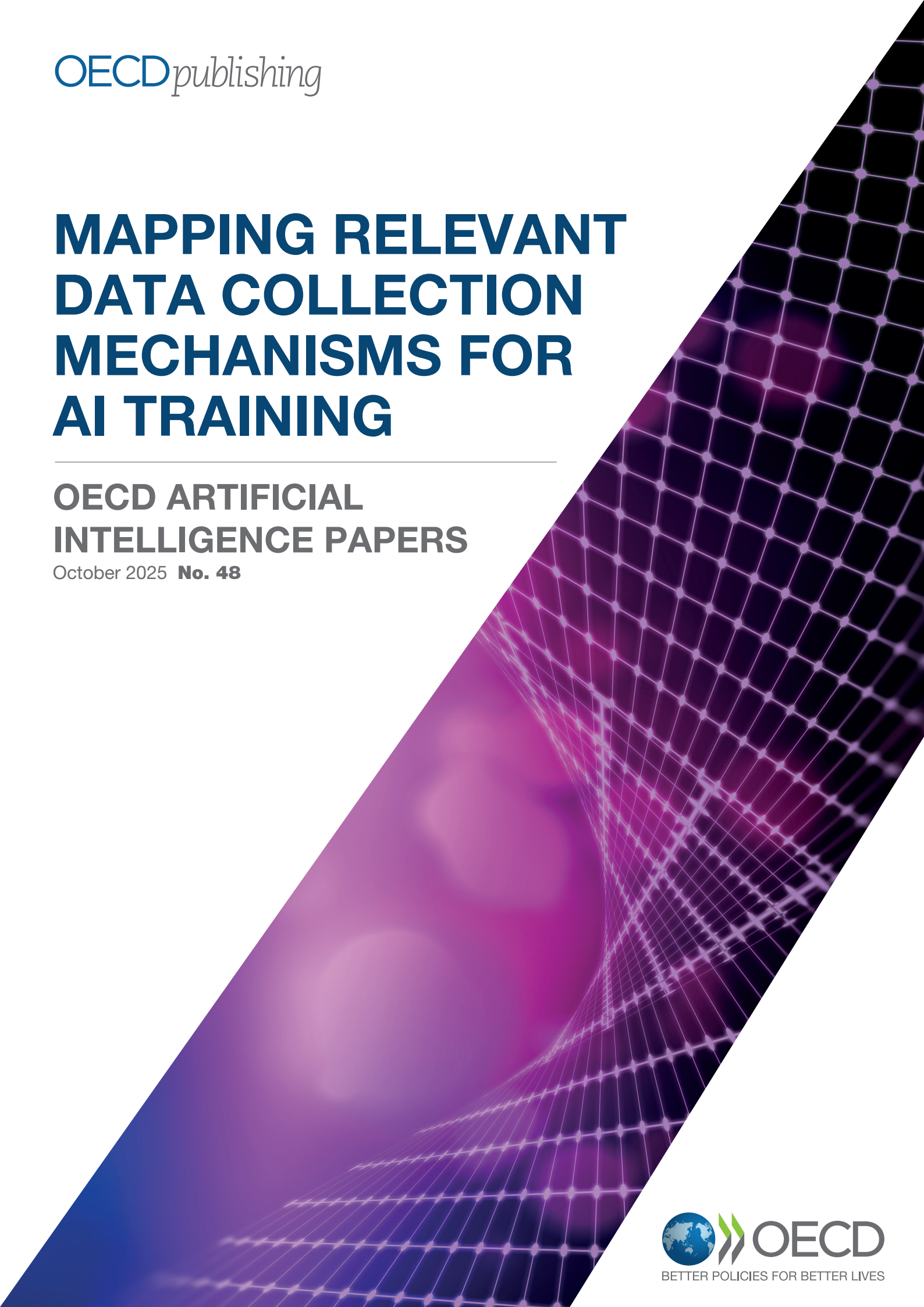
OECD publishing

MAPPING RELEVANT DATA COLLECTION MECHANISMS FOR AI TRAINING

OECD ARTIFICIAL INTELLIGENCE PAPERS

October 2025 **No. 48**



OECD

BETTER POLICIES FOR BETTER LIVES

OECD Artificial Intelligence Papers

Mapping relevant data collection mechanisms for AI training

Foreword

This paper was written by Sergi Gálvez Duran, under the direction of Clarisse Girot, with support from Christian Reimsbach-Kounatze. Editorial review and assistance for publication were provided by Andreia Furtado. It incorporates feedback from delegates of the OECD Digital Policy Committee (DPC) as well as the delegates from its Working Party on Data Governance and Privacy (DGP). The Global Partnership on Artificial Intelligence (GPAI) also discussed this work during its 2024 Fall Plenary. The author gratefully acknowledges valuable comments from Limor Shmerling Magazanik, Sarah Bérubé, Antonia von Born-Fallois, Celine Caira and Kasumi Sugimoto. The paper benefited significantly from the insightful contributions of the members of the OECD Expert Group on AI, Data, and Privacy. This paper was approved and declassified by written procedure by the Digital Policy Committee (DPC) on 18 July 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/DGP(2025)4/FINAL

Abstract

When developing AI systems, practitioners often focus on model building, while sometimes underestimating the importance of analysing the diverse data collection mechanisms. However, the diversity of mechanisms used for data collection deserves closer attention since each of them has different implications for AI developers, data subjects, and other rights holders whose data has been collected. This policy paper maps the principal mechanisms currently used to source data for training AI systems and proposes a taxonomy to support policy discussions around privacy, data governance, and responsible AI development.

Table of contents

FIGURES

BOXES

Executive summary

The performance and reliability of Artificial Intelligence (AI) models are closely linked to the quality and diversity of the data used in their training. While the technical aspects of model development often take centre stage, **the underlying methods for sourcing and assembling training data are equally relevant.** Different data collection mechanisms bring distinct advantages and challenges, not only for AI developers seeking robust and more representative models, but also for individuals whose data may be included. In practice, AI developers often employ multiple data collection mechanisms concurrently to build comprehensive training datasets. **Understanding these mechanisms is essential for advancing trustworthy AI systems and for addressing privacy and data governance considerations in the development process.**

Accordingly, this paper **maps and proposes a taxonomy of the principal mechanisms currently used to source data for training AI systems. This taxonomy aims to provide a basis for future analysis on the privacy and data governance implications of each mechanism.**

The taxonomy organises these key data collection mechanisms into the following structure:

1. **Data collected directly from individuals and organisations**
    - **Provided and observed data**: a growing volume of training data originates from data submitted by individuals or passively collected during their interactions with AI systems, particularly in business-to-consumer (B2C) settings such as chatbots, virtual assistants, and automated helpdesks. Additionally, some AI developers, such as social media platforms, may leverage data provided or observed from individuals across their broader portfolio to support AI model training.
    - **Voluntary data donations:** although still emerging, voluntary data contributions from individuals or organisations offer the potential to enrich training datasets with diverse, real-world information that may otherwise be difficult to access.
2. **Data collected from third-party providers**
    - **Commercial data licensing:** data licensing agreements with organisations offer another avenue for AI developers to access datasets. Data marketplaces and data brokers play a relevant role as data intermediaries in this ecosystem, offering access to a wide variety of third-party data.
    - **Non-commercial practices:** AI developers may also obtain datasets through non-commercial means. Open data initiatives, encompassing both public and private sector data released under open licenses, are key sources for the development of AI models. Significant contributors in this context are dataset publishers who curate and organise datasets from various sources and make them freely and openly available. Given the need for large and diverse datasets to support AI training processes, data scraping has emerged as a widely adopted data collection mechanism to address these demands.

By developing this taxonomy, **the paper offers policymakers and stakeholders a structured approach for policy discussions on privacy, data governance, and trustworthy AI development**. The output underscores the complexity and variety of data collection mechanisms that AI developers rely on, noting that emerging approaches involving secure processing environments and tools such as Privacy-Enhancing

Technologies (PETs) offer ways to improve the usability of these data collection mechanisms while safeguarding privacy and other rights and interests such as intellectual property. This taxonomy sets the groundwork for further analysis on how to balance the growing demand for AI training data (in terms of volume and variety) while also accounting for privacy and data governance aspects such as data quality and traceability.

1 Introduction

Building AI machine-learning models often (albeit not systematically) requires large volumes of data, which may purposefully or inadvertently include personal data. While specific data needs vary depending on the type and purpose of the model, greater access to data generally enables AI models to perform better as they have the ability to learn from data points and patterns inferred in an iterative process (OECD, 2022[1]).Having varied and high-quality data (e.g. accuracy, completeness, consistency, reliability, validity, timeliness) is equally important in supporting the development of robust and trustworthy AI systems, as better data may help address biases, reduce errors, and limit unintended outcomes.

When developing AI systems, practitioners often focus on model building –including weights and parameters– while sometimes underestimating the importance of analysing the diverse data collection mechanisms. However, the diversity of mechanisms used for data collection deserves closer attention since each of them has different implications for AI developers, data subjects, and other right holders whose data has been collected (OECD, 2022[1]). In that regard, the OECD Recommendation on Artificial Intelligence (hereafter, the "OECD AI Principles") (OECD, 2019[2]) emphasise the importance of respecting privacy and of prioritising certain data collection mechanisms that provide better quality datasets and give data subjects more control over their data compared to other mechanisms. As explored below, methods such as data scraping frequently occur without individuals' knowledge, making it difficult to exercise applicable rights, such as statutory rights to be informed, access, delete, and correct their data.

Previous OECD work, notably "Protecting Privacy in a Data-driven Economy" (OECD, 2014[3]) and "Enhancing Access to and Sharing of Data" (OECD, 2019[4]), categorises data based on how it is collected, distinguishing between volunteered data, observed data, derived data, and acquired data. A data provenance categorisation is also included in the OECD's Framework for the Classification of AI systems (OECD, 2022[1]). As AI systems advance, they require not only large amounts of data but also diverse and high-quality datasets. This dual need raises a key question: how can we facilitate access to and sharing of data for AI training –both in terms of volume and quality– while safeguarding individuals' data interests?

In this context, in 2024 the OECD decided to carry out a structural analysis on the various mechanisms used to collect personal data for training AI systems. Such analysis is especially relevant today, as privacy and data governance frameworks are increasingly shaping the ways AI practitioners collect data for building AI systems. It will also remain important in the future, as, despite the uncertainty surrounding the exact forms of AI to come, they will continue to rely heavily on data.

Building on previous OECD work and key multi-disciplinary insights from the OECD.AI Expert Group on AI, Data and Privacy, this paper presents a selection of commonly used data collection mechanisms for AI training. It serves two objectives. First, it outlines the key data collection mechanisms used in AI training, highlighting their main attributes. The goal is to provide policymakers, regulators, and practitioners with the background needed to better understand both the potential benefits and challenges inherent to each mechanism. Second, it lays the foundation for future research on the privacy and data governance implications of these mechanisms, as well as policy initiatives that seek to balance privacy and data governance needs with the growing demand for data in AI training.

2 Data collection mechanisms for AI training

Governments around the world are shaping diverse policies to support and guide the development of trustworthy and human-centric AI systems. In June 2024, the European Union passed the Artificial Intelligence Act (European Union, 2024[5]), and the European Commission recently released the second draft of the General-Purpose AI Code of Practice (European Comission, 2024[6]). Korea's National Assembly has also approved the "Act on the Development of Artificial Intelligence and the Establishment of Trust" (Korea National Assembly, 2024[7]), which is set to take effect in January 2026. In December 2024, the Brazilian Federal Senate passed Bill No. 2338/2023 (Brazil Fenderal Senate, 2024[8]), which also aims to regulate the development and use of AI. In May 2025, Japan enacted the Act on the Promotion of Research, Development and Utilization of Artificial Intelligence-Related Technologies (Japan National Diet, 2025[9]), which emphasizes promoting AI innovation while managing risks through government coordination and voluntary industry compliance. A common thread across these policy frameworks is the recognition of data governance as relevant for ensuring the development of trustworthy AI. In this context, international policy initiatives such as the G7 Voluntary Code of Conduct (G7, 2023[10]) stress the importance that AI developers implement measures that (i) enable dataset traceability; (ii) maintain high-quality training data; (iii) safeguard both personal data and intellectual property rights; and (iv) enhance transparency and accountability by publishing assessments of the AI model's impact on privacy and personal data protection. The OECD played a pivotal role in advancing the G7 Code of Conduct and the Reporting Framework the latter of which was launched as an online platform to monitor the voluntary adoption of the Code of Conduct by organisations in February 2025 (OECD, 2025[11]).

A key aspect of AI data governance is the process of collecting data for building AI training datasets. The growing size and complexity of modern large-scale AI models, particularly Large Language Models (LLMs),[1] requires extensive datasets for effective training. Most notably, public domain[2] or licensed research datasets may not always be sufficient to support the scale required for training most LLMs today (Lee, 2023[12]). As a result, the size of LLMs creates data-sourcing challenges. For example, a major concern is that modern LLMs are primarily trained on data scraped from the web (Feder Cooper, 2023[13]). Numerous lawsuits and investigations have been initiated worldwide, alleging that these datasets were collected in violation of applicable copyright, privacy, and data protection laws, which have yet to be resolved.

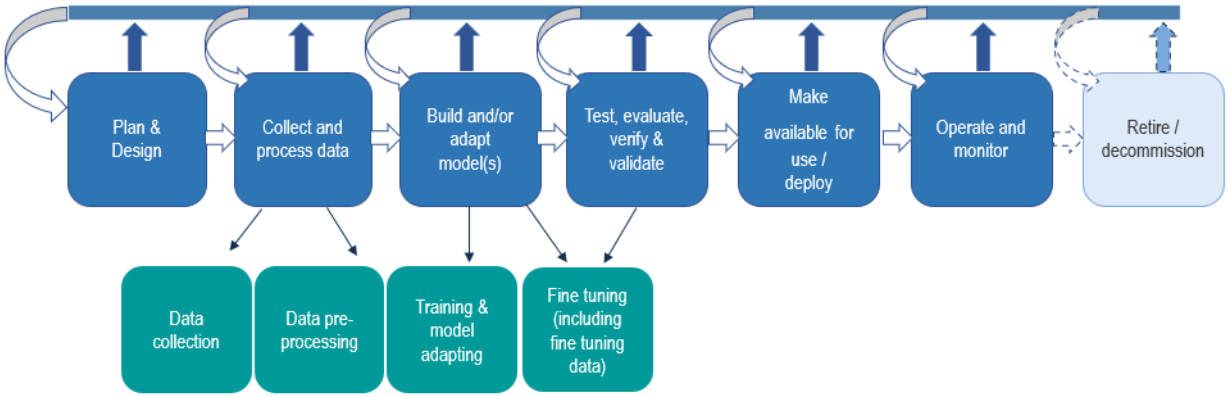
While the attention of policymakers has been largely focused on data scraping practices for LLMs, data scraping is not the only way to collect large amounts of data today. As no single marketplace or centralised repository exists for obtaining training data, firms developing AI models rely on multiple data sources. Openly licensed training datasets are another popular mechanism for obtaining AI training data. An example is Project Gutenberg, a digital library that offers over 70,000 free eBooks, primarily consisting of older literary works that are in the public domain (Gerlach, 2018[14]). Each data collection mechanism comes with varying levels of accessibility, interoperability, reliability, and quality – attributes that are not necessarily mutually exclusive. For instance, the high accessibility of openly licensed datasets does not guarantee that the provenance of the underlying data is well-documented. Another aspect of this situation

is that multiple actors often contribute to assembling AI training datasets. For example, the ImageNet dataset was originally created by researchers at Stanford University, who collected millions of diverse images from around the world (Deng, 2009[15]). These researchers also annotated the images, a process that can also be performed by organisations that specialise in labelling and annotating data. Once annotated, the dataset was curated and made publicly available by the non-profit organisation ImageNet (ImageNet, n.d.[16]), serving as a resource for training various AI models on tasks such as image classification, object detection, and facial recognition. A developer might use the dataset as provided by ImageNet or make further modifications to it during the training stage.

There is often limited visibility into the datasets used to build AI models, making it difficult to assess their quality. For instance, the 2024 Foundation Model Transparency Index outlines 23 subdomains[3] that evaluate different aspects of the development and deployment of foundation models, including data access (Bommasani, 2024[17]). Of all the foundation model developers assessed in this index, only one has openly shared its training data (Bommasani, 2024[17]). Notably, between the publication of the first report in October 2023 and the follow-up in May 2024, developers improved their transparency scores across all subdomains except for data disclosure (Bommasani, 2024[17]). The lagging progress in transparency scores around data can be somewhat explained by the legal risks incurred by businesses when disclosing the datasets used to build AI models, as well as concerns over proprietary information, since the quality of a model is often closely tied to the value of the data it is trained on. Beyond the legal complexities and uncertainties related to privacy regulation and intellectual property rights, there can be also concerns regarding market entry barriers to data collection, as some companies may have privileged access to datasets obtained from other activities in digital markets (Competition & Markets Authority, 2023[18]). There is also growing discussion about the stock limitations of publicly available data accessible via the internet as a resource for AI development. Studies suggest that if the current trend of expanding training dataset sizes for LLMs specifically continues, AI models may exhaust the supply of public human-generated text data between 2026 and 2032 (Pablo Villalobos, 2022[19]). This potential data scarcity raises questions about the long-term viability of relying predominantly on publicly available data from the internet for AI training, enhancing the importance of considering a broad variety of data collection mechanisms. That said, there is a growing industry trend toward developing smaller, more efficient foundational models that rely on less data or more carefully curated datasets. As a result, while the potential of “running out of data” is a noteworthy concern, it may not represent an unmanageable barrier.

Decisions about the data used to build the training dataset are taken at the initial stage of the model development. The development of AI machine-learning models typically involves four phases: data collection, pre-processing, training, and model fine-tuning (OECD, 2024[20]). During the data collection process, AI developers decide the quantity and type of data to include in the training dataset, as well as where to collect the data from. The choices made during the development of the dataset will impact the AI model’s output. Next, the data pre-processing phase converts each data point into a usable format for the machine-learning model. These data points are combined and assembled into a dataset for training. The phase of training (or pre-training) is the initial stage where a “foundation model” is built using a broad dataset. This foundation model serves as the base upon which further adjustments and optimisations are made during the fine-tuning phase, which typically requires using domain-specific datasets.[4]

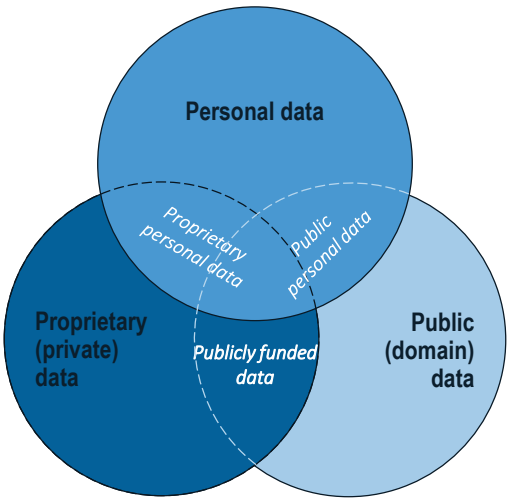
Figure 2.1. The AI model development lifecycle



Source: OECD illustration based on AI system lifecycle (OECD, 2024[20]) (OECD, 2025[21])

Both training and fine-tuning datasets may include data that span overlapping domains. (OECD, 2019[4]) distinguishes between the following three domains of data: i) the *personal domain*, which covers all personal data "relating to an identified or identifiable individual" for which data subjects have privacy interests; ii) the *proprietary domain*, which covers all proprietary data for which there is typically an economic interest to exclude others; and iii) the *public domain*, which includes all data free to access and re-use.[5] These three domains are not mutually exclusive and do often overlap as illustrated in Figure 2.2, reflecting the various overlapping stakeholder interests and applicable data governance frameworks (OECD, 2019[22]; 2022[23]). For instance, public data can include data that are not protected by IPRs or any other rights with similar effects, as well as publicly funded proprietary datasets that have been permissively licensed.

Figure 2.2. The personal, proprietary and public domains of data



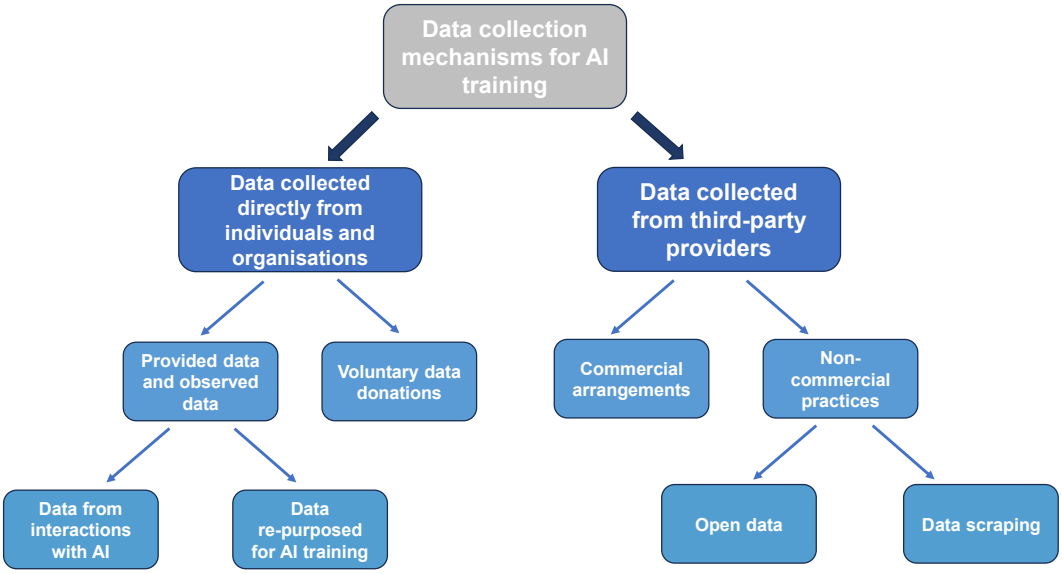
Source: (OECD, 2019[4]) Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, OECD Publishing, Paris, https://doi.org/10.1787/b4d546a9-en

The training phase focuses on the size and diversity of the training dataset because machine-learning models with general capabilities need large and varied data. Large-scale datasets provide models with a

solid base of general knowledge, allowing them to generate a diverse range of outputs and be highly scalable. Pre-trained models may be optionally fine-tuned by using specialised datasets to gain specific capabilities. During the fine-tuning phase, the required data volume is typically smaller than in the training phase, as its purpose is to tailor a pre-trained model to specific tasks and domains.

The following sections focus on data collection mechanisms used for both the training and fine-tuning phases of the machine-learning model's development. They outline relevant data collection mechanisms for AI training, with a particular focus on the attributes of each method. This taxonomy first differentiates based on the sources[6] from which AI developers obtain data: i) directly from individuals and organisations; and ii) from third-party providers. Within these two primary data sources, we identify key different data collection mechanisms, outlined in the figure below and discussed in greater detail in the subsequent sections:

Figure 2.3. Key data collection mechanisms for AI training



Note: This figure presents an overview of the key data collection mechanisms for AI training highlighted in the document. It does not necessarily capture all of the different ways of collecting data for AI training – the AI ecosystem is dynamic, and new mechanisms are likely to continue to evolve as new business players and models enter the AI field.

It is important to note that PETs play a relevant role in enabling the collection of data for AI development, reducing data governance and privacy risks. By enhancing the confidentiality of data collection and use, PETs support the effective operation of the mechanisms outlined in this taxonomy. Synthetic data, for example, is increasingly being considered as a means to collect data confidentially (OECD, 2025[24]). This taxonomy focuses specifically on mechanisms for collecting real-world data, which often form the basis for generating synthetic datasets.

3 Data collected directly from individuals and organisations

3.1. Provided data and observed data

Provided data refers to information originating from direct actions taken by an individual, whereby the person is fully aware of the actions that lead to data origination (OECD, 2014[3]) These data are typically collected through direct user actions such as filling out forms, signing up for services, providing feedback, or uploading content online. It might include a variety of data types, from basic personal information to more complex inputs like preferences, ratings, or user-generated content (e.g., photos, text, and videos). While the individuals concerned may be unaware of the implications of providing these data, the fact that these data are being created should be obvious – or at least intuitive (OECD, 2014[3]).

Observed data are data which have been observed by others and recorded in a digital format. These data can be recorded either at the moment of their creation or transmitted to a digital carrier after observation (OECD, 2014[3]). This type of data collection is typically done through digital tracking (e.g., online cookies, location tracking apps) or physical surveillance (e.g., CCTV cameras). While individuals may be made aware of the creation of observed data (e.g., due to active engagement), much of the creation of observed data may go unnoticed.

In the context of AI systems, the distinction between provided and observed data is often blurred. When individuals interact with AI systems, both types of data are commonly collected and processed simultaneously. For instance, when a user enters a prompt into a chatbot (provided data), the system may also record metadata such as timestamps, interaction duration, device information, or user engagement patterns (observed data) during the same session. This practical overlap means that, in real-world applications, it can be difficult to draw a strict line between the two categories. Therefore, a more detailed analysis at the subcategory level is often necessary.

Accordingly, in the next section we distinguish between two scenarios: firstly, data provided by or observed from individuals during direct engagement with AI services or products; and secondly, data originally collected in the context of other digital activities but subsequently used for AI training.

3.1.1. Data provided by or observed from individuals when engaging directly with AI systems

An increasingly relevant mechanism for training AI systems is the use of information generated from prompts in business-to-consumer (B2C) AI tools such as chatbots, virtual assistants, and automated customer service systems. These AI chatbots increasingly generate “interaction data” – content produced when an AI system engages with a user. As AI systems become more integrated into conversational settings, the volume and importance of this interaction data continue to grow. Interaction data can also be acquired through business-to-business (B2B) integrations with hosted services, where AI functionalities

are embedded into products via direct partnerships or API access. For example, ChatGPT is integrated into Microsoft Bing search through a strategic partnership between Microsoft and OpenAI.

Access to interaction data is not limited to companies operating under these B2C and B2B business models. This kind of data, which is particularly useful for fine-tuning AI models, is available to AI developers through marketplaces such as promptbase.com.

In addition to interaction data, companies offering AI services and products also have access to user-provided or observable data, such as account-level information or metadata. These companies might find it advantageous to leverage this type of information to train their AI models.

3.1.2. Data (provided and/or observed) re-purposed for AI training

In addition to data collected during the use of AI tools, some AI developers may also have access to data (either provided by users or observed from their interactions) from their other digital services and may seek to leverage this data to train AI models. This is the case for owners of large platforms that host user-generated content (FTC, 2024[25]). These platforms have very powerful business incentives to use content such as social media posts, comments, reviews, purchase history, or user feedback to improve their AI models.

By leveraging *user-provided data*, AI developers can design systems and processes that presumably respond to real-world needs and preferences. This data helps improve the performance of AI applications in industries like healthcare, education, finance, or travel, where personalisation and responsiveness to individual needs are relevant. For example, AI-powered features offered by Booking.com to personalise travel planning are trained using anonymised data from hotel reservations and bookings details (Goldenberg, 2021[26]).

Observed data, particularly when collected through online tracking technologies plays a relevant role in training AI models designed for personalised experiences. AI systems learn from this data to understand individual preferences and predict future behaviour. For example, e-commerce platforms like Amazon feed customer shopping data –such as preferences, searches, and browsing behaviour– into their LLMs to deliver more personalised product recommendations (Levine, 2024[27]).

In some instances, this information may have originally been collected for other digital services, not AI model training. Whether such secondary use is permissible depends on the privacy and data protection laws of the applicable jurisdiction, and typically requires an assessment of the relevant legal basis for processing. For example, Brazil's *Autoridade Nacional de Proteção de Dados Pessoais* (ANPD) issued a preventive measure on July 2, 2024, requiring Meta to immediately stop using personal data of Brazilians from its social media platform to train its generative AI model, citing concerns about the legal basis for such processing under Brazil's data protection law (Badillo, 2024[28]). A forthcoming OECD paper will provide further analysis on these considerations.

3.2. Voluntary data donations

Personal or non-personal data may be voluntarily provided by data subjects or by data holders for AI training purposes, without compensation, for the benefit of society. This practice is sometimes referred to as "data donations" or "data altruism", as in legislation such as the European Union's Data Governance Act (DGA) (European Union, 2022[29]).[7] It involves individuals (or organisations) giving consent to share their data with researchers, non-profits, and other entities for purposes like improving healthcare or enhancing public services (Kirstein, 2023[30]).

While these practices may raise questions,[8] data donations can also play a relevant role in training AI systems by providing access to varied, real-world data that would otherwise be difficult or expensive to

obtain. For instance, in the field of healthcare, environmental monitoring, and education, voluntary data donations can help create richer and more comprehensive datasets (Hirsch MC, 2020[31]). This aligns with the OECD AI Principles which encourage governments to promote data trusts and other mechanisms to support the safe, fair, legal and ethical sharing of data. It also aligns with the OECD Recommendation on Enhancing Access to and Sharing of Data (OECD, 2021[32]), which supports different types of data collaborations and the harnessing of new and existing data sources to foster data driven scientific discovery and innovations across the private and public sectors.

In fields like healthcare and genomics, access to large-scale datasets is crucial for developing AI-driven solutions. In this context, data donations may be used to allow people to contribute their personal data (e.g., genetic information, or health data) to public interest research initiatives, speeding up discoveries in drug development, personalized medicine, and epidemiology. An example in this field is the UK Biobank, a large-scale biomedical database and research resource containing genetic, lifestyle and health information and biological samples from half a million UK participants (UK Biobank, n.d.[33]). Interestingly, a study measuring the willingness to share digital health data for research in Germany and Israel revealed that a significant majority of citizens in both countries (82% of Germans and 81% of Israelis) expressed positive attitudes toward the creation of a centralised database for medical research, as long as the data are anonymised (Weisband, 2023[34]) (OECD, 2025[35]). In this context, is worth noting the Data Use and Access Act (DUAA), which received Royal Asset on 19 June 2025. The DUAA is intended by the UK government to enhance the UK's digital strategy and unlock the value of data, harnessing its capabilities to boost public services and contribute to the UK economy. Among its provisions, the DUAA clarifies that individuals can give “broad consent” to an area of scientific research (ICO, 2025[36]).

Similarly, governments and public institutions often lack the data needed to improve public services, develop smarter cities, or tackle societal challenges. It is argued that data altruism can contribute to supporting these efforts by providing the necessary data to train AI models that enhance issues such as public infrastructure, transportation or education. For example, citizens voluntarily sharing data on their energy consumption or transportation habits can help train AI systems used for optimizing city planning, improving traffic management, or reducing energy waste in smart cities. A notable case of data altruism, concerns project DECODE. In this initiative, citizens of Barcelona collected data on noise, air pollution, temperature, and humidity using environmental sensors placed inside and outside their homes. This valuable data was subsequently made accessible to the public through what are known as “data commons” (Sagarra, 2019[37]).

4 Data collected from third-party providers

When data are not directly sourced from the data subject, AI developers seeking data typically acquire it through commercial and non-commercial practices. This section examines these two mechanisms as means of supplying data to the AI economy.

4.1. Data collected from third parties based on commercial arrangements

Commercial data licensing is one way AI developers obtain access to data. For example, OpenAI has signed licensing agreements with news outlets, including The Associated Press (O'Brien, 2023[38]) and Axel Springer (OpenAI, 2023[39]), as well as with social media platforms like Reddit (OpenAI, 2024[40]). Many AI developers, including Meta, have also signed data license agreements with the image distribution platform, Shutterstock (Shutterstock, 2023[41]).

AI developers may also rely on intermediaries, such as data marketplaces, to access third-party datasets. These data marketplaces facilitate the exchange of data between suppliers and AI developers. For illustrative purposes, data marketplaces of this kind include platforms such as defined.ai, appen.com, or AWS Data Exchange.

Commercial data agreements for supporting AI development can also involve business models that monetise data access and sharing, such as data brokers. Data brokers collect and aggregate data from various sources and sell these data to third parties. One important distinguishing factor between data brokers and data market providers is that data brokers are actively engaged in the collection of additional data and their aggregation, while data market providers are passive intermediaries through which data controllers, including brokers, can offer their data sets (OECD, 2019[4]). Data brokers have long been part of the data-driven economy and, in certain jurisdictions, their practices have been analysed by regulatory agencies. One of their current activities is supplying AI-ready datasets for AI model training (PR Newswire, 2024[42]).

Datasets for AI development may also be commercially supplied by businesses as part of their secondary business activities. For example, an agricultural equipment manufacturer may not use the sensor data collected from tractors but can sell it to interested buyers for AI developers working on precision farming solutions. Recently, Bayer signed a partnership with Microsoft to develop an AI model fine-tuned with Bayer's data. This model is intended to offer insights into agronomy and crop protection and will be licensable to Bayer's distributors, AgTech startups, and possibly competitors. Among other things, the AI model will be capable of responding questions on topics such as insecticide components and product suitability for cotton (Bousquette, 2024[43]).

4.2. Data collected from third parties based on non-commercial practices

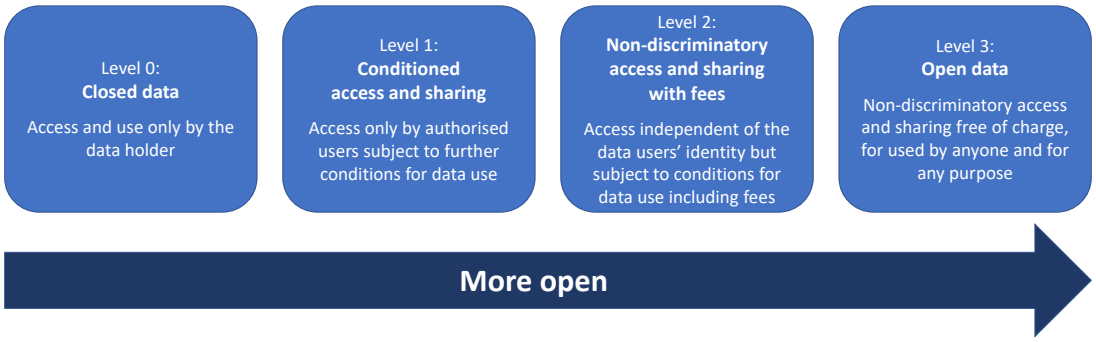
4.2.1. Open data arrangements

Open data[9] has rapidly become a prominent approach to enhance access to data (OECD, 2019[4]). Making data from both public and private sectors accessible under open data licenses is also becoming increasingly important for the growth of the AI economy. In the public sector, examples include the EU's Open Data Directive (European Union, 2019[44]), or the Singapore's Smart Nation initiative (Smart Nation Singapore, n.d.[45]). The Korean government has also taken an active role in creating and releasing AI training datasets to foster domestic innovation, aiming to establish a strong foundation for private sector development of AI models. For instance, it launched the AI Hub platform, which provides open access to a wide array of datasets (AI Hub, 2025[46]). A notable example from the private sector is Microsoft's COCO (Common Objects in Context) dataset, which contains over 330 000 images with detailed annotations for 80 object categories, accessible under a CC-BY 4.0 license for the annotations (Common Objects in Context, 2025[47]).

Advancements in AI development are also facilitated by dataset publishers that curate and organise data from various sources and provide free and open access to it. In this regard, ImageNet, a large-scale image openly licensed dataset (J. Deng, 2009[48]), has been foundational for training many computer vision models. Likewise, recent breakthroughs in protein folding achieved by the AlphaFold deep learning system was only made possible thanks to open access to the Protein Data Bank, which was established almost half a century ago (Liesenfeld, 2023[49]). Text-based repositories like Common Crawl (Common Crawl, n.d.[50])also provide vast quantities of web data for training natural language processing (NLP) models. There are also repositories of fine-tuning datasets that are openly accessible, such as the Hugging Face datasets library (Hugging Face, n.d.[51]).

Although (unrestricted) open data may be desirable, the conditions governing data sharing result in varying degrees of openness, as represented in Figure 4.1 and explained in Box 4.1 below.

Figure 4.1. The degrees of data openness



Source: Based on OECD (2015[52]), *Data-Driven Innovation: Big Data for Growth and Well-Being*, https://doi.org/10.1787/9789264229358-en.

Box 4.1. The degrees of data openness

Data openness is not a 'binary concept'. It is rather a continuum of various degrees of openness, ranging from closed access and use only by the data holder (Level 0 in Figure 3), to:

- Level 1: conditioned access and sharing arrangements where data access and sharing are subject to "terms that 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". For example, bilateral data licensing agreements.
- Level 2: non-discriminatory data access and sharing arrangements, where data can be accessed and shared for fees, but "based on terms that are independent of the data users' identities". For example, paywalled content, such as articles behind a subscription on news websites or datasets sold through online data marketplaces, typically falls into this category. Anyone willing to pay can access the data, regardless of who they are. Initiatives such as the European Health Data Space and national Health Data Hubs, as another example, allow public bodies to provide data access under defined conditions and within secure processing environments. These arrangements may also include fees to cover operational costs.
- Level 3: open data (arrangements) as the extreme form of data openness, which are "non-discriminatory data access and sharing arrangements, where data are 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, 2025[53]) For example, scientific datasets released under Creative Commons licenses, which are freely available to all.

Source: Based on (OECD, 2025[53]), Enhancing Access to and Sharing of Data in the Age of Artificial Intelligence, and (OECD, 2015[52]), Data-Driven Innovation: Big Data for Growth and Well-Being, https://doi.org/10.1787/9789264229358-en

The degree of data openness is contingent upon a variety of factors: licensing and usage restrictions, legal concerns over privacy, data protection, national security, and intellectual property rights. Additionally, the process of preparation, curation, and hosting of large-scale open datasets is expensive. Sustaining high-quality open data requires ongoing investment in both human expertise and technological infrastructure, making long-term financial support a challenge (Baack, 2025[54]).

It is important to note that the data openness framework applies on the one hand, to the original data sources, and on the other hand, to the dataset created from a combination of these data sources. This means a training dataset may include data with varying levels of openness, yet the dataset remains closed. Many popular LLMs have been trained on data with varying degrees of openness (such as web-scraped and licensed data) but their training datasets remain proprietary. A description of the dataset may be available, but the full dataset cannot be accessed or used. For example, OpenAI's GPT-4 technical report states that GPT-4 was pre-trained "*using both publicly available data (such as internet data) and data licensed from third-party providers*" (OpenAI, 2023[55]). However, the training dataset itself remains closed. This may be due to a combination of factors, such as commercial interests in keeping datasets proprietary and intellectual property or licensing considerations related to how data are sourced, organised, and shared. Potential policy responses to support data availability for AI development will be explored in a forthcoming OECD paper.

4.2.2. Scraping of publicly available internet data

Data scraping, also called web scraping, is the automated extraction of publicly accessible web data using a software agent ("bot"). Using software agents for automated web browsing is not new and has been applied in many contexts for years. For example, airline flight comparison sites use "screen scraping" to scan airline websites for prices (Whitaker, 2024[56]). Similarly, "web crawlers" are used by search engines like Google to index online content and link users to relevant pages (Google, n.d.[57]).

From the perspective of how data are captured, screen scraping can be distinguished from web crawling. Screen scraping involves downloading data displayed on websites, while web crawling focuses on linking websites through keyword and metadata analysis (without necessarily downloading the content of those websites). Further analytical differences between these processes are discussed in the paper (OECD, 2025[21]) "Intellectual property issues in artificial intelligence trained on scraped data".

The appeal of data scraping as a data collection tool has led to the widespread use of scraping bots for commercial and non-commercial purposes, including AI model training. Interestingly, traditional machine-learning models were primarily trained using structured datasets curated for specific tasks. LLMs, however, need larger and more diverse datasets to capture the complexities of natural language; therefore, they are usually trained on web-scraped data (Lee, 2023[12]).

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 (OECD, 2025[21]). The growing demand for web-scraped data has also given rise to organisations like Common Crawl or LAION, called AI data aggregators, that scrape data and share it with third parties (OECD, 2025[21]). In fact, AI aggregators play a remarkable role in facilitating the scraped data for AI development; a study analysing 47 LLMs published between 2019 and October 2023 found that at least 64% of them were trained using data from Common Crawl (Baack, 2024[58]).

With respect to the type of data obtained through scraping, it is important to note that scraped content may involve not only personal data directly sourced from or observed about individuals, but also personal data shared publicly online by third parties. For example, images, names, or comments may be posted by social media users about other individuals – such as tagging someone in a photo or mentioning them in a comment. In these cases, the data being scraped refers to individuals who did not themselves upload the content, illustrating that scraped datasets may include personal data originating from both the data subject and third parties.

While data may be accessible on a website, that does not automatically mean it qualifies as open data that can be freely re-used (Global Privacy Assembly, 2024[59]). The legality of data scraping for training AI models is currently under debate in various jurisdictions, with particular focus on intellectual property, cybersecurity, privacy, and data governance issues. Meanwhile, web hosting companies are increasingly implementing technical and legal measures, such as robot exclusion protocols (robots.txt) and explicit prohibitions in their websites' terms of use, to mitigate web scraping activities.

5 Conclusions

This paper highlights the key data collection mechanisms used by firms in developing AI models, focusing on the sources from which data are typically obtained: i) directly from individuals and organisations, and ii) from third-party providers. Within these two broad categories, we identify distinct mechanisms that play a relevant role in the creation of AI training datasets.

An increasingly significant mechanism for AI training is the use of data generated from interactions with B2C AI tools, such as chatbots, virtual assistants, and automated customer service systems. Interaction data are also gathered through B2B integrations, where AI functionalities are embedded into products via direct partnerships or API access.

In addition to the data collected during the use of AI tools, certain developers may also have access to data from their other digital services and may seek to leverage this data for AI model training.

Commercial data licensing is another common practice, where AI developers gain access to data through licensing agreements with organisations that provide third-party datasets. Data marketplaces and data brokers are also important resources for AI developers, offering access to a wide variety of third-party data.

Open data are becoming increasingly important for training AI models. The concept of data openness operates on two levels: the original data sources and the datasets created from these sources. This means that while a training dataset may incorporate data with diverse levels of openness, the resulting dataset itself can remain closed and proprietary.

The growing popularity of data scraping as a data collection tool has led to its widespread use for AI model training. However, the legality of data scraping is currently a subject of debate in various jurisdictions, with key concerns around intellectual property, cybersecurity, privacy, and data governance. In parallel, web hosting companies are increasingly implementing technical and legal measures to limit or prevent scraping activities.

While voluntary data donations are not yet a major source for AI model training and their actual voluntary nature could be challenged in some circumstances, they could eventually play a significant role in AI model training, enabling access to diverse, real-world datasets otherwise difficult and costly to obtain.

It is also important to note that emerging tools such PETs play a relevant role in enabling data collection for AI development. By making data collection and use safer, PETs reduce privacy and governance risks and support the effective operation of the mechanisms outlined in this taxonomy.

These findings underscore the complexity and variety of data collection mechanisms that AI developers rely on, as well as the evolving legal challenges that might come with them. The landscape is constantly shifting, and continued exploration of these mechanisms is essential for shaping policies that balance AI development with privacy and data governance needs.

References

AI Hub (2025), *https://aihub.or.kr/*. [46]

Baack, B. (2025), "Towards Best Practices for Open Datasets for LLM Training", *arXiv:2501.08365*, https://arxiv.org/abs/2501.08365. [54]

Baack, S. (2024), *Training Data for the Price of a Sandwich*, https://foundation.mozilla.org/es/research/library/generative-ai-training-data/common-crawl/ (accessed on 19 August 2025). [58]

Badillo, M. (2024), *Processing of Personal Data for AI Training in Brazil: Takeaways from ANPD's Preliminary Decisions in the Meta Case*, https://fpf.org/blog/processing-of-personal-data-for-ai-training-in-brazil-takeaways-from-anpds-preliminary-decisions-in-the-meta-case/ (accessed on 19 August 2025). [28]

Bommasani, K. (2024), *The Foundation Model Transparency Index v1.1*. [17]

Bousquette, I. (2024), *It's a Legacy Agriculture Company—And Your Newest AI Vendor*, https://www.wsj.com/articles/bayer-microsoft-generative-ai-90754f54 (accessed on 19 August 2025). [43]

Brazil Fenderal Senate (2024), *Brazil's Artificial Intelligence Legal Framework*. [8]

Common Crawl (n.d.), *https://commoncrawl.org/*. [50]

Common Objects in Context (2025), *https://cocodataset.org/*. [47]

Competition & Markets Authority (2023), *AI Foundation Models: Initial Report*. [18]

Deng, J. (2009), "ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. pp. 248-255. [15]

European Comission (2024), *Second Draft General-Purpose AI*, https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts. [6]

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*. [5]

European Union (2022), *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)*, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R0868. [29]

European Union (2019), *Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast)*, https://eur-lex.europa.eu/eli/dir/2019/1024/oj/eng (accessed on 19 August 2025). [44]

Feder Cooper, L. (2023), "Report of the 1st Workshop on Generative AI and Law", *Yale Law & Economics Research Paper*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4634513. [13]

FTC (2024), *AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive*, 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. [25]

G7 (2023), *Hiroshima Process International Code of Conduct for Organizations*. [10]

Gerlach, F. (2018), "A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics", *arXiv:1812.08092*, https://arxiv.org/abs/1812.08092v1. [14]

Global Privacy Assembly (2024), *Concluding joint statement on data scraping and the protection of privacy*, https://www.priv.gc.ca/en/opc-news/speeches-and-statements/2024/js-dc_20241028/ (accessed on 15 May 2025). [59]

Goldenberg, L. (2021), "Booking.com Multi-Destination Trips Dataset", *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2457 - 2462, https://doi.org/10.1145/3404835.3463240. [26]

Google (n.d.), . [62]

Google (n.d.), *Google Search Central*, https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers (accessed on 19 August 2025). [57]

Hirsch MC, R. (2020), "Rare diseases 2030: how augmented AI will support diagnosis and treatment of rare diseases in the future", pp. 740-743, https://doi.org/10.1136/annrheumdis-2020-217125. [31]

Hugging Face (n.d.), . [51]

ICO (2025), *The Data Use and Access Act 2025 (DUAA) - what does it mean for organisations?*, https://ico.org.uk/about-the-ico/what-we-do/legislation-we-cover/data-use-and-access-act-2025/the-data-use-and-access-act-2025-what-does-it-mean-for-organisations/#innovate (accessed on 26 June 2025). [36]

ImageNet (n.d.), *http://www.image-net.org/about*. [16]

J. Deng, W. (2009), "ImageNet: A large-scale hierarchical image database", *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. Miami, FL, USA, pp. 248-255, https://doi.org/10.1109/CVPR.2009.5206848. [48]

Japan National Diet (2025), *Act on the Promotion of Research and Development and Utilization of Artificial Intelligence-Related Technologies*. [9]

Kirstein, L. (2023), *Mobility data as a commons – towards a common mobility data infrastructure*, https://fsr.eui.eu/mobility-data-as-a-commons-towards-a-common-mobility-data-infrastructure/#_ftnref6 (accessed on 19 August 2025). [30]

Korea National Assembly (2024), *Act on the Development of Artificial Intelligence and the Establishment of Trust*. [7]

Lee, C. (2023), “AI and Law: The Next Generation”, https://blog.genlaw.org/explainers/. [12]

Levine (2024), *How Amazon is using generative AI to improve product recommendations and descriptions*, https://www.aboutamazon.com/news/retail/amazon-generative-ai-product-search-results-and-descriptions (accessed on 19 August 2025). [27]

Lieberman, B. (2020), *Climate TRACE to track real-time global carbon emissions*, https://yaleclimateconnections.org/2020/08/climate-trace-to-track-real-time-global-carbon-emissions/. [60]

Liesenfeld, A. (2023), “Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators”, *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI 2023*, https://dl.acm.org/doi/10.1145/3571884.3604316. [49]

Lorenz, P. (2023), *Initial policy considerations for generative artificial intelligence*, https://doi.org/10.1787/fae2d1e6-en. [66]

O’Brien, M. (2023), *ChatGPT-maker OpenAI signs deal with AP to license news stories*, https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a (accessed on 19 August 2025). [38]

OECD (2025), *Enhancing Access to and Sharing of Data in the Age of Artificial Intelligence*, OECD Publishing, Paris. [53]

OECD (2025), *Facilitating the secondary use of health data for public interest purposes across borders*, OECD Publishing, Paris, https://doi.org/10.1787/d7b90d15-en. [35]

OECD (2025), *Intellectual property issues in artificial intelligence trained on scraped data*, OECD Publishing, Paris, https://doi.org/10.1787/d5241a23-en. [21]

OECD (2025), *Launch of the Hiroshima AI Process (HAIP) Reporting Framework*, https://www.oecd.org/en/events/2025/02/launch-of-the-hiroshima-ai-process-reporting-framework.html (accessed on 7 February). [11]

OECD (2025), *Sharing trustworthy AI models with privacy-enhancing technologies*, OECD Publishing, Paris, https://doi.org/10.1787/a266160b-en. [24]

OECD (2024), *Explanatory Memorandum on the Updated OECD Definition of an AI System*, OECD Publishing, https://doi.org/10.1787/623da898-en. [20]

OECD (2023), *AI language models: Technological, socio-economic and policy considerations*, OECD Publishing, Paris, https://doi.org/10.1787/13d38f92-en. [65]

OECD (2022), *Guide for Data Governance Policy-making*, OECD Publishing, Paris, https://doi.org/10.1787/40d53904-en. [23]

OECD (2022), “OECD Framework for the Classification of AI systems”, *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. [1]

OECD (2021), *Mapping data portability initiatives, opportunities and challenges*, OECD Publishing, Paris, https://doi.org/10.1787/a6edfab2-en. [64]

OECD (2021), *Recommendation of the council on Enhancing Access to and Sharign of Data*, https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0463 (accessed on 19 August 2025). [32]

OECD (2019), *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies*, OECD Publishing, Paris, https://doi.org/10.1787/b4d546a9-en. [4]

OECD (2019), *Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies*, OECD Publishing, Paris, https://doi.org/10.1787/b4d546a9-en. [22]

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed on 19 August 2025). [2]

OECD (2015), *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris, https://doi.org/10.1787/9789264229358-en. [52]

OECD (2014), *Protecting Privacy in a Data-driven Economy: Taking Stock of Current Thinking*, https://one.oecd.org/document/DSTI/ICCP/REG(2014)3/en/pdf. [3]

OpenAI (2024), *OpenAI and Reddit Partnership*, https://openai.com/index/openai-and-reddit-partnership/ (accessed on 19 August 2025). [40]

OpenAI (2023), *Axel Springer to deepen beneficial use of AI in journalism*, https://openai.com/index/axel-springer-partnership/ (accessed on 19 August 2025). [39]

OpenAI (2023), *GPT-4 Technical Report*, https://arxiv.org/pdf/2303.08774. [55]

Pablo Villalobos, A. (2022), "Will we run out of data? Limits of LLM scaling based on human-generated data", https://arxiv.org/abs/2211.04325. [19]

PR Newswire (2024), *LexisNexis Unveils Nexis Data+, a Single-API Platform Giving Organizations Unprecedented Access to Gen AI-Approved Licensed News Content and High-Quality Company Data*, https://www.prnewswire.com/news-releases/lexisnexis-unveils-nexis-data-a-single-api-platform-giving-organizations-unprecedented-access-to-gen-ai-approved-licensed-news-content-and-high-quality-company-data-302325319.html. [42]

Sagarra, H. (2019), *Final report on the Barcelona pilots, evaluations of BarcelonaNow and sustainability plans*. [37]

Shutterstock (2023), *Shutterstock Expands Long-standing Relationship with Meta*, https://www.prnewswire.com/news-releases/shutterstock-expands-long-standing-relationship-with-meta-301719769.html (accessed on 19 August 2025). [41]

Smart Nation Singapore (n.d.), . [45]

UK Biobank (n.d.), *https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank*. [33]

US Department of Commerce (2024), *Preparing Open Data for the Age of AI*, https://www.commerce.gov/news/blog/2024/01/preparing-open-data-age-ai. [61]

US House of Representatives (2024), *Bipartisan House Task Force Report on AI: Guiding principles, forward-looking recommendations, and policy proposals to ensure America continues to lead the world in responsible AI innovation*. [63]

Weisband, S. (2023), *GIHF-AI Study 2023, Trust in the use of health data - a comparison between Germany and Israel*. [34]

Whitaker, G. (2024), *Ryanair wins Booking.com 'screen scraper' case*, https://www.aviationbusinessnews.com/industry-news/ryanair-wins-booking-com-screen-scraper-case/ (accessed on 19 August 2025). [56]

Notes

[1] LLMs are advanced machine-learning algorithms developed and refined through a process that encodes information from the input data into their model parameters. This allows them to learn and perform tasks such as translation, content generation, or image recognition (OECD, 2023[65]) (Lorenz, 2023[66]).

[2] Public domain covers all data whose sharing are not restricted by IPRs or any other rights with similar effects. It is therefore here understood more broadly than merely being free from copyright protection (OECD, 2019[4]). The public domain can overlap with the personal and proprietary domains (see Figure 2).

[3] The 2024 Foundation Model Transparency Index includes 23 subdomains that assess various aspects of foundation models, ranging from their development to deployment: data, labor impact, data access, compute, code, environmental impact, model properties, capabilities, limitations, risks, trustworthiness, model mitigations, evaluations, model updates, release process, distribution channels, user interface, data protection, feedback mechanisms, documentation, usage policies, affected geographies, and impact (Bommasani, 2024[17])

[4] In some cases, such as with reasoning models, fine-tuning may instead involve reinforcement learning from human feedback, rather than being trained solely on domain-specific datasets.

[5] See Note 2.

[6] While the taxonomy focuses on data collection sources, it is important to note that the ways in which data are made available can vary significantly across these mechanisms. In some cases, data are acquired outright – for example, through commercial licensing agreements, where the licensee may gain ownership or broad reuse rights. In other cases, data are only accessible under defined conditions, such as within secure processing environments, where it can be analysed but not downloaded or retained.

[7] Under the Data Governance Act, entities can register as “data altruism organisations”, which will manage and facilitate the sharing of data for public good, ensuring that the data are used responsibly and in compliance with privacy regulations.

[8] The voluntary nature of data donations may be influenced by factors such as digital literacy, awareness of data rights, and understanding of associated risks. Participation patterns may vary across demographic groups, and the context of donation requests could affect the genuinely voluntary nature of consent,

[9] ‘Open data arrangements’ refers to non-discriminatory data access and sharing arrangements, where data are 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, 2021[32]).