks, coordination of common practices could streamline compliance for both AI developers and rights holders while maintaining flexibility across different operational and regulatory contexts. The "data scraping code of conduct" could be aligned with this template and the EU GPAI Code of Practice to help facilitate a coordinated, cross-jurisdictional approach to transparency. This could streamline compliance for AI developers and potentially simplify protection for rights holders by coordinating best practices for data management.

Since data scraping encompasses a range of activities, a “data scraping code of conduct’’ could outline preferred practices that align with policy objectives. Specifically, it could differentiate among the various techniques for data collection, data processing, and data storage, as discussed above. For instance, as discussed in Section 1, some AI data aggregators utilise links to make AI input data available, rather than relying primarily on copies. To the extent desirable, a “data scraping code of conduct’’ could address such practices.

The standard definitions included in the code could help streamline its implementation and ensure consistency, while leaving room for flexibility to accommodate different legal and regulatory contexts.

And specific provisions for AI data aggregators

The “data scraping code of conduct” could also include provisions for AI data aggregators, including their compliance, documentation, and transparency activities. For instance, the code could provide an opportunity to help rein in practices of making pirated data available on AI data aggregator sites. It also might address the practice of linking to such data. One key aspect could be to implement effective control mechanisms that enable rights holders to manage and safeguard their data, ensuring compliance with IP rights. Data aggregators could be encouraged to adopt standardised transparency tools, such as dataset cards or provenance documentation, to provide clear information about datasets, including licensing terms and any restrictions related to data usage. By offering dataset creators the ability to restrict or approve access to IP-protected data, platforms can play a key role in ensuring lawful use and better transparency.

For example, platforms like Hugging Face have implemented measures such as dataset cards that detail the licensing, provenance, and intended use of the dataset, making it easier to control the re-use of licensed or protected datasets hosted on its platform by third parties (Hugging Face, 2024). In turn, for third parties, these cards enable them to understand their legal obligations and restrictions. Additionally, gated datasets, which restrict access to datasets until specific information is provided or approval is given help rights holders control who can access their datasets (Hugging Face, 2024). A similar approach could be incorporated into the “data scraping code of conduct”, ensuring that aggregators provide transparency about the provenance of data, including whether any IP-protected data is involved, and enforce appropriate permissions for its use.

The code could also address how remediation mechanisms should be implemented when data scraping activities result in violations of IP and possibly other rights. To the extent appropriate, the code may differentiate among commercial, government, and non-commercial actors and use cases.

The development of the code should include a broad range of stakeholders to ensure it promotes fair competition and innovation. This includes rights holders, both established AI data aggregators and new market entrants from commercial, non-commercial and other sectors. Such inclusive participation can help prevent the code from inadvertently creating barriers to entry.

But also transparency requirements for users of scraped data

The “data scraping code of conduct” could seek commitments directed to LLM developers and other users of scraped data (AI operators). These commitments could address aspects unique to their role in the AI development ecosystem, while building upon the transparency frameworks discussed above. AI operators could be encouraged to document the provenance of datasets and any associated rights or restrictions. This could allow developers to trace the source of data and ensure they comply with IP rights.

To the extent applicable, the “data scraping code of conduct’’ could also consider how these commitments might vary among commercial, non-commercial and government actors and use cases. The following are examples of other topics that could be addressed in this portion of the code.

Table 2. Preliminary terms for IP related issues in a potential data scraping code of conduct for AI developers and operators

Topic	Potential principles for the code of conduct
Data acquisition	AI operators would commit to obtain scraped data only from entities and individuals who have agreed to, and abided by, the data scraping code. The registry and reporting mechanisms, discussed above, could help AI Operators comply with this commitment.
Compliance	AI operators could commit to comply with applicable laws and contracts and not alter or delete any rights management or similar information embedded in data before it is scraped. The code could also include specific provisions directed to privacy and cybersecurity compliance and practices (addressed in a separate report).
Ethical considerations	AI operators could be expected to avoid scraping or using scraped data in unlawful ways. Policymakers may choose to address certain ethical considerations as well.
Use limitations (research and non-commercial uses)	These terms could be addressed in a similar fashion, as described above in sub-section “A voluntary code of conduct to help address issues posed by data scraping” on page 29.
Other Restrictions	The code could address other possible use restrictions, including whether and to what extent scraped data can be further distributed (and to whom). To the extent desirable, the code could differentiate between using scraped data for training, fine-tuning, and/or other purposes.
Documentation and transparency	AI operators could commit to appropriate documentation and disclosures practices about their AI training data, notably scraped data. For example, AI operators could consider periodically disclosing in in standardised formats, how scraped data is used, stored, processed and shared.
Technical safeguards	AI operators could commit to techniques to prevent trained AI models from producing certain types of AI-generated outputs, without appropriate consents, such as AI-generated outputs that are (i) substantially or confusingly similar to the scraped data, (ii) mimic or substantially replicate a person’s name, image, voice, or likeness. Technical measures could also be used to help identify AI-generated outputs.
Non-IP issues	The code could provide guidance on non-IP issues (discussed elsewhere) such as privacy as well as on cybersecurity practices and to document and make certain disclosures about such practices. The code could also address remediation techniques to be used, as discussed above, when data scraping violates (or leads to violations of) appliable laws or policies.
End user agreements and behaviours	AI operators could agree to use reasonable efforts to educate users about these prohibitions as well as appropriate uses of their AI systems. AI operators could use reasonable efforts to track and appropriately address violators of these terms. The code could address these commitments in more detail.

Note: This table is for illustrative purposes and is not exhaustive.
Source: OECD.AI

Technical tools that protect IP rights, enable rights holders to control access to their data more easily, and that support licensing mechanisms can be encouraged

Policymakers can encourage the development of standard and widely accessible technical tools that protect IP rights, enable rights holders to control access to their data more easily, and that support licensing mechanisms. The EASD Recommendation [OECD/LEGAL/0463] calls on Adherents to foster the adoption of these technical tools “including data access control mechanisms and privacy enhancing technologies, through which data can be accessed and shared in a safe and secure way between approved users, combined with legally binding and enforceable obligations to protect the rights and interests of data subjects and other stakeholders” (OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data, 2021). These tools could potentially build upon ongoing work by various standards organisations, (e.g. the W3C TDM Reservation Protocol).

A frequently mentioned technical tool is the robots.txt protocol, which is widely used to inform web crawlers about which parts of a website should not be scraped. It has long served as mechanism for limiting data scraping, but it may not necessarily be legally enforceable or technically binding, depending on the facts and circumstances. Recent research suggests that since mid-2023, many web domains have adopted robots.txt files to address AI-related data scraping (Shayne Longpre, 2024) However, significant inconsistencies exist in their application. For example, web crawlers from well-known developers of advanced AI systems are frequently restricted, while lesser-known entities often bypass such restrictions. In addition, certain sectors, such as news platforms, forums, and social media, are more likely to use these safeguards than personal or small e-commerce websites. There is also a disconnect between the restrictions stated in website terms of service and the actual technical measures in place, as many websites do not correctly configure their robots.txt files to reflect contractual restrictions.

In response, policymakers can encourage the development of more enforceable tools for data access control, as robots.txt's voluntary nature limits its effectiveness in preventing unauthorised scraping.

Rights management safeguards and transparency requirements and tools could be further developed in many jurisdictions via standardised tools and mechanisms

As outlined in Section 4, the EU allows rights holders to opt-out of text and data mining for commercial purposes under specific conditions. Developing standardised opt-out tools and other mechanisms (including standard mechanisms to track how scraped data is used) could make it easier for rights holders to protect their rights. It could also streamline compliance for organisations collecting or using data via data scraping. Some efforts are already underway to promote the development of new opt-out mechanisms (Keller, 2024). Any such developments will need to carefully balance the interests of rights holders, data users, and the broader public interest in AI innovation.

Appropriately standardising these tools could provide other benefits as well. It could offer an opportunity to holistically address IP as well as privacy and data protection concerns. Developing new technical tools could also help standardise transparency practices pertaining to AI training data, further incentivising responsible AI data sharing practices.

Table 3. Preliminary AI operator IP related tools for a potential data scraping code of conduct

Additional potential tool functionality	Needs such tool might address
Conditional access tool	Conditioned data access refers tools that permit data access or sharing subject to terms that may include limitations on the users authorised to access the data (discriminatory arrangements), conditions for data use including the purposes for which the data can be used, and requirements on data access control mechanisms through which data access is granted. Policymakers might encourage stakeholders to explore tools that enable rights holders to easily communicate conditions for access to data for data scraping purposes across multiple platforms or by multiple data scraping tools.
Automated contracting	Contracts can make commitments and understandings enforceable. In addition to developing standard contract terms, as discussed below, policymakers could encourage the development of technical tools that facilitate automated contracting and automated monitoring of contract compliance. This could make it easier for rights holders to seek redress should the need arise, particularly if coupled with appropriate processes for addressing disputes. However, it is important to include human oversight and intervention, particularly for complex or high-stakes disputes, to ensure access to justice, especially for vulnerable groups. If rights holders have reliable contracting and enforcement mechanisms, they may be more willing to opt-in or grant conditional opt-in consents. This might draw upon efforts to develop the W3C TDM Rights Reservation Protocol.
Direct payment mechanisms	To help make transactions potentially more efficient, policymakers might want to explore whether mechanisms could be used or developed to facilitate direct payments to rights holders, including potentially for the initial use and possibly permitted downstream uses. Policymakers can also look at what incentives may help encourage development of solutions in this space.
Monitoring compliance with data scraping code and contracts	Policymakers may want to encourage the development of tools that can help monitor compliance with a data scraping code and contractual compliance.

Note: This table is for illustrative purposes and is not exhaustive.
Source: OECD.AI

Standard contract terms could help chart a responsible path for data scraping

The development of appropriate standard contract terms requires careful consideration of a number of factors and could greatly benefit from the establishment of a global multi-stakeholder process that would allow different viewpoints to be taken into account. Having a common understanding of appropriate terminology to facilitate the development of appropriate standard contract terms could be a necessary first step in developing these contract terms. In this process, it would be helpful to consider technical, legal and/or other terms that may already exist in different jurisdictions. The need for common terminology is particularly important given the cross-border nature of AI and the territorial nature of IP rights, and the fact that legal terms may have significantly different meanings across jurisdictions. For example, the concept of "enjoyment purposes" plays a specific role in Japanese copyright law, while terms such as "compensation" and "remuneration" may be interpreted differently in various legal systems.

As noted above, policymakers in the United States, United Kingdom, and Japan, have highlighted the potential for contracts to help chart a responsible path for data scraping. The Global Partnership on AI (GPAI) IP Advisory Committee project on responsible AI data and model sharing supports this too, particularly when standard contract terms are complemented with codes of conduct, technical tools, and education (Tiedrich, Lee; Avdulla, Alban;, 2023). Given the diversity of stakeholders and the varying approaches across different fields and jurisdictions, it may be desirable to develop a range of different standards contract terms to accommodate different needs and bargaining positions. These terms could be tailored to address various use cases from non-profit research to commercial applications. Importantly, these standard contract terms could serve as an optional starting point and would not obviate the right of organisations to negotiate bespoke arrangements, when appropriate.

Given the usefulness of Standard Contract Clauses (SCCs) in other contexts, this approach seems promising. SCCs have helped manage cross-border data flows among countries with different privacy laws. Standard open-source and Creative Commons licenses have also helped unlock innovation and opportunities for small and medium-sized enterprises (Creative Commons, 2024).

Many companies are also turning to contracts to help manage data rights, particularly for AI training data (OECD, Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, 2019). Major platforms such as The New York Times, X, Zoom, Instacart, have updated their terms of service to specifically address AI data scraping (Ostwal, 2023) (Mehta, 2023) (Heath & Fried, 2023). Additionally, AI companies are forming direct partnerships with content providers: OpenAI has established agreements with AP News and Shutterstock, News Corp, and the Financial Times (O'Brien, 2023) (Reuters, 2024) (Shutterstock, 2023), while Google has entered an agreement with Reddit (Tong, Wang, & Coulter, 2024). The scope of these contracts is not public but may include copyright licenses, access to data, and/or other aspects related to AI training data acquisition,

The importance of these contractual approaches is further highlighted by the Statement on AI Training, signed by over 6,500 creators. This movement further underscores the growing importance of license agreements as a tool to protect creators' rights while allowing the use of their works in AI training (Statement on AI training, 2024). Implementing standard contract terms could have multiple benefits: they could help prevent unfair treatment of individuals and smaller and medium-sized enterprises (SMEs) as recognised in the EASD Recommendation (OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data, 2021). Standard contract terms could also reference the “data scraping code of conduct” and/or technical tools.

Finally, standard contract terms can be drafted with policy objectives in mind, helping to translate policies into practice. Policymakers also could express their views about making the standard contract terms fair and reasonable in a way that addresses significant inequities in bargaining power. As highlighted in the GPAI IP Advisory Committee work, developing standard contract terms will significantly benefit from multi-stakeholder collaboration. Multi-stakeholder collaboration on developing the standard contract terms may also lead to greater adoption and practical application across jurisdictions.

Raising awareness about IP issues and data scraping also plays a critical role

Raising awareness about data scraping, and its legal and societal implications, also plays a critical role, in part by facilitating and encouraging responsible behavior. Rights holders, data subjects, data producers, data holders and data users should be empowered with information on how they can protect and manage their rights. For example, rights holders may benefit by having more information about the current legal landscape across jurisdictions pertaining to their rights and emerging policy measures that may help them protect their rights. Equally important, other stakeholders in the AI data ecosystem should understand the IP laws and policies pertaining to their data-related activities.

Additionally, users of AI systems also need education on how they can help protect rights holders. This could include directions on how to avoid prompts that are likely to circumvent technical safeguards or violate rights. AI operators could develop short videos or other content explaining to users how their systems should be used. As mentioned above, AI operators also could incorporate these restrictions into their end user agreements.

Raising stakeholder awareness also builds upon other OECD policy recommendations. In line with the OECD AI Principles highlight the importance of transparency and explainability of AI systems as outlined in Principle 1.3: “AI actors are encouraged to commit to transparency and responsible disclosure concerning AI systems. This commitment involves providing meaningful information, appropriate to the context and consistent with the state of the art, to: i) foster a general understanding of AI systems, their capabilities, and limitations; ii) make stakeholders aware of their interactions with AI systems, including in the workplace; iii) where feasible and useful, provide plain and easy-to-understand information on the sources of data/input, factors, processes, and/or logic that led to the predictions, content, recommendations, or decisions made by AI systems, thereby enabling those affected to understand the output; and iv) offer information that enables those adversely affected by an AI system to challenge its output" (OECD, 2024).

The EASD Recommendation also highlights that "Adherents should strive to ensure that stakeholders are fully informed as to their rights (including their right to information and to obtain redress), responsibilities and respective liabilities in case of violations of privacy, intellectual property rights, competition laws, or other rights and obligations" (OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data, 2021) It further calls on Adherents to "promote the development of the data-related skills and competencies needed, including by workers and public servants, to harness the benefits of data access, sharing."

Annex A. Selected copyright exceptions in different jurisdictions

Text and data mining jurisdictions

In **Japan**, under the Copyright Act, TDM is authorised broadly for both commercial and non-commercial uses, particularly for "non-enjoyment purposes" (i.e., when the content is not meant to satisfy intellectual or emotional needs) (Ueno, 2021). However, technological protection measures (TPMs) or contractual terms can override these exceptions. This exception allows for the use of copyrighted works for AI training, thereby facilitating innovation in fields like machine learning. Japan has published 2 reports on the issue of AI and IP, “the General Understanding on AI and Copyright in Japan” (Published by the Legal Subcommittee under the Copyright Subdivision of the Cultural Council) and interim report on AI and IP (Study Group established by the Secretariat of the Intellectual Property Strategy Headquarters). These reports maintain the current TDM regime, while seeking to clarify the application of legal rules. In addition, they encourage businesses to adopt technical tools to reduce the risk of IP infringements, and to obtain licenses remunerating creators even for uses allowed under the Copyright Act (Japan Copyright Office, and Secretariat of Intellectual Property, 2024).

In the **United Kingdom**, the Copyright Designs and Patents Act of 1988 permits TDM solely for non-commercial research (e.g., the non-commercial carve-out) or with the rights holder’s permission. Contractual overrides are not authorised for the Non-Commercial Carve-Out. Additionally, the United Kingdom’s Intellectual Property Office (IPO) had established a working group to develop a voluntary code of practice on copyright and AI, aiming to address barriers for AI firms/users and protect rights holders, including potentially through the use of contractual arrangements and a code of practice. However, consensus on code of practice or other approaches on AI and copyright have not been reached, and future proposals, potentially including legislation, may emerge (Department for Science, Innovation & Technology, 2024).

The **European Union** has two key instruments to address copyright issues related to AI: the Directive on copyright and related rights in the Digital Single Market (Directive (EU) 2019/790, “DSMD”) and the European Union Artificial Intelligence Act (the ‘EU AI Act’) (European Union, 2024).

The DSMD introduces two exceptions that allow the use of protected works for TDM activities under certain conditions:

- Article 4(3) of the DSMD allows TDM of lawfully accessible works for both commercial and non-commercial purposes, provided the entity has legal access (e.g., through licenses or public availability). However, rights holders can opt out of this usage via contracts, declarations, machine-readable means, or terms and conditions. This opt-out mechanism only applies if rights holders explicitly reserve their rights against such usage.
- Art. 3 introduces Non-Commercial Carve-Out to the general TDM opt-out regime. The EU Non-Commercial Carve-Out is specifically tailored for scientific research. This provision allows research organisations and cultural heritage institutions to conduct TDM for scientific purposes, covering both natural and human sciences, provided they have lawful access to the works or subject matter.

> Unlike the exception set out in Article 4 DSMD, copyright holders cannot override or opt-out of this exception by any contractual or technical restrictions.

As the DSMD is a European Union directive, not a regulation, it does not apply directly in European Union member states. Member states must enact laws implementing the directive into their national law. Hence, although harmonised in their purpose, the national laws vary in their implementation of the opt-out mechanism. As a result, there is no uniform way to opt-out. While some member states have made machine-readable opt-outs mandatory, others have made them optional, and some do not address this method of communication at all (Nobre, 2024). Additionally, there is inconsistency in how member states interpret key terms such as ‘research’ purposes and apply the conditions for the exceptions like the non-commercial carve out. This lack of full harmonisation sometimes leads to inconsistent legal interpretations and challenges in applying these exceptions uniformly across the EU (Hutukka, 2023) (European Commission, 2024) . The European Commission’s Directorate-General for Communications Networks, Content and Technology is considering updating the TDM opt-out mechanism to provide a standardised, machine-readable way to better empower rights holders to prevent the use of copyrighted works to train AI models.

The EU AI Act complements the DSMD by focusing on the responsibilities of providers of general-purpose AI models. The EU AI Act requires these providers to implement a policy to comply with EU copyright and related laws. This includes using state-of-the-art technologies to identify copyrighted works and comply with the TDM opt-out mechanism established in Article 4(3) of the DSMD.

Additionally, providers are required to “draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model” (Art. 53(1)(d) AIA). This summary, which will be publicly disclosed, is expected to enable rights holders to effectively exercise their rights. The GPAI providers will have to list “the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used” (Recital 107). The EU AI Office will develop a template for making such disclosures.

A notable aspect of the EU AI Act is its extraterritorial reach. It requires providers of general-purpose AI models to comply with EU copyright laws, even if the model is trained outside the EU, as long as the output is used in the EU market (Art.2(1)(c)). This means that general purpose AI models trained outside the EU will need to comply with both copyright laws of the jurisdictions where the training occurs and EU copyright laws if the models are used in the EU. This provision underscores the importance of coordinated copyright approaches across jurisdictions and introduces new compliance requirements for AI providers serving the EU market.

Singapore's Copyright Act of 2021 introduced a purpose-based exception permitting the copying of copyrighted materials for computational data analysis purposes, applicable to both commercial and non-commercial use). The exception is confined to specific acts (copying and, in narrow circumstances, communication) and can be used only for the purpose of computational data analysis. It expressly restricts the use of copyright material for any other purpose, failing which the exception will not apply to the original use for computational data analysis (Section 244(2)(b) of the Copyright Act).

This exception applies regardless of any contract terms that might otherwise restrict such activities. However, conditions are in place to protect the commercial interests of copyright owners, such as restrictions on sharing copies and limitations on the use of copied works for specific purposes.

Additionally, TDM activities may also be exempted under Singapore's general 'fair use' exception (formerly known as the 'fair dealing' exception). The Copyright Act makes it clear that the computational data analysis exception and the fair use exception operate independently of one another (Section 184 of the Copyright Act 2021), meaning that TDM activities can potentially qualify under either or both exceptions, assuming the conditions of each are met.

Singapore's Model AI Governance Framework for Generative AI provides a basis for global conversation to address generative AI concerns while maximising the space for continued innovation. It identifies the importance of balancing copyright with data accessibility and encourages policymakers to foster open dialogue amongst all relevant stakeholders to understand the impact of the fast-evolving generative AI technology and ensure that potential solutions are balanced and in line with market realities (AI Verify Foundation; Infocomm Media Development Authority, 2024[55]).

Finally, Singapore recently introduced a new regulatory regime for collective management, which promotes more effective and transparent licensing solutions (Intellectual Property Office of Singapore, 2024).

Fair use and fair dealing

Australia's Copyright Law sets out specific 'fair dealing' exceptions for the purposes of: research or study; criticism or review; parody or satire; reporting the news; and provision of legal advice/ a legal practitioner, registered patent attorney or registered trademarks attorney giving professional advice (Australian Government, 1968).

These exceptions do not apply to all types of copyright material. The Copyright Act provides that 'fair dealings' for these specified purposes may be made with the following copyright material: literary, dramatic, musical or artistic works; adaptations of literary; dramatic or musical works; and audio-visual items. Where the use of a 'substantial part' or more of the work, adaptation, or audio-visual item constitutes a 'fair dealing', there is no infringement of the copyright in that specific copyright material.

Canada's Copyright Act sets out two exceptions to copyright infringement could potentially apply in the context of data scraping: 1) the fair dealing exception for research (section 29); and, 2) the exception for temporary reproductions for technological processes in (section 30.71). In a lawsuit involving a text and image web-crawler, a Canadian court applied fair dealing (Century 21 Canada Limited Partnership v Rogers Communications Inc., 2011 BCSC 1196). In that case, the crawler gathered text and photos from websites to populate the defendant's own website. The court found the activities infringing but uncertainty remains as to whether the same analysis would apply to other types of data scraping activities conducted for different purposes. As for the exception for technological processes, the Copyright Board of Canada interpreted this provision as "intended to capture copies that happen automatically, or without the direct control of the user", and that are automatically deleted once the technological process is completed (Copyright Board of Canada, 2016).

As in the case of fair dealing, there is uncertainty as to whether and to what extent this exception would apply to TDM activities. For example, while some TDM activities may require making ephemeral copies, others may require copies of works to be stored indefinitely, which would make this provision inapplicable. In October 2023, the federal government launched a consultation on copyright and generative AI (Government of Canada, 2023). One of the purposes of the consultation was to determine whether the Copyright Act should be amended to clarify its application to TDM activities, and how existing exceptions for fair dealing for research and temporary reproductions for technological processes might be adapted to TDM. Following the closure of the consultation period, updates from the federal government remain pending.

Israel's Copyright Act adopted the fair use doctrine modelled after United States law. The Israeli Ministry of Justice issued an opinion confirming that its fair use standard largely covers data scraping for training AI models except if the AI model's exclusive purpose is to imitate a single artist. Also, the opinion does not apply to the output of the machine learning process which might potentially violate copyright even if the machine learning process itself was not infringing (Ministry of Justice, 2022).

Korea likewise incorporated fair use modelled on the United States law into its Copyright Act (Republic of Korea, 2017).

The **United States** Copyright Act includes a fair use exception that allows for limited use of a work without the copyright holder's consent "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research" (17 U.S. Code §107) - . To determine whether the fair use exception applies in a copyright infringement lawsuit, the statue enumerates four factors for courts to consider: :

- the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;
- the nature of the copyrighted work;
- the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and;
- the effect of the use upon the potential market for or value of the copyrighted work.

Substantial United States litigation has emerged in the AI context concerning copyright and other claims, including whether data scraping falls within the statutory fair use exception. Fair use decisions focus on the underlying facts presented in each case.

The Digital Millennium Copyright Act (DMCA) also is applicable to AI data scraping. For instance, the DMCA can impose criminal and civil liability on persons who circumvent a "technological measure that effectively controls access to a copyrighted work by "avoid[ing], bypass[ing], remov[ing], deactivat[ing], or impair[ing] a technological measure, without the authority of the copyright owner" (17 USC 1201(a)). The DMCA thus creates a regime that is somewhat similar to opt-out mechanisms in TDM laws, as discussed above. Following the 2021 triennial rulemaking process required by the DMCA to determine the need for temporary exemptions to 17 USC 1201(a)(1)(A) (prohibition of circumvention of technological measures that effectively control access to copyrighted works), there is currently a temporary exemption for certain TDM activities involving scholarly research and teaching (37 CFR 201.40(b)(4), (5)). United States copyright laws also prohibit the intentional removal of copyright management information (CMI) as well as knowingly distributing or importing works after the CMI has been removed (17 U.S. Code § 1202).

References

Schaul, K., Chen , S., & Tiku , N. (2024, April 19). *Inside the secret list of websites that make AI like ChatGPT sound smart*. Retrieved May 27, 2024, from The Washington Post: https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/?tid=wa_softregwall_auth

Achille, A., Kearns, M., Klingenberg, C., & Soatto, S. (2023). AI Model Disgorgement: Methods and Choices. Retrieved from https://arxiv.org/pdf/2304.03545

Australian Government. (1968). Copyright Act 1968. Retrieved from https://www.legislation.gov.au/C1968A00063/2019-01-01/text

B. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. Retrieved from https://arxiv.org/abs/2005.14165

Baack, S. (2024). A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3630106.3659033

Bommassani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., & Liang, P. (2024). The Foundation Model Transparency Index v1.1. Retrieved from https://crfm.stanford.edu/fmti/paper.pdf

Chason, R. (2024). *With French under fire, Mali uses AI to bring local language to students*. Retrieved from Washington Post : https://www.washingtonpost.com/world/2024/04/13/mali-books-artificial-intelligence-ai/

Chen, P., Wu, L., & Wang, L. (2023). AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. *Applied Sciences*. Retrieved from https://www.mdpi.com/2076-3417/13/18/10258

Clark, J., & Perrault, R. (2022). *AI Index Report 2022.* Retrieved from https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf

Clarke, L. (2023, October). *An AI firm harvested billions of photos without consent. Britain is powerless to act*. Retrieved May 2024, from Politico: https://www.politico.eu/article/ai-ruling-obstruct-british-efforts-protect-citizens-images-us-data-harvesting/

Common Crawl. (2024). *Common Crawl*. Retrieved May 27, 2024, from https://commoncrawl.org/

Copyright Board of Canada. (2016). Re: Sound, CSI, Connect/SOPROQ, Artisti – Tariff for Commercial Radio. Retrieved from https://decisions.cb-cda.gc.ca/cb-cda/decisions/en/366778/1/document.do

Copyright Office. (2023). *Artificial Intelligence and Copyright.* Retrieved from https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright

Creative Commons. (2024). *About CC licenses*. Retrieved from https://creativecommons.org/share-your-work/cclicenses/

Data Provenance. (2023, May 2027). The Data Provenance Initiative: A Large Scale Audit of Dataset

Licensing & Attribution in AI. Retrieved May 27, 2024, from https://arxiv.org/abs/2310.16787v3

Department for Science, Innovation & Technology. (2024, February). *A pro-innovation approach: Government response to consultation.* Retrieved from https://assets.publishing.service.gov.uk/media/65c1e399c43191000d1a45f4/a-pro-innovation-approach-to-ai-regulation-amended-governement-response-web-ready.pdf

Drexl, J., Hilty, R., Desaunettes-Barbero, L., Globocnik, Gonzalez Otero, B., Hoffmann, J., . . . Wiedemann, K. (2021). Artificial Intelligence and Intellectual Property Law: Position Statement of the Max Planck. *Max Planck Institute for Innovation & Competition Research Paper No. 21-10*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3822924

EDPB. (2024, May 23). *The European Data Protection Board has established a ChatGPT task force and published its first report on the group's activities.* Retrieved from https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf

EleutherAI. (2024, May 27). *EleutherAI*. Retrieved May 27, 2024, from https://www.eleuther.ai/

Engineering at Meta. (2021, June 21). *Consolidating Facebook storage infrastructure with Tectonic file system*. Retrieved May 27, 2024, from https://engineering.fb.com/2021/06/21/data-infrastructure/tectonic-file-system/

EUR-Lex. (2019). *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.).* Retrieved from http://data.europa.eu/eli/dir/2019/790/oj

European Commission. (2024). *AI Act: Participate in the drawing-up of the first General-Purpose AI Code of Practice*. Retrieved from https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice

European Commission. (2024, March). *Improving access to and reuse of research results, publications and data.* Retrieved from https://media.licdn.com/dms/document/media/D4D1FAQGsRjNdA1eIBw/feedshare-document-pdf-analyzed/0/1715868941529?e=1717027200&v=beta&t=57ljtga1Go5Hbl0CEZKbs0FzUZbr_it36tP322mtt_o

European Parliament. (2018). *The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market – Technical aspects.* Retrieved from https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/604942/IPOL_BRI(2018)604942_EN.pdf

European Union. (1996). *Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 an.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

Federal Trade Commission. (2024, February 13). *AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive*. Retrieved May 19, 2024, from https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive

Fei, L. (2024). A comparative study on public interest considerations in data scraping dispute. *International Journal of Law in Context*, 1-13. doi:10.1017/s1744552324000156

G7. (2023). Hiroshima Process International Code of Conduct for Advanced AI Systems. Retrieved from https://www.mofa.go.jp/ecm/ec/page5e_000076.html

Government of Canada. (2023). Consultation paper: Consultation on Copyright in the Age of Generative Artificial Intelligence. Retrieved from https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-paper-consultation-copyright-age-generative-artificial-intelligence

Government of Canada. (2023). *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems.* Retrieved from https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document

Government of Japan. (2024). *The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI*. Retrieved from https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html#:~:text=Amid%20the%20growing%20global%20debate%20over%20advanced%20artificial,aim%20of%20promoting%20safe%2C%20secure%2C%20and%20trustworthy%20AI.

Guadamuz, A. (2023, May 5). *Photographer sues LAION for copyright infringement*. Retrieved 2024, from https://www.technollama.co.uk/photographer-sues-laion-for-copyright-infringement

Guttridge-Hewitt, M. (2023). *How web scraping helps collect sustainability information*. Retrieved from https://environmentjournal.online/features-opinion/how-can-web-scraping-help-collect-sustainability-information/

Hall, R. S., Vassilev, A., Greene, K., Perine, L., & Patrick, A. B. (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* NIST. Retrieved from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

Heath, R., & Fried, I. (2023, August 18). *Terms-of-service land grab: Tech firms seek private data to train AI*. (AXIOS, Editor) Retrieved 2024, from https://www.axios.com/2023/08/18/ai-legal-user-data

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M., & Liang, P. (2023). Foundation Models and Fair Use. *Stanford University*. Retrieved from https://arxiv.org/pdf/2303.15715

Hugging Face. (2024). *Dataset Cards*.

Hugging Face. (2024). *Datasets: imagenet-1k*. Retrieved May 2024

Hugging Face. (2024). *Gated datasets.* Retrieved from Dataset Cards

Hutukka, P. (2023). Copyright Law in the European Union, the United States and China. *IIC - International Review of Intellectual Property and Competition Law, 54*(7), 1044-1080. doi:10.1007/s40319-023-01357-0

ICO. (2023). *Joint statement on data scraping and data protection.* Retrieved May 17, 2024, from https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf

IFFRO. (2024, May). *Code of Conduct*. Retrieved 2024, from https://ifrro.org/page/code-of-conduct/

IFFRO. (2024). *Relationship between Reproduction Rights Organisations*. Retrieved May 2024, from https://ifrro.org/page/rro-relationship/

Intellectual Property Office of Singapore. (2022). *Copyright: Factsheet on Copyright Act 2021.* Retrieved from https://www.ipos.gov.sg/docs/default-source/resources-library/copyright/copyright-act-factsheet.pdf

Intellectual Property Office of Singapore. (2024). *Class Licensing Scheme for Collective Management Organisations (CMOs)*. Retrieved May 19, 2024, from https://www.ipos.gov.sg/about-ip/copyright/copyright-owners/collective-management-organisations

ISO. (2022). *ISO 22989.* Retrieved from https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en

Israeli Ministry of Justice. (2022). *Opinion: Uses of copyrighted materials for machine learning.*

J. Gervais, D. (2022). AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Production of AI Machines. *Vanderbilt Law School Publications*. Retrieved from https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=2276&context=faculty-publications

Japan Copyright Office, and Secretariat of Intellectual Property. (2024). *General Understanding on AI and Copyright in Japan.* Retrieved from https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf

Keller, P. (2024). *Considerations for opt-out compliance policies by AI model developer.* Retrieved May 2024, from https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf

LAION. (2024, May 27). *LAION*. Retrieved May 27, 2024, from https://laion.ai/

Lee, N. T., & Lai, S. (2022). *The U.S. can improve its AI governance strategy by addressing online biases*. Retrieved from Brookings: https://www.brookings.edu/articles/the-u-s-can-improve-its-ai-governance-strategy-by-addressing-online-biases/

Leffer, L. (2024). *Artists Are Slipping Anti-AI ‘Poison’ into Their Art. Here’s How It Works*. Retrieved from https://www.scientificamerican.com/article/art-anti-ai-poison-heres-how-it-works/

Levi, S., Epstein, M., & Feirman, J. (2024). *District Court Adopts Broad View of Copyright Preemption in Data Scraping Case.* Retrieved from https://www.skadden.com/insights/publications/2024/05/district-court-adopts-broad-view?sid=4b85270f-c143-44b5-807f-34e19c66e69d

Macgence. (2024, March 5). *A Comprehensive Guide to AI Training Data Collection*. Retrieved May 27, 2024, from Medium: https://medium.com/@macgenceai/a-comprehensive-guide-to-ai-training-data-collection-01d9a329d96c#:~:text=In%20AI%20training%2C%20well%2Dknown,computer%20vision%2C%20and%20speech%20recognition

Mammen, C., Collyer, M., Dolin, R., Gangjee, D., Melham, T., Mustaklem, M., . . . Wang, V. (2024). Creativity, Artificial Intelligence, and the Requirement of Human Authors and Inventors in Copyright and Patent Law. *SSRN Electronic Journal*. doi:10.2139/ssrn.4892973

Mehrotra, D., & Courts, A. (2024). *Amazon Is Investigating Perplexity Over Claims of Scraping Abuse*. Retrieved from Wired: https://www.wired.com/story/aws-perplexity-bot-scraping-investigation/

Mehta, I. (2023). *X updates its terms to ban crawling and scraping*. (TechCrunch, Editor)

Metz, C., Kang, C., Frenkel, S., Thompson, S. A., & Grant, N. (2024). *How Tech Giants Cut Corners to Harvest Data for A.I.* Retrieved from The New York Times: https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html

Mickle, T. (2024). *Scarlett Johansson Said No, but OpenAI’s Virtual Assistant Sounds Just Like Her*. Retrieved from https://www.nytimes.com/2024/05/20/technology/scarlett-johannson-openai-voice.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Retrieved from https://arxiv.org/abs/1301.3781

Ministry of Justice. (2022). *Opinion: Uses of copyrighted materials for machine learning.* Retrieved May 19, 2024, from https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf

Ministry of Justice, Japan. (2024). *Japanese Law Translation Database System, "著作権法（一部未施行）Copyright Act (Partially unenforced)".* Retrieved from https://www.japaneselawtranslation.go.jp/en/laws/view/4207

Ministry of Law and the Intellectual Property Office of Singapore . (2024). *2024 PUBLIC CONSULTATION ON PRESCRIBED EXCEPTIONS IN PART 6, DIVISION 1 .* Retrieved from https://www.mlaw.gov.sg/files/2024_Public_Consultation_on_Prescribed_Exceptions_in_Part_6__Division_1_of_the_Copyright_Regulations_2021.pdf

National Library of Medicine. (2024). Web Scraping Definition. Retrieved from https://www.nnlm.gov/guides/data-glossary/web-scraping

Neuburger , J. (2022, December 8). *hiQ and LinkedIn Reach Proposed Settlement in Landmark Scraping Case*. Retrieved from https://natlawreview.com/article/hiq-and-linkedin-reach-proposed-settlement-landmark-scraping-case

Nobre, T. (2024). *The Post-DSM Copyright Report: research rights .* Retrieved from https://communia-association.org/2024/02/05/the-post-dsm-copyright-report-research-rights/

O'Brien, M. (2023, July 19). *ChatGPT-maker OpenAI signs deal with AP to license news stories*. (A. news, Editor) Retrieved May 2024

OECD. (2018). OECD Guidelines for Multinational Enterprises on Responsible Business Conduct. Retrieved from https://www.oecd.org/publications/oecd-guidelines-for-multinational-enterprises-on-responsible-business-conduct-81f92357-en.htm

OECD. (2019). *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies.* OECD Publishing, Paris. doi:10.1787/276aaca8-en

OECD. (2021). *Recommendation of the Council on Enhancing Access to and Sharing of Data.* Retrieved 03 06, 2023, from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463

OECD. (2022). OECD Framework for the Classification of AI systems. In *OECD Digital Economy Papers.* OECD Publishing, Paris. doi:10.1787/cb6d9eca-en

OECD. (2023). Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI. In *OECD Digital Economy Papers.* OECD Publishing, Paris. doi:10.1787/2448f04b-en

OECD. (2023). *G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI.* OECD Publishing, Paris. doi:10.1787/bf3c0c60-en

OECD. (2024). *"AI, data governance and privacy: Synergies and areas of international co-operation".* Paris: OECD Artificial Intelligence Papers, No. 22, OECD Publishing, https://doi.org/10.1787/2476b1a4-en.

OECD. (2024). *Recommendation of the Council on Artificial Intelligence.* Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD. (2024). *Report on the implementation of the OECD Recommendation on Artificial Intelligence.* Retrieved from https://one.oecd.org/document/C/MIN(2024)17/en/pdf

OECD.AI. (2024). *OECD AI Incidents Monitor*. Retrieved from https://oecd.ai/en/incidents-methodology

Ostwal, T. (2023, August 10). *The New York Times Updates Terms of Service to Prevent AI Scraping Its Content*. (Adweek, Editor) Retrieved May 19, 2024, from https://www.adweek.com/media/the-new-york-times-updates-terms-of-service-to-prevent-ai-scraping-its-content/

Project Gutenberg. (2024). Retrieved from https://www.gutenberg.org/

Ray, P., & Clark, J. (2024). *Artificial Intelligence Index Report 2024.* Retrieved from https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf

Reisner, A. (2023, May 19). *Revealed: The authors whose pirated books are powering generative AI*. (T. Atlantic, Editor) Retrieved 2024, from The Atlantic.

Republic of Korea. (2017). Copyright Act. Retrieved from https://elaw.klri.re.kr/eng_service/lawView.do?hseq=42726&lang=ENG

Reuters. (2024, April 29). *OpenAI to Use FT Content for Training AI Models in Latest Media Tie-Up*.

Retrieved May 19, 2024, from https://www.reuters.com/technology/financial-times-openai-sign-content-licensing-partnership-2024-04-29/#:~:text=The%20Financial%20Times%20has%20signed%20a%20deal%20with,latest%20media%20tie-up%20for%20the%20Microsoft-backed%20%28MSFT.O%29%20startup.

S. Gillis, A. (2023). *Screen Scraping*. Retrieved from https://www.techtarget.com/searchdatacenter/definition/screen-scraping

S. Gillis, A. (2024). *Definition: web crawler*. Retrieved from TechTarget: https://www.techtarget.com/whatis/definition/crawler

Shayne Longpre, R. M.-M. (2024). Consent in Crisis: The Rapid Decline of the AI Data Commons. Retrieved from https://arxiv.org/abs/2407.14933

Shutterstock. (2023, July 11). *Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data*. Retrieved May 19, 2024, from https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., & Russell Authur, B. B. (2024). *Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research.* Retrieved from https://arxiv.org/abs/2402.00159

State of Israel, Ministry of Justice. (2022). *Opinion: Uses of copyrighted materials for machine learning.*

Statement on AI training. (2024). *Statement on AI training*. Retrieved from https://www.aitrainingstatement.org/

Study Group on Intellectual Property Rights in the Era of AI. (2024). *Interim Report of the Study Group on Intellectual Property Rights in the AI Era.*

Teven, L. S., Angela, F., Christopher, A., Ellie, P., Suzana, I., Daniel, H., . . . Sasank, P. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual. Retrieved from https://arxiv.org/pdf/2211.05100

The Australian Government the Treasury. (2023). *Screen scraping – policy and .* Retrieved from https://treasury.gov.au/sites/default/files/2023-08/c2023-436961-dp.pdf.

The Australian Government the Treasury. (2023). *Screen scraping – policy and regulatory implications.*

The Authors Guild. (2023). *The Authors Guild, John Grisham, Jodi Picoult, David Baldacci, George R.R. Martin, and 13 Other Authors File Class-Action Suit Against OpenAI*. Retrieved from https://authorsguild.org/news/ag-and-authors-file-class-action-suit-against-openai/

The Intellectual Property Office. (2024). *Copyright and AI: Consultation.* Retrieved from https://assets.publishing.service.gov.uk/media/6762c95e3229e84d9bbde7a3/241212_AI_and_Copyright_Consultation_print.pdf#:~:text=This%20consultation%20sets%20out%20our%20plan%20to%20deliver,to%20the%20training%20of%20AI%20models%20is%20disputed.

Tiedrich, Lee; Avdulla, Alban;. (2023). *How Can Standard Contract Terms Advance Responsible AI Data and.* Retrieved May 19, 2024, from https://gpai.ai/projects/blogs/Committee%E2%80%99s%20AI%20data%20and%20model%20sharing%20project-(Lee%20Tiedrich%20and%20Alban%20Avdulla)-(03-12-2023).pdf

Tiku, N. (2023, October 25). *AI researchers uncover ethical, legal risks to using popular data sets*. Retrieved May 27, 2024, from The Washington Post: https://www.washingtonpost.com/technology/2023/10/25/data-provenance/

Tong, A., Wang, E., & Coulter, M. (2024, February 22). *Exclusive: Reddit in AI content licensing deal with Google*. (Reuters, Editor) Retrieved May 19, 2024

U.S. Copyright Office. (2024). *Copyright and Artificial Intelligence, Part 1: Digital Replica.* Retrieved from https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf

U.S. Presidential Executive Order. (2025). *Fact Sheet: President Donald J. Trump Takes Action to Enhance America's AI Leadership*. Retrieved from https://www.whitehouse.gov/fact-sheets/2025/01/fact-sheet-president-donald-j-trump-takes-action-to-enhance-americas-ai-leadership/

Ueno, T. (2021). The Flexible Copyright Exception for 'Non-Enjoyment' Purposes – Recent Amendment in Japan and Its Implication. *GRUR International, 70*(2), 145-152. doi:10.1093/grurint/ikaa184

UGC Principles. (2007). Principles for User Generated Content Services. Retrieved from https://ugcprinciples.com/

US Executive Office of Science and Technology Policy. (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* Retrieved from https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

US Government Publishing Office. (2010). *Limitations on exclusive rights: Fair use. Sec. 107 in United States Code, 2006 Edition, Supplement 4, Title 17 - Copyrights .* The Office of the Law Revision Counsel of the United States House of Representatives,.

Villalobos, P., Sevilla , J., Heim, L., Besiroglu, T., Hobbhahn, M., & Ho, A. (2022). Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. Retrieved from https://arxiv.org/abs/2211.04325

Vincent, J. (2023). Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content. Retrieved from https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit

Wakelee-Lynch, J. (2023). *AI's Impact on Artists*. Retrieved from https://magazine.lmu.edu/articles/mimic-master/

Wiggers, K. (2024). LinkedIn scraped user data for training before updating its terms of service. Retrieved from https://techcrunch.com/2024/09/18/linkedin-scraped-user-data-for-training-before-updating-its-terms-of-service/

Willman, C. (2023). *AI-Generated Fake 'Drake'/'Weeknd' Collaboration, 'Heart on My Sleeve,' Delights Fans and Sets Off Industry Alarm Bells*.

WIPO. (2020). WIPO Conversation on Intellectual Proterty and Artificial intelligence - Second session. Retrieved from https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf

WIPO. (2021). *WIPO Good Practice Toolkit for Collective Management Organizations (The Toolkit) : A Bridge between Rightholders and Users.* (W. I. Organization, Ed.) World Intellectual Property Organization. doi:10.34667/tind.44374

WIPO. (2024). *Hamburg Regional Court, Germany [2024]: Robert Kneschke v. LAION e.V., Case No. 310 O 227/23, Germany.* Retrieved from https://www.wipo.int/wipolex/en/judgments/details/2381

WIPO. (2024). *https://www.wipo.int/web/trade-secrets#:~:text=In%20general%2C%20to%20qualify%20as%20a%20trade%20secret%2C,of%20confidentiality%20agreements%20for%20business%20partners%20and%20employees.* Retrieved from Trade Secrets: https://www.wipo.int/web/trade-secrets#:~:text=In%20general%2C%20to%20qualify%20as%20a%20trade%20secret%2C,of%20confidentiality%20agreements%20for%20business%20partners%20and%20employees.

WIPO. (2024). *Summary of the Berne Convention for the Protection of Literary and Artistic Works (1886)*. Retrieved from https://www.wipo.int/treaties/en/ip/berne/summary_berne.html

WIPO. (n.d.). Is Artificial Intelligence Collides With The Trademark Law?´. Retrieved from https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/ind_revella.pdf

WIPO. (n.d.). *WIPO-Administered Treaties*. Retrieved from https://www.wipo.int/treaties/en/

Notes

[1] This paper does not include patents or inventions within its scope. This paper should not be construed as authorising or suggesting policies or practices inconsistent with international obligations concerning intellectual property rights.

[2] The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), established under the WTO, sets out minimum standards for IP protection that member countries must follow. It covers areas such as copyright, patents, and trademarks, creating a harmonised framework for international trade in IP. WIPO-administered treaties, including the Berne Convention for the Protection of Literary and Artistic Works (copyright) and the Paris Convention for the Protection of Industrial Property (patents and trademarks), further contribute to the global harmonisation of IP laws. These international agreements facilitate the protection of IP rights across jurisdictions while allowing for some flexibility in national laws. The WIPO Copyright Treaty and WIPO Performances and Phonograms Treaty (collectively known as the WIPO Internet Treaties) were designed to update and supplement existing international treaties on copyright and related rights (the Berne and Rome Conventions) and contain provisions relevant to data scraping that include: (1) the prohibition against formalities, which some stakeholders raise in the context of “opt outs”; (2) the 3-step-test, which sets forth boundaries for exceptions and limitations; (3) protections for rights management information; (4) technological protection measures; (5) and separate protections for compilations of data or other material

[3] According to OECD AI Principle 1.5, AI actors must be accountable for the proper functioning of AI systems. This includes ensuring traceability of datasets and decisions throughout the AI system lifecycle and adopting a systematic risk management approach to address potential risks. Such risks encompass issues related to intellectual property rights, privacy and harmful bias.

[4] This report does not include within its scope patents or inventions. This paper should not be construed as authorising or suggesting policies or practices inconsistent with international obligations concerning intellectual property rights.