GPAI
ASSOCIATED
PROJECT

From scraping to ethical sharing: Initial considerations for Virtuous Innovative Approaches and Data Use Collaboration in AI Training (VIADUCT)

December 2025

Inria
CENTRE OF THE GPAI
EXPERT COMMUNITY

This report was produced by the Inria Centre of Expertise for International Cooperation on AI in the context of the VIADUCT initiative, under the steering of an advisory group composed of volunteer members selected for their recognized expertise in ethical data sharing and collaboration practices. It was not subject to approval by GPAI and OECD members and should not be considered to reflect their position.

A core part of the OECD/GPAI expert community, the Expert Support Centres are nationally funded AI-focused entities in three countries: Canada (CEIMIA), France (Inria) and Japan (NICT).

Acknowledgements

This report was written by Yann Dietrich, Marie Langé and Bertrand Monthubert, and coordinated by Jean Constantin, Viaduct initiative lead at Inria, as part of a GPAI-associated project.

We are grateful to the OECD's AI & Emerging Technologies Division and to the Data Flows, Governance & Privacy Division, as well as Inria's Centre of Expertise for International Cooperation on AI for their thorough review and constructive comments.

Please cite this publication as:
Constantin, J., Dietrich, Y., Langé, M., Monthubert, B. (2025).
From Scraping to Ethical Sharing:
Initial Considerations for Virtuous Innovative Approaches and Data Use Collaboration in AI Training. Inria.

Table of Contents

Executive Summary

The rapid development of artificial intelligence (AI) relies on access to vast volumes of data throughout its lifecycle. The sourcing of this data has relied on legally and ethically contentious practices, particularly the indiscriminate scraping of publicly available and often copyrighted content. Popular datasets like CommonCrawl and LAION 5B contain copyrighted works and personal data used without explicit permission or compensation for data holders. This approach has triggered a global backlash, with over 50 lawsuits filed against AI developers and increasing technical barriers against scraper robots. Leaders in the AI industry now warn of *"peak data"*, as public human-generated content will soon be exhausted. This scarcity conflicts with AI's ever-growing appetite for high quality expert data to support increasingly advanced applications.

Data for AI is not uniform but spans multiple domains and governance regimes which can evolve or overlap depending on contexts and jurisdictions[1]. Each of these regimes: copyrighted content, personal data, trade secrets, government data, and open data, is constrained by distinct legal and technical restrictions. Copyrighted materials require permission from holders, yet enforcement of opt-out decisions remains inconsistent. Personal data is protected under GDPR, demanding anonymisation and clear legal grounds for processing, while trade secret datasets are shielded by confidentiality agreements. Government data, though mandated to be open, often remains inaccessible due to sensitivity or infrastructure limitations. Open data, while legally permissive, suffers from fragmentation and underinvestment. These disparities create a fragmented landscape where data sharing is hindered by transaction costs, confidentiality requirements, and misaligned incentives.

Efforts to address these challenges have produced partial solutions. Opt-out mechanisms like ai.txt and TDMRep allow data holders to declare preferences but lack standardisation. Privacy preserving techniques enable secure data processing but at high computational cost. Licensing agreements can bring legal clarifications but are hindered by contractual complexity. Data attribution models, designed to compensate data holders, remain impractical at scale. No single solution suffices, highlighting the need for context specific approaches that balance innovation with data holders' interests.

Fostering ethical data sharing is not trivial and requires addressing multiple technical, economic and legal obstacles. The VIADUCT initiative proposes an experimental approach, engaging with data holders and AI developers to characterize constraints and explore innovative data sharing approaches.

[1] The scope of this report is limited to the European Union (EU) legal framework.

Introduction

As artificial intelligence (AI) systems continue to advance at remarkable speed, questions surrounding the provenance and governance of the data used for their development have taken on renewed importance. Recent analyses indicate that a substantial share of training datasets relies on large-scale web-scraping, a practice that has shaped the development of contemporary AI models. The OECD's 2025 policy paper *Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data* highlights that widely used corpora (such as the CommonCrawl dataset) were assembled from online content, including material protected by intellectual property and personal data rights. At the same time, most major AI developers have been opaque on their training datasets, with transparency declining steadily over the past years (Wan et al., 2025). The combination of rapid technological progress, rising data needs, and evolving legal and societal expectations have led to a turning point for the AI ecosystem; one that invites reflection on new models of **data governance**[2] within and in between data and AI stakeholders.

Growing pressures on traditional data sources for AI have brought forth the need to rethink how data is accessed and shared. Data sharing, "*the process of making an organization's data resources available to multiple applications, users and other organizations*" (Mucci, 2025), has been traditionally tackled as a technical problem with the development of communication technologies from early computers to modern platforms. Yet, the legal, economic and social dimensions of data sharing must also be addressed to foster sustainable and ethical data sharing mechanisms for AI development. In this perspective, VIADUCT[3], a GPAI-associated project, proposes innovative approaches to support the emergence of ethical data sharing models for AI at the intersection of law, economics, and technology.

During 2025, 25 interviews[4], as well as two workshops were conducted with **data holders,** AI developers (**data users**), **data intermediaries** and experts to qualify current data sourcing practices, and explore ethical alternatives. This report is based on the issues and solutions raised by the interviewed stakeholders, and is completed by an in-depth literature review.

Section 1 examines why current AI data sourcing practices have become increasingly unsustainable and argues that more collaborative approaches between data holders and AI developers are necessary to enable long-term AI innovation. Section 2 shows that "*data for AI*" is not monolithic but a mosaïc of objects, legal regimes and constraints that shape data sharing modalities and governance. Finally, section 3 offers a detailed review of existing solutions to foster and support ethical, transparent and secure data sharing for AI. The analyses presented in this report are primarily based on the legal framework of the European Union (EU), and may not apply uniformly to other jurisdictions.

[2] *See Appendix 3 for definitions of all core terms, which appear in bold throughout the report.*

[3] VIADUCT: *Virtuous Innovative Approaches and Data Use Collaboration for AI Training*

[4] See the detailed list of stakeholders consulted in appendix 1

Section 1. The AI data dilemma: data scraping and scarcity

Artificial Intelligence and Machine Learning technologies have accelerated dramatically since the early 2010s, driven by advances in computing power and the explosion of accessible training data. A turning point came with the release of OpenAI's ChatGPT in November 2022, which sparked a global race to develop and deploy increasingly powerful generative AI models. By 2024, private investments in AI had surpassed 252 billion USD worldwide (Kariuki, 2025) as major technology platforms, startups and researchers started competing with their own algorithms vying for talents, computing resources and data.

Indeed, data has become a strategic asset, essential to every stage of the development and deployment of AI models. From pre-training to fine-tuning, from real-time grounding to ongoing improvement, data fuels the entire AI value chain. This growing dependence on data raises pressing questions about how AI models are fed, what types of data are used, under what conditions, and with what societal consequences.

1.1. Data at every step of the AI development process

Data, often described as the "fuel for AI" (Ilya Sutskever, 2024) has been the object of ever-growing appetite from AI model developers. It is used at every step the development and usage of AI systems as described in figure 1.

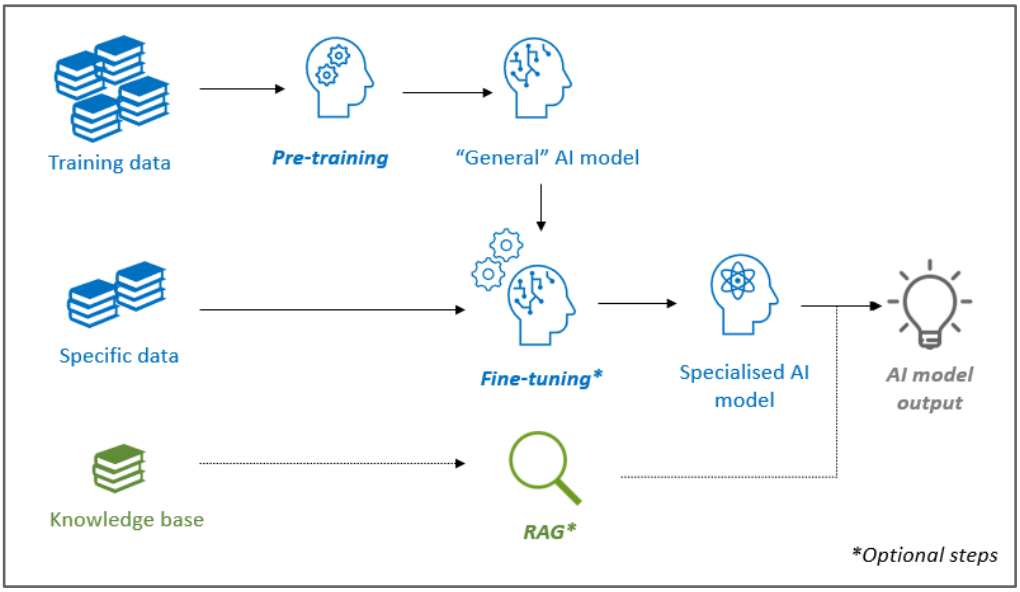

Figure 1. AI model development process

AI models usually undergo a first pre-training phase where they are fed with large volumes of generalist data to master basic abilities such as syntax and vocabulary for Large Language Models (LLMs) or object recognition for image models. Pre-training has relied on general purpose corpora of texts, images, videos, sounds or software code, often available online on dedicated platforms such as HuggingFace or Kaggle. Some popular training datasets include

CommonCrawl, Google's C4, Wikipedia (Liu et al., 2024) for text or LAION-5B for images. Pre-training datasets have followed AI models' rising complexity, with their size doubling every six months since 2010 (Rahman, 2024).

Pre-training is often followed by a second phase called fine-tuning, which aims at specializing a model on a certain task, topic or sector. This phase can be performed by the pre-trainer or by another organisation. For instance, OpenAI's GPT models were fine-tuned on text pairs (instructions and answers) to master fluid conversation with users. Such datasets may be collected from existing sources or purpose-built. Common general instruction datasets included Stanford's Alpaca_data, Tsinghua's OpenChat, while other datasets such as MedDialog specialize on medicine, DISC-Law-SFT on legal reasoning or OpenMathInstruct-1 on mathematical problem reasoning.

As shown in figure 1, data can also be used to generate content without being incorporated into the model itself. This occurs with techniques such as Retrieval-Augmented Generation (RAG). RAG architectures ground a model's outputs by enriching its prompts with relevant information and documents. This enables the model to provide more accurate, contextual and fresh answers while bypassing costly retraining.

1.2. Untrustworthy web-scraping practices as the foundation of modern AI models

In its 2025 report, *Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data,* the OECD highlights deep issues regarding the composition and sourcing of popular training datasets. Indeed, many of the datasets commonly used for AI training were scraped from publicly accessible websites, processed and aggregated by initiatives such as Common Crawl and LAION. Web scraping refers to "*the automated extraction of data from the web, online databases and from other sources using automated software tools or scripts*" (OECD, 2025, p.16). This practice is usually systematic, indiscriminate and unilateral, escaping any form of governance. Hence, the harvest of online content by scraper bots has often encompassed public domain content, copyrighted works and GDPR-protected personal data alike (Figure 2).

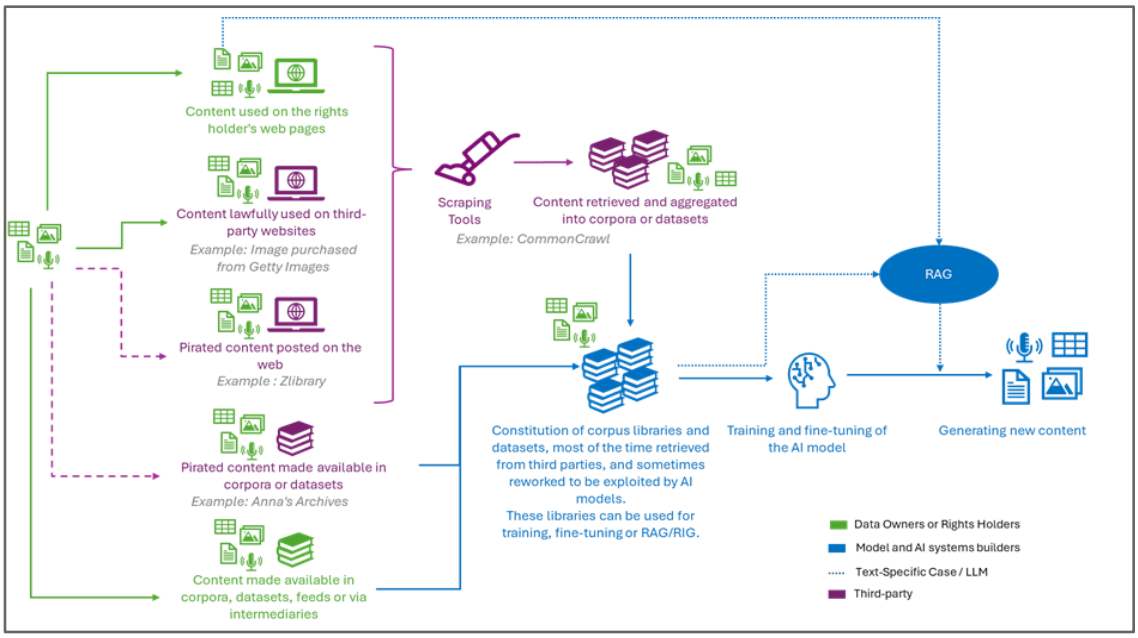

Figure 2. The data value chain in AI systems involves multiple stakeholders

CommonCrawl remains one of the most extensive scraped datasets containing billions of webpages. Many other popular datasets, like Google's C4, or Inria's OSCAR corpus are directly derived from CommonCrawl, with additional processing and filtering for undesirable content such as hate speech, abusive language, duplicates or error messages (Baack, 2024). In addition to scraped datasets, companies may also aggregate proprietary data to train their own models or to license to AI developers. This is particularly the case for large platforms and social media which can easily leverage user posts to train their AI models (e.g. xAI for Grok, LinkedIn) or sell them to a third-party (e.g. Reddit to Google). Due to the volume and complexity of these datasets, and despite heavy reprocessing and curation efforts, undesirable, protected content is likely to remain.

Indeed, scraped training datasets like Books3, Google's C4, or LAION-5B contain large amounts of copyrighted materials such as books, press articles, photographs, movies or software code (Reisner, 2023; Schaul et al., s. d.). In a step further, while banned in several European countries, Meta accessed and downloaded protected books from the "shadow" Library Genesis (Richard Kadrey v.. Meta Platforms Inc., 2025). Similarly, the image dataset CommonPool was found to contain multiple instances of personal and sensitive information, including on children, such as individuals' names, religions, sexual orientation, portraits, identity documents or resumes, which had been scraped from public online sources like social media or news articles (Hong et al., 2025).

Identifying the use of protected content has been challenging for data holders and governments as AI developers rarely disclose their training datasets. The complexity of the scraped data value chain, coupled with the "black box" nature of AI models reinforce this opacity and unclear legal status. Going against this trend, some AI initiatives like Ai2's OLMO

or BigScience's BLOOM have sought to build fully open models by releasing weights, code and training data.

1.3. The sustainability of data sourcing practices in the AI industry in question

Despite the massive harvesting of online data, Villalobos et al. (2024) notes that "*by the end of this decade, [...] the current reliance on public human text data for training ML models may become unstainable*". This prediction has been echoed in comments throughout the industry: "*We've achieved peak data, and there will be no more*" (Ilya Sutskever, 2024), "*We've already run out of data*" (Neema Raphael, 2025). This shortage of public training data is accentuated by a growing political and cultural backlash against AI developers, as exemplified by the silent album "Is This What We Want?" released by 1000 British music artists in favor of greater artist protection. In this direction, content and data holders have initiated over fifty lawsuits in the United States (e.g. The Authors Guild v. OpenAI & Microsoft, 2023; Reddit v. Anthropic, 2025), in Canada (e.g. Canadian News Media Companies v. OpenAI), in the United Kingdom (Getty v. Stability AI, 2025), in France (SNE, SGDL and SNAC v. Meta, 2025) or in India (ANI V. OpenAI, 2024). Conversely, websites and social media have started fortifying their online contents by updating their Terms & Conditions to ban scraping and AI training by third-parties (Mehta, 2025) and by blocking scraper bots (Fletcher, 2024). The ban of scraper bots serves a double purpose: protect contents and prevent bots from overwhelming website servers.

The deepening scarcity of new online public training data clashes directly with AI developers' ambitions to train ever-more performing models. Sophisticated models are increasingly integrated into organizations' processes and decision making, most notably through agentic AI. To effectively and safely manage complex tasks, such as financial fraud detection, customer service automation, or pathology diagnosis, AI models must develop specialized industry expertise. As a result, high value use cases will demand increasingly more specialized, expert, and high-quality data for training, development, and deployment. Alternatives such as synthetic data generation are under study but have been shown to collapse model performance (Shumailov et al., 2024) or require seeding with real data (Seddik et al., 2024).

Unlike classical training data, commonly sourced from public web scraping or open datasets, expert data often resides in closed organization databases: corporate processes, financial documents, engineering files, patient X-rays, customer data, support tickets, etc. "Closed data", stored on private cloud or on-premise servers are inaccessible to web crawlers and scrapers, yet they account for 95% of all digital content (Greengard, 2025). These data are diverse in objects, structures and subject to multiple legal regimes, including trade secrets, personal data, intellectual property or government data. Such data offers advantages for AI model training: it encompasses a broader range of formats, reflects industry expertise, and describes real world processes. Finally, organization data often follows more stringent governance standards than online public contents, limiting the risks of inaccuracies, disinformation, low quality contents or model poisoning.

However, the lack of equitable value-sharing mechanisms, ensuring data holders also benefit from AI-generated economic gains, and inadequate confidentiality protections have fostered a climate of distrust. This low trust environment has led private organizations to prioritize data protection, with cybersecurity risks and intellectual property infringement ranking among their top concerns in AI adoption (McKinsey, 2025).

The multiplication of challenges affecting traditional sources of AI data raises concerns about the long-term viability of current AI industry practices. To ensure continued and more ethical AI innovation, alternative data sourcing approaches must be developed. VIADUCT seeks to address this need by engaging with both data holders and AI developers to characterize both parties' constraints, expectations and terms, to then define and experiment new, mutually beneficial mechanisms for data sharing.

Section 2. Data for AI: A mosaïc of objects, legal regimes and constraints

Data in the context of AI is far from monolithic: it encompasses a wide variety of domains, from copyrighted texts to personal information, trade secrets, government documents, and open datasets — each governed by distinct legal regimes and shaped by specific technical, economic, and ethical constraints. Understanding this heterogeneity is essential to designing fair, secure, and efficient data sharing mechanisms for AI development. A one-size-fits-all solution does not appear feasible nor desirable. Section 2 introduces a typology of data governance regimes based on European laws: copyrighted, personal, proprietary, public sector and open data. Each regime establishes specific legal, economic, and technical constraints on data sharing, thereby determining their degree of openness, ranging from fully closed to fully open. These regimes are neither mutually exclusive nor static: a single dataset may simultaneously fall under multiple regimes (e.g. a copyrighted article containing personal information), which may evolve depending on context. This typology draws on the European Data Space Support Centre's *Blueprint* (Data Spaces Support Centre, 2025) and the OECD's report on *"Enhancing Access to and Sharing of Data"* (OECD, 2019). It will serve as a foundational analytical tool to guide VIADUCT's data sharing projects.

2.1. Data typologies in the EU legal context

2.1.a. Copyrighted contents and works

Copyrighted content represents a major category of data used in AI training and development. EU copyright law encompasses all creative works including texts, images, videos, sounds (InfoSoc Directive 2001/29/EC, 2001), as well as software code (Software Directive 2009/24/EC, 2009). Such content can be found on public websites or in private databases, in both cases copyright protections apply uniformly. Copyright protection grants authors ("data holders") exclusive rights to reproduce, communicate and distribute their works. It also protects against technological circumvention and remains in force for seventy years after the author's death. Copyright protection is wide and protects extensive arrays of works including social media posts whose rights are owned by users but automatically licensed to platforms under their terms and conditions.

Protection may extend to databases subject that the selection and the arrangement of the database meet the originality requirement. A “sui generis” database protection can also apply to protect substantial investments in obtaining, verifying, or presenting contents (Database Directive 96/9/EC, 1996). Sui generis database protection expires fifteen years after completion.

For AI, special copyright exception regimes were defined to authorize “text and data mining” (TDM) processing under specific circumstances (DSM Directive (EU) 2019/790, 2019). TDM processing on copyrighted content is always permitted for scientific research, and allowed for non-research purposes unless data holders explicitly opt-out via appropriate means, such as machine-readable metadata. However, enforcing data holders’ opt-out decisions can be particularly difficult due to the lack of industry standards and opaque practices by AI developers. Moreover, the current opt-out system does not support more complex scenarios such as opt-in with financial compensation conditions. No direction out of european legislations provides guidance on how to define compensation in the case of AI training and technical solutions efficiently supporting such mechanisms have yet to emerge.

2.1.b. Personal data

Personal data refers to information “relating to an identified or identifiable individual” for which data subjects (“data holders”) have privacy interests (OECD, 2025b). EU’s GDPR sets an obligation for all personal data processing (e.g. AI training, inference), taking place in the EU or performed on EU legal subjects, to be lawful, fair, transparent, confidential and minimal. The law defines clear grounds for lawful processing including data subject consent, contract performance, vital and public interest as well as controllers’ legitimate interests. To be lawful, a processing must remain within the scope of specified, explicit and legitimate purposes. Any further processing would require new legal grounds, including renewed data subject content.

Hence, data subjects retain control over their data even after transferring it to a controller organization (“data user”) or making it public. They may need to consent to further processing if it is incompatible with the original purpose, and in certain cases, they can request that the controller organization erase or transfer their data. Stricter restrictions apply for sensitive data including a person’s racial or ethnic origins, political opinions, religious beliefs, health data, etc. Data processing for AI (e.g., LLM/ML training) or RAG is subject to the same GDPR requirements and should be limited to what is necessary and relevant for the lawful purpose, whether based on consent or another legal basis.

The Digital Omnibus, recently proposed by the European Commission, confirms that AI-related data processing can rely on legitimate interest, but must still respect safeguards, including balancing the controller’s interests with the data subject’s rights. This effectively shifts the standard AI processing purpose from user consent (opt-in) to legitimate interest where individuals must actively opt-out. It further lowers regulatory barriers by introducing a "subjective" definition of personal data, treating pseudonymized data as non-personal if the controller organization cannot re-identify it, and permitting the use of sensitive data for specific AI tasks like bias detection.

2.1.c. Trade secrets and proprietary data

A proprietary dataset can be considered a trade secret if it constitutes information which is "*not known or readily accessible*", "*has commercial value because it is secret*", and "*has been subject to reasonable steps [...] to keep it a secret*" (Directive EU/2016/943, 2016). Disclosure and usage is unlawful and can be remedied if trade secrets were accessed without authorization, via dishonest commercial practices or in breach of a confidentiality agreement. On top of this regime, copyright law, the GDPR or contract law may offer additional layers of protection depending on the nature of the data and the contractual framework (e.g. NDAs).

Trade secrets, such as proprietary databases, may be shared to third-party ("data users") by holding organizations ("data holders") with their explicit approval and under strict confidentiality measures like licensing agreements, confidentiality clauses and adequate cybersecurity. The absence of such measures when sharing a trade secret may result in the lifting of legal protections. For AI development, proprietary datasets may be shared with third-parties for collaborative R&D projects, in bilateral commercial transactions, or through specialized data marketplaces (e.g. Snowflake Marketplace) as long as proper measures are taken to maintain secrecy. Any sharing or transaction should be accompanied by a licensing agreement with data users detailing modalities, conditions and authorized usages.

2.1.d. Public sector data

Public sector bodies in the EU, including state, regional, local authorities, are required to make their documents available for commercial and non-commercial reuse (Open Data Directive UE/2019/1024, 2019). Public sector data should be made available by electronic means in "*open, machine-readable, accessible, findable and reusable [formats], with their metadata*". This obligation does not concern copyrighted documents, trade secrets, sensitive information (national security, defence, statistical or commercial confidentiality) and personal data. Access to public sector data cannot be limited via exclusive licensing conditions, nor can it be charged beyond the processing costs incurred by administrations.

Specifically, data produced as part of publicly funded research should be open by default following FAIR[5] principles and made available for both commercial and non-commercial reuse. More restrictive regimes, such as personal data or trade secrets, may overlap with openness requirements. In such cases, technical and governance solutions may be leveraged to mitigate sharing risks: limited access, safe and secure environments, desensitisation, metadata or data stewardship and ownership (OECD Committee for Scientific and Technological Policy, 2021).

The OECD highlights key principles to enhance access and reuse of public sector data including openness, transparent conditions for reuse, complete data catalogues, data quality and integrity, transparent and consistent pricing and non-exclusivity (OECD Digital Policy Committee, 2008). Public sector entities globally have taken significant

[5] *FAIR: Findable, Accessible, Interoperable, Reusable*

steps to open their data for reuse via platforms such as data.europa.eu for the European Union, data.gouv.fr in France, data.overheid.nl in the Netherlands, open.canada.ca in Canada or data.go.jp in Japan among others. Yet, full and effective government data sharing remains hindered by persistent challenges. Many key datasets (e.g. health, taxes, police, defense) describe sensitive information precluding unrestricted dissemination. Additionally, the costs associated with collecting, processing and publishing ever-more complex and voluminous datasets pose a significant barrier for the public sector. Building and maintaining the necessary infrastructure requires resources and expertise that are often lacking in smaller local government bodies. Consequently, government data, even though open de jure, often remains de facto unavailable.

2.1.e. Open data

Open data refers to non-discriminatory data access and sharing arrangements, where data is machine readable and can be accessed and shared, free of charge, and used by anyone for any purpose subject, at most, to requirements that preserve integrity, provenance, attribution, and openness (OECD Committee for Scientific and Technological Policy et al., 2021). In Europe, public domain encompasses all contents with expired copyrights (70 years after author's death), government documents, non-copyrightable ideas and facts, and all contents under open licenses. Open licenses authorise content to be freely used, redistributed, sold and modified, free of charge by any person or organization including private companies. Creative Commons (CC) is a popular open framework which offers standard licenses with varying conditions for reuse (Creative Commons, s. d.). For instance, CC BY allows for full content reuse with mandatory credits to authors, CC BY-SA offers the same conditions with the obligation to share derivative work under the same terms ("share alike"). Other CC licenses are more restrictive and prohibit commercial uses (CC BY-NC) or all derivative works (CC BY-ND). License conditions are triggered for any AI processing when no TDM exception apply and in case of training memorization.

Hence, in most cases open data content can be freely used for pre-training, fine-tuning or RAG, and may also be republished as part of open software development. Multiple initiatives, like Pleias' Common Corpus (Langlais et al., 2025), have sought to aggregate open data in AI-ready datasets as a legal and risk-free alternative. These open datasets, even though much smaller, offer legal certainty and transparency for AI developers.

Legal restrictions on open data are limited, yet high investments are required for digitalisation, aggregation, processing and dissemination. In a classic "*tragedy of commons*", open data governance can be fragmented, leaving gaps for issues such as low quality, low standardization and poor infrastructure. Additionally, many public domain resources remain inaccessible on closed databases, this is especially the case in small local government bodies with limited resources and technology expertise to open their data in a safe and efficient way. In a similar way, many public domain documents and archives have not yet been digitalised and thus remain inaccessible for AI use.

"Data Commons" provides a blueprint for collective stewardship and community-driven governance. This model acknowledges that open data, originally thought as a public good (non-rival, low excludability), is vulnerable to over-use jeopardizing trust and the long term sustainability of data ecosystems. Given that data is generated within complex social systems, the communities responsible for its production are best-positioned to govern it democratically, regulate access and maximize its public impact (Tarkowski & Zygmuntowski, 2022).

2.2. Ethical data sharing for AI: Technical, economic and legal constraints

2.2.a. Defining "*ethical data sharing*"

AI systems are not merely technical artefacts, but rather sociotechnical arrangements combining algorithms, data, knowledge, people, organization and economic interests in specific contexts of usage (OECD, 2022b). Similarly, "data" is not only a benign collection of zeros and ones, but reflects the technical, social, political and economic realities in which it is produced and used (Kitchin and Lauriault, 2014). As shown in the previous section, the governance of a dataset can sit at a crossroad between multiple diverging stakeholder interests and legal regimes. These interplays highlight the need for ethics in how data is governed, shared and reused across organizations, particularly for AI. This report introduces an alternative model of data sharing governance: **ethical data sharing**.

Existing literature does not offer a clear and global definition of this concept. Ethics in sharing data has been primarily discussed in relation to personal data, specifically in the context of scientific research (e.g. medical trials). However, as argued above, non-personal data also embed and impact social and economic dynamics, thus motivating the need for a broader ethical framework of data sharing to cover personal data, copyrighted content, business proprietary data, public data and open data alike.

In order to build a global definition of ethical data sharing, this report blends elements of policy texts and gray literature. The European Strategy for Data (EU Commission, 2020) envisions a single market for data guided by European values and fundamental rights. The strategy identifies conditions to realize this ambition: data access equality, governance, data quality, efficient and secure infrastructures and enforceable individual rights. The Global Partnership for Sustainable Development Data (Barbero & McLaren, 2024) lists policy landscape, trust building, value sharing, dependable infrastructures and empowered users as the main ingredients to foster ethical multiparty data sharing. Based on these inputs, **ethical data sharing** can be generally defined as **a set of technical, legal, economic and institutional arrangements which supports compliant, trustworthy and fair sharing of data, for all involved stakeholders, aimed at both commercial and non-commercial reuse, including AI** (figure 3).

To be ethical, a data sharing process must first be compliant with contractual terms and local regulations. However, ethics in data sharing requires principles beyond simple compliance. Trust is the foundation of any consented exchange or transaction for all involved parties. Data holders must trust data users to handle their data in an appropriate and safe way, while data users must receive the guarantee that the data they acquire is compliant, truthful and of high quality. The European Commission highlighted seven key requirements for trustworthy AI systems, including their data: human agency, technical robustness, governance, transparency, accountability and societal well-being (European Commission, 2019). Ethical data sharing's final component, fairness, stipulates that parties should be treated and compensated equitably. This implies that data holders should be credited and receive a share of the value generated with their data proportional to their contributions.

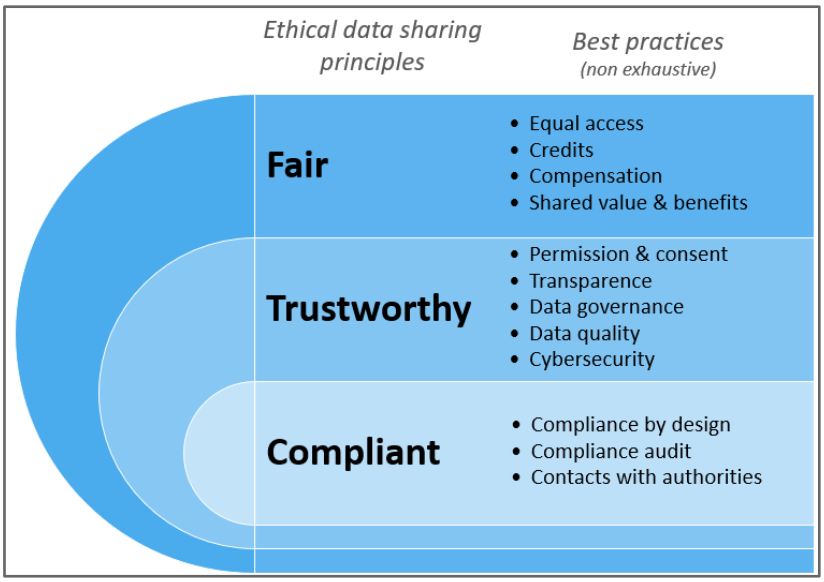

Figure 3. Ethical data sharing's guiding principles

2.2.b. Obstacles to ethical data sharing

Ethical data sharing is constrained by various challenges and restrictions, depending on the nature and type of data at stake (Table 1). While these constraints apply to data sharing for any use case, they are especially pronounced in the context of AI and ML due to the high volumes and complexity of data involved, as well as the "black-box" nature of these algorithms:

- **Transaction frictions**: Sharing large datasets for AI development often entails complex technical, legal and economic logistics. Data transfers may be lengthy and costly processes for organizations requiring expertise in multiple domains: contracts, licenses, cybersecurity, etc. This complexity creates significant barriers to participation in AI data markets, particularly for smaller organizations.
- **Data quality**: Datasets often require heavy reprocessing to be AI-ready: aggregation, formatting, standardization, curation and correction. Datasets may also contain undesirable elements such as disinformation or

discriminatory content. Data quality issues can be complex and expensive to manage, and can introduce significant uncertainty for downstream usages.

- **Digital self-determinatio**n: For personal data, copyrighted content and trade secrets, EU legal texts grant data holders the authority to determine any processing performed on their data including reproduction, modification, commercialization or transfer to another party. Except for certain exceptions, data processing activities require permission from data holders, whether data is stored on private or public databases.
- **Confidentiality**: By nature, personal and trade secret data should remain strictly confidential when processed. Confidentiality requirements may also arise for sensitive government data including personal data or data related to national security. Such data are required to be protected by appropriate technical, organisational and contractual measures when stored, shared across parties and processed for AI.
- **Cost and value compensation:** To foster ethical data sharing with AI developers, fair compensation mechanisms are essential to offset the costs of data extraction and dissemination while aligning incentives with data holders' interests. For copyrighted content and trade secrets, compensation may take the form of monetary remuneration or equity stakes in collaborative AI applications. By contrast, government and open data may only be charged a fee limited to recovering the operational costs associated with data reproduction, anonymization, and dissemination (Open Data Directive UE/2019/1024, 2019). A critical exception exists for personal data, which cannot be treated as a tradeable commodity (European Data Protection Board, 2021), thereby precluding financial incentives.

Data types	EU legal texts	Transaction frictions	Data quality	Digital self-determination	Confidentiality	Cost and value compensation
Copyrighted data	- Database directive (1996) - Infosoc directive (2001) - Software directive (2009) - DSM directive (2019)	X	X	X		X
Personal data	- GDPR (2016)	X	X	X	X	
Trade secrets	- Trade Secret directive (2016) - Data Act (2023)	X	X	X	X	X
Public sector data	- Open data directive (2019) - Data Governance Act (2022)	X	X		X	X
Open data	- Copyright duration directive (2006) - Open data directive (2019) - Data Governance Act (2022) - Open licenses		X			X

Table 1. Data types and their associated data sharing challenges (Source: authors, inspired by Data Spaces Support Centre, 2025)

Section 3. Approaches to facilitate the emergence of ethical and legal data sharing for AI

While the technical and legal challenges of AI data sharing are now well documented, sustainable and ethical solutions remain scarce. The OECD emphasizes, “standardised and widely accessible technical tools” as one of the key levers to overcome data sourcing for AI (OECD, 2025, p.28).

However, an exclusively technical approach cannot be sufficient, and any viable data sharing solution must address three interdependent dimensions: technical feasibility, legal compliance, and, crucially, economic viability. Enabling ethical data sharing, particularly for high-value or sensitive data, requires more than permission and infrastructure. It requires incentives, compensation models, and trust-building mechanisms that can support open data commons, protect data holders, and ensure that the value extracted from data is fairly redistributed. Without economic models that align the interests of data holders and AI developers, even the most sophisticated legal or technical frameworks are likely to fail in practice.

This section explores a range of existing and emerging approaches (see Figure 4) that aim to facilitate responsible data access and reuse across the AI value chain. These include regulatory instruments (e.g., opt-out regimes), technical standards (e.g., metadata-based enforcement), privacy-preserving technologies, licensing mechanisms, smart contracts, and even experimental attribution systems. Each solution is analyzed for its potential to address specific friction points : from consent and confidentiality to transaction costs and economic incentives.

While each obstacle described in section 2 may appear as exclusively technical, legal or economic, any solution must address these aspects interdependently. For example, incentivizing greater data sharing implies the development of new technical frameworks supporting data valorization and equitable value redistribution. Similarly, data quality, though primarily a technical concern, directly influences a dataset’s value and transaction costs. Section 3 explores current and emerging approaches (Table 2) to tackle ethical data sharing obstacles while accounting for the technical, legal and economic dimensions. Each of these solutions must rely on a strong data governance in order to deploy them effectively and maximize their impact.

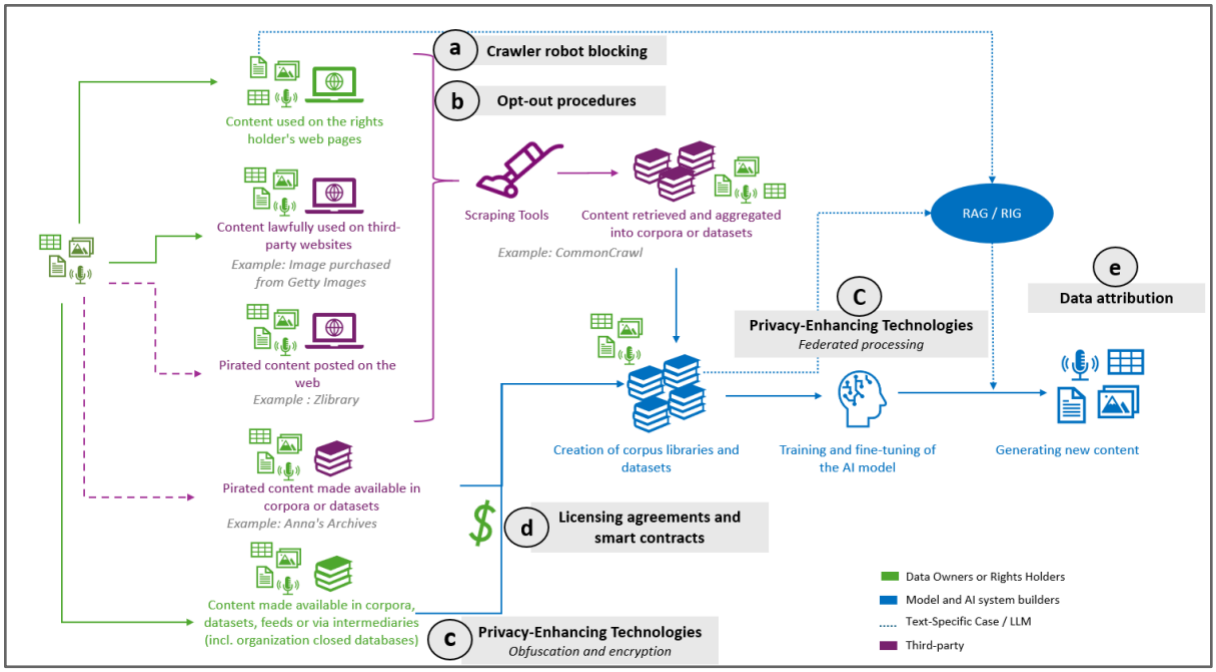

Figure 4. Solutions can support ethical data sharing all along the AI data value chain

3.1. Crawler robots blocking

In an effort to encourage innovation, certain countries and jurisdictions have defined extensive copyright exceptions for AI training and data analysis. Singapore, Israel, Japan or the United States all define broad copyright exceptions for uses falling under computational data analysis (Singapore), acts of non-enjoyment (Japan) or fair use (Israel, United States). These regimes, however, establish certain legal limitations like the lawful access to content used for AI. In 2025, a US court recognized that Anthropic's use of copyrighted books for AI training was fair use, but condemned the AI company for pirating copies (Bartz v. Anthropic, 2025).

In this context, websites have increasingly relied on technical methods such as .htaccess or tools such as Kudurru to block crawler robots and prevent their content from being scraped (Appendix 2). In 2025, robots represented over 30% of global web traffic with 18% growth from 2024 to 2025 (Cloudfare Blog, 2025). Data scraping robots for AI have rapidly grown with OpenAI's GPTBot representing 30% of all AI-dedicated crawlers in 2025 (Cloudfare, 2025). Blocking crawlers can thus protect websites against two issues: unauthorized content scraping and infrastructure overload.

However, crawler robots encompass multiple uses beyond data scraping, including search indexing (GoogleBot), API clients, performance checks and traffic monitoring. Hence, blocking all robot crawlers comes with risks for websites: lower indexation on search engines, decreased traffic and diminished performances. Targeting undesirable robots can prove difficult in a rapidly evolving landscape with new robots frequently deployed. While crawlers from major AI developers and initiatives can easily be targeted, those from smaller AI developers are not known and cannot be blocked easily. The decision to block AI-driven crawler robots whether for scraping content or powering retrieval-augmented generation (RAG), also presents a dilemma: while it

may safeguard protected materials, it risks excluding websites from LLM-generated answers at a time where AI platforms are becoming a dominant driver of user traffic.

3.2. Opt-out procedures

Other jurisdictions like the EU or the UK have attempted to strike a balance between innovation and data holder protection with conditional copyright exceptions for AI. The EU's Digital Single Market directive allows research organisations and cultural heritage institutions to conduct AI-training through TDM of copyrighted content for scientific research, provided they have lawful access. For other (including commercial) uses, TDM is permitted unless data holders have expressly opted out.

In theory, such a framework provides flexibility and supports both creation and innovation. In practice, the volumes of data involved in AI training, the diversity of sources, the black box nature of AI algorithms and the impossibility to "unlearn" data make enforcement challenging. Varied legal and technical approaches have been defined to express and enforce data holders' opt-out decisions: opt-out declaration, metadata or registries for instance. An opt-out solution should satisfy four conditions to be effective:

- **Once and for all AI bots**: opt-out decisions apply to all AI bots and processing.
- **Related to content**: the opt-out decision can be defined for each specific content on a website or in a database.
- **Can be standardized**: the solution can be standardized and used by everyone.
- **Licensing conditions**: the solution allows the data holder to define precise licensing conditions (e.g. commercial/non commercial use, remuneration, royalties).

The table in Appendix 2 analyses some of the solutions and standards which have emerged in the industry over the past years. Data holders have increasingly expressed their opt-out preferences regarding scraping and AI training whether through Terms and Conditions or machine-readable files (e.g. txt, xml) on their websites. The use of robots.txt files, in particular, provides a standardized way for publishers to indicate crawler restrictions for all or specific areas of their websites. Major outlets such as *The Guardian*, *Le Monde*, *Medium*, and *Reddit* have adopted this approach banning certain robots from crawling and scraping their content. Large scraping initiatives like Common Crawl have pledged to comply with these preferences. However, robots.txt targets general crawling without specific conditions for AI usage, and requires to block each robot individually. Spawning.ai's ai.txt (Li et al., 2025), W3C's TDMRep and Really Simple Licensing (RSL) all propose solutions tailored for AI with granular preferences by file types, authorised AI processing and licensing conditions. HTML tags and HTTP headers offer another possibility to embed scraping and AI opt-outs in website code via meta-tags. Under these solutions, data holders' AI preferences are not attached to the content and can be lost when files are downloaded, texts are copied or images are screenshot.

Embedding AI usage preferences directly into media metadata allows for permissions to travel with a file whether it is transferred, copied, or downloaded. Standards like TDMRep and C2PA, both built on the ODRL framework, enable data holders to specify

granular AI usage preferences and embed them in the media files metadata, including PDFs, EPUBs, images, and videos. C2PA goes further by attaching full media credential, covering origin, authorship, and edit history. Such metadata credentials help verify authenticity and identify data holders. Metadata approaches offer clear advantages for AI developers who can efficiently filter non-compliant content. However, metadata remains vulnerable to being stripped or erased. Additionally, certain types of copyrighted content, such as plain text, lack the capacity to carry metadata altogether, limiting the effectiveness of this solution.

Another approach relies on public registries listing opted-out authors and works. The French Graphic Art Author Right Association (ADAGP) has published a web page compiling all authors who have opted-out from AI processing. However, such web pages lack standardization and do not clearly identify protected works. Spawning.ai's "Do Not Train" registry (DNTR) improves on this by cataloging authors and works via URLs. Like other registry-based solutions, it loses effectiveness once content is copied, downloaded, or screenshot and adoption remains limited. ISCC codes offer an innovative solution where any type of content can be hashed in a unique code. ISCC codes for opted-out works can be kept on public registries, AI developers with uncertain content can obtain its corresponding hash code and compare it to codes listed on public registries.

The multiplication of opt-out methods have fostered a confusing landscape with no clear and widely-adopted standard. Such a situation makes compliance difficult for AI developers and can be used as a justification for non-compliance. Moreover, these approaches remain purely declarative, and fail to protect contents from malicious actors determined to plunder websites.

3.3. Privacy-Enhancing Technologies

Privacy-Enhancing Technologies (PETs) are defined as "*collection of digital technologies, approaches and tools that permit data processing and analysis while protecting the confidentiality, and in some cases also the integrity and availability, of the data and thus the privacy of the data subjects and commercial interests of data controllers*" (OECD, 2023). While PETs have traditionally been developed to safeguard personal privacy, these technologies can be leveraged to protect any sensitive data, including trade secrets, personal information, and sensitive government intelligence. PETs can support confidential data sharing for AI in two ways: first they allow to train, fine-tune and test AI models while maintaining them secure and confidential, second they can support safe collaborative development and sharing of AI models (OECD, 2025c). Such dispositions may be mandated by either data holders or AI developers to overcome confidentiality restrictions, mitigate legal risks, enforce greater data control and safety.

For PETs to be deemed effective, they must ensure that reidentification of the original data is impossible for any third party. For personal data, the European Union's GDPR stresses that simple pseudonymization, such as redacting individual names, is insufficient to lift legal restrictions. In its 2025 ruling, the Court of Justice of the European Union reaffirmed that pseudonymisation does not lift a dataset outside of

GDPR if actors in the processing chain have reasonable means to attribute data to individuals and that identifiability risks may vary depending on context and available technologies (*EDPS vs SRB*, 2025). Such rigorous standards can also apply to trade secret and sensitive government data, which must be unidentifiable to be shared freely. This section describes three major types of confidentiality-enhancing technologies: data obfuscation, encryption and federated processing, based on works by the OECD (2023) and Ekitia.

Data obfuscation methods alter the sensitive information contained in datasets making them safer to store and to share across parties. Anonymization tools automatically remove identifying elements from a dataset. Differential privacy strengthens protection by injecting calibrated noise into raw data, guaranteeing that sensitive records cannot be reidentified. Synthetic data generation offers a promising alternative for AI applications by producing artificial datasets that replicate the statistical properties and structure of original data without exposing real individuals or assets. This technique has been used in medical contexts, successfully generating synthetic medical imagery and lab results. However, these methods are not without limitations. Residual reidentification risks persist, while the introduction of noise may degrade a dataset's value for downstream applications.

Encryption methods such as homomorphic encryption enable processing, including AI training and inference, to be performed directly on encrypted data. A data holder can encrypt sensitive datasets, delegate processing to an external entity, and decrypt only the final results, thereby avoiding raw data exposure. However, these methods impose substantial computational overhead, support only a limited set of operations without introducing noise, and may still leak information under certain conditions, posing challenges for large-scale or complex applications.

Federated processing offers a robust AI-development framework to preserve data confidentiality and maintain data holders' control. In these approaches, AI training occurs in a decentralized way directly on data holders' infrastructure. Only final training results (e.g., gradients or updated weights) are shared to a trusted central third-party for aggregation, thus never revealing original data. Secure Multi-Party Computation (SMPC) further enhances security by aggregating training results in an encrypted way without revealing individual contributions to the central third-party. While federated processing holds significant promises, it often demands high computational resources and advanced infrastructures, which may limit scalability and practical adoption.

3.4. Licensing agreements and smart contracts

In 2024, over 30 licensing agreements had been concluded between major publishers and AI companies, including deals such as OpenAI with Le Monde, the Financial Times, and Condé Nast; Perplexity with Time, Der Spiegel, and Getty Images; and Mistral with AFP (Guaglione, 2024). These agreements grant AI companies formal access to publishers' wide ranges of materials for training, fine-tuning, and retrieval-augmented generation (RAG) in exchange for compensation and other benefits. Licensing agreements offer several advantages for both data holders and AI

developers. Data holders take back control and can impose contractual terms to share their data: confidentiality levels, modalities such as confidentiality-enhancing technologies, authorized usages (only RAG, training, etc.), remuneration for their content or author credits. For AI developers, these agreements offer legal certainty, exclusive access to proprietary datasets and archives, streamlined integration of real-time content for RAG, and data quality guarantees.

In the past years, licensing agreements between large publishers and AI developers have reached multi-million-dollar valuations. However, smaller actors have been mostly kept locked out of these opportunities. High licensing costs coupled with the legal and economic complexities of licensing content have created significant barriers for smaller AI developers to sign such agreements. Similarly, smaller publishers and data holders often lack the technical and legal resources to establish such deals, risking exclusion from potential revenue streams and visibility in AI model outputs. The opaque nature of these agreements with limited disclosure on pricing, terms, and exclusivity clauses, raises competition concerns, potentially fostering oligopolistic markets (Federal Trade Commission, 2023; Autorité de la Concurrence, 2024).

Smart contracts can democratize data licensing for AI by automating the negotiation and the execution of standard licensing agreements. They are defined as “digital contracts stored on a blockchain that are automatically executed when predetermined terms and conditions are met” (IBM, 2021). This approach can support more transparent and competitive AI & data markets by reducing legal barriers for smaller players, enabling usage-based pricing and enforcing contractual restrictions automatically. Smart contracts are already a standard on cloud platforms where users can subscribe to plans and pay by usage in an automatic way. Similarly, platforms such as Snowflake’s Data Marketplace have provided datasets under simple smart contracts. Other approaches, like European Data Spaces, have relied on decentralized models (DAO, blockchain) to further empower data holders in sharing and licensing their data. However, several questions remain over the proper valorization of datasets, or the enforcement of contractual clauses once a dataset has been downloaded.

3.5. Data attribution

Some stakeholders have advocated for data attribution solutions which link AI-generated outputs to their most influential training data sources and determine each work’s contribution. If mandated, data attribution solutions could enable fair data holder compensation based on a royalty-like model.

Two primary approaches have emerged: retraining-based and gradient based methods (Hammoudeh & Lowd, 2024). Retraining-based methods, like Shapley values, assess the counterfactual impact of individual training data points through systematic leave-one-out retrainings. While theoretically robust, Shapley values are practically infeasible for large models which would require millions of expensive retraining for each individual data point. Gradient methods, such as Influence Functions (IFs) avoid retraining by leveraging model gradients. However, Influence

Functions remain computationally expensive with limited application for larger models at scale (Zhu & Cangelosi, 2025).

Private companies like Prorata.ai, claim to efficiently attribute AI models outputs to the relevant licensed data sources and compensate data holders accordingly. Prorata.ai's proprietary algorithm, "Gist Attribution", has never been audited or peer-reviewed.

Despite interesting theoretical properties, real-life applications of data attribution methods remain unfeasible, impracticable or unreliable in the current state of research.

Data-sharing solutions	Transaction frictions	Data quality	Digital self-determination	Confidentiality	Cost and value compensation
a. Crawler robot blocking			✓		
b. Opt-out procedures	✓		✓		✓
c. Privacy-enhancing technologies (PETs)	✓		✓	✓	
d. Licensing agreements and smart contracts	✓	✓	✓	✓	✓
e. Data attribution					✓

Table 2: Ethical data sharing approaches

Conclusion

The rapid ascent of AI technologies has unlocked transformative opportunities across all sectors of the economy: from weather forecasting, to healthcare diagnostics, to automated customer support, scientific research and beyond.

This report has shown that modern AI development heavily depends on data at all points of its lifecycle. However, the AI industry has paid little attention to the provenance and composition of their datasets, often containing large volumes of protected contents with no formal permission or compensation for data holders. While advanced models require ever-increasing volumes of high quality data, publicly accessible data sources are declining. These opposing dynamics have raised concerns about the sustainability of current data sourcing practices.

Data is not a monolith. Its access for AI development is governed by diverse legal statuses, technical and economic constraints. Personal data and trade secrets demand strict confidentiality, while copyrighted content and open data face less restrictions. Similarly, personal, trade secret and copyrighted data holders must give their permission to most downstream usage of their data, unlike open and relevant government data which can be used freely. This heterogeneity underscores the need for contextual solutions, tailored to data types, sectors, and regulatory environments.

Recent years have seen a proliferation of standards and solutions to address ethical data sharing challenges: opt-out mechanisms to enforce data holder preferences, privacy-enhancing technologies to enable confidential sharing, or attribution-based models to incentivize fair remuneration. Despite offering valuable approaches, no single solution appears as the “silver bullet” tackling all constraints.

Rather than promoting predefined answers, VIADUCT, a GPAI associated-project proposes an iterative and exploratory process, structured around sector-specific contexts. Future work will focus on qualifying technical, legal and economic bottlenecks on ethical data sharing and identifying potential levers to address them. Some illustrative areas, such as cultural content, media archives, industrial or environmental data, may serve as entry points for experimentation, though no sectoral focus has been fixed at this stage.

A first step will seek to establish an economic framework to understand data as a “factor of production” of AI, moving beyond the simplistic “raw material” metaphor. Many existing data sharing initiatives have focused on legal compliance or infrastructure, overlooking the economic conditions, incentives and business models which can motivate data collaborations. By framing ethical data sharing as a systemic challenge — legal, technical, and economic — VIADUCT seeks to contribute to the construction of trustworthy, equitable, and operational frameworks for the future of AI.

Bibliography

Autorité de la Concurrence. (2024). *Avis relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative* (Nos. 24-A-05). https://www.autoritedelaconcurrence.fr/sites/default/files/integral_texts/2024-07/24a05_merged.pdf

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*, 2199-2208. https://doi.org/10.1145/3630106.3659033

Barbero, M., & McLaren, J. (2024, mai). *Effective and Ethical Data Sharing at Scale*. Global Partnership for Sustainable Development Data. https://www.data4sdgs.org/effective-and-ethical-data-sharing-scale

Creative Commons. (s. d.). About CC Licenses. *Creative Commons*. Visited on December 1st 2025, à l'adresse https://creativecommons.org/share-your-work/cclicenses/

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases, CONSIL, EP, 077 OJ L (1996). http://data.europa.eu/eli/dir/1996/9/oj

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society, EP, CONSIL, 167 OJ L (2001). http://data.europa.eu/eli/dir/2001/29/oj

Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the Legal Protection of Computer Programs (Codified Version) (Text with EEA Relevance), EP, CONSIL, 111 OJ L (2009). http://data.europa.eu/eli/dir/2009/24/oj

Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the Protection of Undisclosed Know-How and Business Information (Trade Secrets) against Their Unlawful Acquisition, Use and Disclosure (Text with EEA Relevance), CONSIL, EP, 157 OJ L (2016). http://data.europa.eu/eli/dir/2016/943/oj

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.), EP, CONSIL, 130 OJ L (2019). http://data.europa.eu/eli/dir/2019/790/oj

Directive (UE) 2019/1024 du Parlement européen et du Conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public (refonte), EP, CONSIL, 172 OJ L (2019). http://data.europa.eu/eli/dir/2019/1024/oj

EDPS vs SRB, No. Case C-413/23 P (CJEU September 4th 2025). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62023CJ0413

European Commission. (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission. (2020). European Strategy for Data. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

Federal Trade ommission. (2023). *Generative AI Raises Competition Concerns*. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns

Fletcher, R. (2024). *How many news websites block AI crawlers?* Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-XM9G-WS87

Greengard, S. (2025). *Deep web | Definition, Search Engines, & Difference from Dark Web | Britannica*. https://www.britannica.com/technology/deep-web

Guaglione, S. (2024, December 27th). 2024 in review : A timeline of the major deals between publishers and AI companies. *Digiday*. https://digiday.com/media/2024-in-review-a-timeline-of-the-major-deals-between-publishers-and-ai-companies/

Hammoudeh, Z., & Lowd, D. (2024). Training Data Influence Analysis and Estimation : A Survey. *Machine Learning*, *113*(5), 2351-2403. https://doi.org/10.1007/s10994-023-06495-7

Hong, R., Hutson, J., Agnew, W., Huda, I., Kohno, T., & Morgenstern, J. (2025). *A Common Pool of Privacy Problems : Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset* (No. arXiv:2506.17185; Version 1). arXiv. https://doi.org/10.48550/arXiv.2506.17185

IBM. (2021, July 27th). *What Are Smart Contracts on Blockchain? | IBM*. https://www.ibm.com/think/topics/smart-contracts

Information Commissioner's Office. (2022). Data Sharing: A Code of Practice. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/data-sharing-a-code-of-practice/

Kariuki, N. (2025). *Artificial Intelligence Index Report, Chapter 4 : Economy*. Stanford University Human-Centered Artificial Intelligence. https://hai.stanford.edu/ai-index/2025-ai-index-report/economy

Kitchin, R., Lauriault, T. (2014). *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. The Programmable City Working Paper 2*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112

Langlais, P.-C., Hinostroza, C. R., Nee, M., Arnett, C., Chizhov, P., Jones, E. K., Girard, I., Mach, D., Stasenko, A., & Yamshchikov, I. P. (2025). *Common Corpus : The Largest Collection of Ethical Data for LLM Pre-Training* (No. arXiv:2506.01732). arXiv. https://doi.org/10.48550/arXiv.2506.01732

Li, Y., Song, W., Zhu, B., Gong, D., Liu, Y., Deng, G., Chen, C., Ma, L., Sun, J., Walsh, T., & Xue, J. (2025). *ai.txt : A Domain-Specific Language for Guiding AI Interactions with the Internet* (No. arXiv:2505.07834; Version 1). arXiv. https://doi.org/10.48550/arXiv.2505.07834

Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2024). *Datasets for Large Language Models : A Comprehensive Survey* (No. arXiv:2402.18041). arXiv. https://doi.org/10.48550/arXiv.2402.18041

Mehta, I. (2025, June 5th). X changes its terms to bar training of AI models using its content. *TechCrunch*. https://techcrunch.com/2025/06/05/x-changes-its-terms-to-bar-training-of-ai-models-using-its-content/

Mucci, T. (2025). What is data sharing? *IBM Analytics*. https://www.ibm.com/think/topics/data-sharing

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

OECD. (2022a). *Going Digital to Advance Data Governance for Growth and Well-being*. OECD Publishing. https://doi.org/10.1787/e3d783b0-en

OECD. (2022b). *OECD Framework for the Classification of AI systems*. OECD Publishing. https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html

OECD. (2023). *Emerging privacy-enhancing technologies : Current regulatory and policy approaches* (OECD Digital Economy Papers No. 351; OECD Digital Economy Papers, Vol. 351). https://doi.org/10.1787/bf121be4-en

OECD. (2025a). Intellectual property issues in artificial intelligence trained on scraped data. *OECD Artificial Intelligence Papers*. https://doi.org/10.1787/d5241a23-en

OECD. (2025b). *Mapping relevant data collection mechanisms for AI training* (48e éd., OECD Artificial Intelligence Papers) [OECD Artificial Intelligence Papers]. https://doi.org/10.1787/3264cd4c-en

OECD. (2025c). *Sharing trustworthy AI models with privacy-enhancing technologies* (38th ed., OECD Artificial Intelligence Papers) [OECD Artificial Intelligence Papers]. https://doi.org/10.1787/a266160b-en

OECD Committee for Scientific and Technological Policy. (2021, January 20th). *Recommendation of the Council concerning Access to Research Data from Public Funding*. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347

OECD Committee for Scientific and Technological Policy, OECD Digital Policy Committee, & OECD Public Governance Committee. (2021, October 6th). *Recommendation of the Council on Enhancing Access to and Sharing of Data*. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463

OECD Digital Policy Committee. (2008, April 30th). *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information*. OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0362

Rahman, R. (2024, June 19th). *The size of datasets used to train language models doubles approximately every six months*. Epoch AI. https://epoch.ai/data-insights/dataset-size-trend

Reisner, A. (2023, August 19th). Revealed : The Authors Whose Pirated Books Are Powering Generative AI. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/

Schaul, K., Chen, S. Y., & Tiku, N. (s. d.). *Inside the secret list of websites that make AI like ChatGPT sound smart*. Washington Post. Visited on October 23rd, at URL: https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Seddik, M. E. A., Chen, S.-W., Hayou, S., Youssef, P., & Debbah, M. (2024). *How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse* (No. arXiv:2404.05090). arXiv. https://doi.org/10.48550/arXiv.2404.05090

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755-759. https://doi.org/10.1038/s41586-024-07566-y

Tarkowski, A., & Zygmuntowski, J. (2022). *Data Commons Primer*. https://openfuture.eu/wp-content/uploads/2022/07/220723data-commons-primer.pdf

Tomé, J., Pacheco, J., & Azevedo, C. (2025, July 1st). *From Googlebot to GPTBot : Who's crawling your site in 2025*. The Cloudflare Blog. https://blog.cloudflare.com/from-googlebot-to-gptbot-whos-crawling-your-site-in-2025/

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). *Will we run out of data? Limits of LLM scaling based on human-generated data* (No. arXiv:2211.04325). arXiv. https://doi.org/10.48550/arXiv.2211.04325

Wan, A., Klyman, K., Kapoor, S., Maslej, N., Longpre, S., Xiong, B., Liang, P., & Bommasani, R. (2025). *The 2025 Foundation Model Transparency Index*.

Zhu, H., & Cangelosi, A. (2025). *Revisiting Data Attribution for Influence Functions* (No. arXiv:2508.07297; Version 1). arXiv. https://doi.org/10.48550/arXiv.2508.07297

Appendix 1: Stakeholders interviewed as part of this

Participant	Organisation	Position
Alexandra Bensamoun	Paris Saclay	Law professor specialized in digital regulation and IP law, missioned by French Ministry of Culture
Cédric Manara, Sarah Clédy	Google	Head of Copyright and Gov Affairs & Public Policy Manager
Florent Rimbert	Alliance de la Presse d'Information Générale	Responsable pôle développement numérique
Paul Keller	Open Future	Director of Policy
Laurent LeMeur	EDRLab	Director, CTO, EDRLab
Yann Dietrich	Atos	Head of IP
Agata Ferreti	IBM	AI Alliance Lead for Europe
Pierre Gronlier	Gaia-X	Chief Innovation Officer
Till Klein	AppliedAI Institute for Europe	Trustworthy AI lead
Tom Vaughan	CommonCrawl	Principal Engineer
Robert Kroplewski	Polish Minister of Digital Affairs for Information Society	Plenipotentiary, GPAI expert
Adrien Basdevant	Entropy Law	Lawyer, expert on new technologies, data and innovation
Alexandre Martinelli, Phi Hung Le	La Javaness	CEO and CDO
Etienne Bernard, Alexandre Constantin	Numind	CEO and ML scientist
Anastasia Stasenko	Pleias	Co-founder
Djame Seddah	Inria	Senior researcher in CS & NLP
Romain Azais	Inria	Researcher in applied mathematics
Patrick Armengaud	Inria	PEPR manager
Jonathan Pacifico	Cellenza	ML engineer and Chocolatine LLM developer
Michel-Marie Maudet, Jean-Pierre Lorre, Julie Hunter	Linagora	CEO, research director and NLP senior researcher
Sebastian Posth	Liccium	CEO
Bertrand Monthubert	Ekitia	Founder and GPAI expert
Laurent Philippe	BoostAeroSpace	CTO
Kai Meinke	DeltaDAO, Pontus-X	Director
Estelle Gueville	Yale University	History researcher
Edmond Baranes	Université de Montpellier	Economics professor

report

Appendix 2: AI opt-out approaches and tools

Opt-out techniques	Details	Once for all AI bots?	Related to content?	Can be standardized?	Licensing conditions?	Comments
.htaccess	This server configuration file can be used to block AI engine robots on the site.	x	x	✓	x	This solution is difficult for website owners to maintain, as a row must be added to each new AI engine robot identified
Bot Blocking Tool	Examples: botscorner.fr, kudurru.ai	x	x	x	x	
Website's Terms & Conditions	Integration into the General Terms and Conditions of the website	✓	x	x	✓	Is not machine-readable
XML	Provision of an XML (theoretically machine-readable) expressing the authors' content use policy (example of the ADAPGP)	✓	x	x	✓	Does not correspond to a standard.
robots.txt	Use of the "robots.txt" file to indicate to known crawler robots that they are not authorized to scrape website content.	x	x	✓	x	Difficult to maintain for website owners, because you have to add a row to each new AI engine robot identified, while maintaining indexing capabilities.
ai.txt	Provision of a "ai.txt" file to be put on the website, to indicate a refusal of use of the site's content by AI engines.	✓	x	✓	✓	Allows website's owners to address all AI engines, but since it is not a standard, only stability.ai and Hugging Face respect it.
Meta Tags	Use of the meta tags "noai", "noimageai" in the web pages HTML or "no-cache", "noarchive" in their HTTP headers to indicate an opt-out or refusal to allow content to be downloaded.	✓	✓	✓	x	
Really Simple Licensing (RSL)	Machine-readable framework for publishers to syndicate content for third-party clients and crawlers in exchange for traffic	✓	✓	✓	✓	
TDMRep	A W3C working group has proposed the TDMRep protocol to express the data use policy in the context of TDM in a generic and "machine-readable" way.	✓	✓ EPUB, PDF x Other	✓	✓	Based on the ODRL ontology
C2PA	Use of C2PA metadata	✓	✓	✓	✓	Partly based on the ODRL ontology
Web page	Creation of a web page listing authors opposing the use of their works (example with the ADAGP)	✓	x	x	x	It is difficult to make the link between a content and the listed authors.
Do Not Train Registry (DNTR)	Central repository of media URLs for which rightsholders have expressed AI training opt-out	✓	✓ images, videos, files with URLs x Other	x	x	Limited adoption with Stability AI and Hugging Face. Only works for media with URL available.

ISCC registry	Use of the ISCC standard	✓	✓	✓	x

Appendix 3: Glossary

Data governance: "Diverse arrangements, including technical, policy, regulatory or institutional provisions, that affect data and their creation, collection, storage, use, protection, access, sharing and deletion across policy domains and organisational and national borders. Efforts to govern data take many forms. They often seek to maximise the benefits from data, while addressing related risks and challenges, including to rights and interests." (OECD, 2022a, p.13)

Data holder: "Party who, according to domestic law, is competent to decide about the contents and use of (personal and non-personal) data regardless of whether or not such data are collected, stored, processed or disseminated by that party or by an agent on its behalf"(OECD, 2019, p.35). Depending on the data governance regime, a data holder may be (1) a data subject, (2) an individual owning a copyright, (3) an organization owning a copyright, (4) an organization holding proprietary data, (5) a public sector body holding government data.

Data intermediary: "Entity enabling data holders to share their data or gathering data, so it can be re-used by potential data users. They may also provide additional added-value services such as data processing services, payment and clearing services and legal services, including the provision of standard licence schemes" (OECD, 2019, p.36). Examples may include scraping initiatives like CommonCrawl, data sharing platforms like Hugging Face or data marketplace like Snowflake's Data Marketplace .

Data processing permission: Any freely given, specific, informed and unambiguous indication of the data holder's wishes, by a statement or by a clear affirmative action, signifies agreement to the processing, including AI processing. In the case of copyrighted content, permission can take the form of a licensing agreement; for personal data, the GDPR refers to permission as "consent".

Data sharing: "The process of making an organization's data resources available to multiple applications, users and other organizations." (Mucci, 2025)

Data user: "Party responsible to generate social and economic value by leveraging shared data for use cases such as analytics or AI model training ("AI developer"). Within the EU's GDPR, data users are identified as "controllers" who decide on the data processing and relevant legal grounds, and "processors" who perform the processing on behalf of the controllers. They may include (1) consumers, who directly access data about them that are controlled by businesses; (2) citizens, who access public-sector data made available by governments via their open data initiatives; (3) researchers that access scientific data made available via open science project; and (4) businesses that access data provided through e.g. data partnerships, open data or data portability initiatives." (OECD, 2019, p.35)

Ethical data sharing: a set of technical, legal, economic and institutional arrangements which support compliant, trustworthy and fair sharing of data, for all involved stakeholders, aimed at both commercial and non-commercial reuse, including AI.