Received: 6 October 2025 | Revised: 16 January 2026 | Accepted: 11 February 2026

DOI: 10.1002/aaai.70054

SPECIAL TOPIC ARTICLE

# An actionable framework for AI-ready data

**Neil Majithia**[1] | **Thomas Carey-Wilson**[1] | **Elena Simperl**[1,2] | **Nigel Shadbolt**[1,3]

[1]Open Data Institute, London, UK

[2]Department for Informatics, King's College London, London, UK

[3]Department of Computer Science, University of Oxford, Oxford, UK

**Correspondence**
Neil Majithia, Open Data Institute, London, UK.
Email: neil.majithia@theodi.org

**Abstract**

Data is the foundation of AI. Poor-quality data drive up costs and can lead to hidden problems for AI models, especially in complex fields such as healthcare and manufacturing. Meanwhile, biased data negatively affect the performance of AI models, and untested evaluation datasets can result in false positives or overestimates of model accuracy. For data publishers to realize their true potential in supporting the AI ecosystem and its impacts, they should take measures to ensure that their datasets support AI practitioners' needs; in other words, their data should be made AI-ready. In this article, we present a framework for data publishers to follow to make their datasets AI-ready. The framework provides specific, actionable guidance based on previous work and experience at the Open Data Institute and augmented with insights from literature and discussions with a range of experts. We first define AI-ready data before briefly discussing a selection of frameworks in the literature and where they are insufficient. We then provide a visual snapshot of our framework for AI-ready data, and a subsequent in-depth discussion of its criteria. Finally, we demonstrate the usage of our framework with a number of example datasets. We conclude by discussing the further steps that should be taken for the entire open data ecosystem to be made AI-ready in order to realize its true potential in supporting an innovative future.

## INTRODUCTION

Data have a foundational role in artificial intelligence, meaning any inadequacies in datasets can have disproportionately large negative effects on AI development, performance, and reliability. Subtle quality issues such as mismatches in formatting can lead even the most sophisticated AI systems astray via "data cascades", an effect discussed by Sambasivan et al. (2021), in which small data-level errors compound into strong, downstream impacts on the training of a model that undermine its subsequent performance and trustworthiness in high-stakes domains. Similarly, biases present in training data will inevitably surface as biases in AI model behavior (Mehrabi et al. 2021), and poor quality evaluation datasets can lead to overestimates of models' accuracy and abilities (Niven and Kao 2019).

The "garbage in, garbage out" paradigm, therefore, holds: Usage of datasets with inadequate oversight of their properties will result in inadequate AI models and systems. The organizations that train models therefore employ staff and resources to focus on data quality, cleanliness, and assurance, aiming to mitigate any data cascades. However, data publishers should also take steps to ensure that their datasets are designed and packaged in ways that support the needs of the AI ecosystem; in other words, they should ensure that their data are *AI-ready*.

With institutional experience spanning decades of research, our consensus at the Open Data Institute (ODI) defines the AI-readiness of a dataset along several dimensions. It involves the dataset's **technical suitability for machine learning** (ML) (such as being in a proper format and structure), its **overall quality and adherence to standards** (to ensure robustness, consistency, and reliability), its **legal compliance** (addressing licensing, privacy, and consent requirements), and its **responsible collection** (so that the data are representative and gathered in a transparent, accountable way).

However, these dimensions present nothing more than a definition; they do not represent actionable guidance for enhancing the AI-readiness of a dataset. In this article, we propose a framework for AI-ready data that provides concrete steps for data publishers and repository providers to take to achieve genuine, holistic AI-readiness.

In the following sections, we first evaluate the current landscape of data frameworks and their adequacy for the AI ecosystem. We then introduce our framework, built from the combined insights of literature, interviews with domain experts, and our own experience as researchers in the field of data-centric AI. We demonstrate the usage of our framework on a selection of datasets, evaluate it, and finally conclude with a discussion about further steps that should be taken for the entire open data ecosystem to be made AI-ready.

## FAIR AND OTHER FRAMEWORKS FOR DATA READINESS

Existing frameworks that offer guidance on organizational data practices across various domains could be adopted for AI-readiness, but many lack actionable specificity or topical relevance to AI (Table 1).

Take, for example, the FAIR data principles (Wilkinson et al. 2016), arguably the most renowned guidelines for data practices. The principles state that data, metadata, and surrounding infrastructure should be made Findable, Accessible, Interoperable, and Reusable for the global digital ecosystem to truly enable data-led research. Underneath each of these principles, specific criteria are laid out, advocating for data publishers to, for instance, describe their data with rich metadata.

Although the FAIR principles present valuable guidelines for data preparation and publishing, discussions with AI practitioners generally reveal a consensus that the principles are aspirational goals rather than actionable guidance for publishers of AI data, telling users what to aim towards but not how to get there in practice. For example, Koesten et al. (2020) learn a series of GitHub-based publication practices that improve the reusability component of FAIR datasets to provide concrete recommendations for data holders sharing their data on that platform. Furthermore, the principles cannot address AI-specific concerns such as ensuring data is formatted for ML pipelines, addressing dataset bias, or tracking versions of a continuously updated dataset. FAIR was not built for these purposes and, importantly, although the principles therefore do not constitute AI-readiness, they are still important objectives for modern digital infrastructure to design towards.

Recognizing FAIR's insufficiencies in AI contexts, various groups have proposed extensions or new frameworks to build on the principles. For example, the FAIR-R principles were recently introduced by Verhulst, Zahuranec, and Chafetz (2025) to explicitly address AI-readiness, adding another "R" (standing for AI-readiness) principle to the FAIR acronym. The criteria underneath this fifth principle remain conceptual in the paper, offering high-level discussion topics without concrete operational guidance and thereby limiting their applicability.

While FAIR was not created for AI-readiness, more explicit AI-readiness frameworks also face similar limitations. For example, Hiniduma et al. (2025) combine traditional data quality metrics (such as completeness or duplication) with AI-specific ones (such as fairness) into a quantitative assessment for AI-ready data, but this assessment cannot be used as actual guidance for dataset design and publication. Similarly, the Afzal et al. (2020) Data Readiness Report and the Castelijns, Maas, and Vanschoren (2020) ABC framework both provide dimensions for AI-readiness assessment rather than design, with the latter focused specifically on datasets themselves rather than metadata or surrounding infrastructure.

AI-readiness frameworks with more explicit guidance are often designed specifically for application in certain domains and subject matters. For example, the Bridge2AI task force offers actionable, practical guidelines for biomedical datasets to be made AI-ready (Clark et al. 2024) with helpful advice for data publishers, albeit focusing on the sociotechnical environment around a dataset rather than its specific properties. The Schwabe et al. (2024) METRIC framework is also in the field of medical data, although diametrically opposite to Bridge2AI, centring on intrinsic data quality metrics found across medical literature (specifically in the context of bias) rather than taking a holistic view of data practices. Finally, a report by the U.S. Department of Commerce (2025) provides specific action items for making open government data holistically AI-ready; though the report's recommendations are strong and well evidenced, they are limited to their context in government rather than their application to enterprise or scientific data.

In conclusion, despite the number of data frameworks (whether concerned with AI-readiness or not) that have been published, none offer actionable, holistic, and

**TABLE 1** Comparison of AI data readiness frameworks and their criteria.

| Framework | Criteria for data to follow | Critique |
|---|---|---|
| FAIR Principles (Wilkinson et al. 2016) | Findability, Accessibility, Interoperability, Reusability | High-level guidelines rather than actionable guidance, especially for nuances of AI data |
| FAIR-R (Verhulst, Zahuranec, and Chafetz 2025) | FAIR principles; provenance and transparency; quality assurance; interoperable infrastructure; annotation and enrichment; ethical and legal compliance; equitable access, data commons, and AI integration considerations | Remains conceptual rather than actionable |
| AIDRIN (Hiniduma et al. 2025) | Quality; understandability; structural quality; value; fairness and bias; governance; AI-application-specific metrics | Assessment mechanism rather than guidance for publishers |
| Data Readiness Report (Afzal et al. 2020) | Various quality and readiness metrics | Outlines data characteristics and quality transformations without specific instructions |
| ABC Framework (Castelijns, Maas, and Vanschoren 2020) | Understandability; trustworthiness; optimization; transparency | Focused on properties of datasets rather than metadata or surrounding infrastructure |
| Bridge2AI recommendations (Clark et al. 2024) | FAIRness; provenance; characterization; premodel explainability; ethics; sustainability; computability | Focused on the sociotechnical environment surrounding a dataset rather than the dataset/metadata/infrastructure themselves |
| METRIC Framework (Schwabe et al. 2024) | Timeliness; representativeness; informativeness; consistency; measurement process | Explicit focus on mitigating biases |
| AI & Open Government Data Assets (U.S. Department of Commerce 2025) | Documentation; data and metadata format; data storage and dissemination; data licensing and usage; data quality and integrity | Focus on government context |

All frameworks were reviewed for relevance, scope, and applicability in AI contexts. FAIR-R and Bridge2AI include AI-specific enhancements over traditional principles.

domain-agnostic guidance for the purposes of making data more AI-ready.

The consequence of this gap is that organizations still struggle with convenient questions: How can we publish our dataset so that an AI engineer can plug it into a training pipeline with minimal friction? How do we structure and annotate our data so that a future researcher (or even an automated AI agent) can understand its contents and limitations at a glance? How do we keep our published data up-to-date and version-controlled so that AI models built on it remain relevant as the data evolves? These are the kinds of questions not fully answered by existing high-level frameworks. There is a need for a more actionable framework that bridges the divide between high-level principles and the day-to-day reality of preparing data for AI.

## OUR FRAMEWORK FOR AI-READY DATA

To address this need, we have developed a practical framework for AI-ready data (Figure 1) that builds on earlier ideas, focusing on concrete steps that data publishers can take. In creating this framework, our team combined insights from multiple sources: We reviewed the growing literature on data readiness and responsible AI, we interviewed five domain experts (including data scientists, data infrastructure engineers, and AI ethicists), and we drew on over a decade of ODI's experience in open data across various sectors. The result is a framework that centers on making data "AI-ready by design." In other words, instead of treating AI compatibility as an afterthought (something to worry about after data is collected or when a data scientist runs into problems), the framework encourages publishers to bake AI considerations into the entire lifecycle of a dataset—from how it is collected and formatted, to how it is documented and made accessible to others.

A key characteristic of the ODI's AI-ready data framework is its holistic structure. Earlier, we noted the multiple dimensions of dataset readiness, including technical, quality, legal, and ethical aspects. In practice, these dimensions intersect and influence one another. Rather than treating them as separate silos, the ODI framework organizes recommendations into three interconnected components: Dataset Properties, Metadata, and Surrounding

| Category | Criteria | Sub-criteria | Guidance and Examples |
|---|---|---|---|
| 1) Dataset | a) Following international standards and norms | | For example, use ISO-3 codes for countries or ISO-8601 for timing data. |
| | b) Semantic and logical consistency across entries | | As 'heart attack' and 'cardiac arrest' are synonymous terms, only one should be used in a medical dataset. In domain-specific cases like this, labels should adhere to internationally recognised vocabularies such as ICD-10, SNOMED CT, or similar standards. |
| | c) Identifiable class and source imbalance | | CommonCorpus, an aggregate text dataset for AI training, clearly presents the source of each entry. |
| | d) De-identification and Anonymisation where necessary | | Transactions are anonymised with Principal Component Analysis in the Credit Card Fraud Detection dataset to ensure confidentiality without compromising quality. |
| | e) Appropriate file format | | The comma-separated value or .csv (particularly .csv on the Web, or CSVW) format is commonly preferred for structured datasets. Formats like Apache Parquet, or .rdf for graph-based data, also offer advantages. Selection should reflect the intended AI application and interoperability needs. |
| 2) Metadata | a) Machine-readable metadata format | | Using the machine-readable Croissant metadata standard provides discoverability and interoperability for a dataset. |
| | b) The dataset served to users with attached metadata | | API queries for a dataset should return its metadata alongside it, as seen with the WorldPop API. |
| | c) Basic technical specifications | i) Modalities | A dataset should clearly describe the data types (such as text, image, video, time series) within it. |
| | | ii) Dimensionality | A user should know a dataset's number of rows and columns and any nested data or layering. |
| | | iii) Semantics | Defining how columns and rows should be interpreted ensures users clearly understand the data, thereby using it better and more responsibly. |
| | | iv) Bias | The Croissant Responsible AI extension allows authors to describe potential biases in their dataset. |
| | | v) Basic summary statistics | Users appreciate descriptions of dataset column distribution, including calculations of averages, ranges and variances. |
| | | vi) Synthetic data | Any synthetic data-generation methods or machine annotation of data points should be demarcated. |
| | d) Supply chain information | i) Collection | Ontologies like Prov-O or Croissant-RAI enable clear descriptions of datasets' provenance, including their collection and processing. |
| | | ii) Preprocessing | |
| | e) Legal and sociotechnical information | i) Licence name(s) and URL(s) | Dataset usage benefits when clear statements regarding its socio-technical information, including the name of its licence and a URL link, are included in its attached metadata. Statements on intended or permitted users, and notices about data protection, ensure AI practitioners have complete confidence in using a dataset. |
| | | ii) Intended access controls | |
| | | iii) Data protection declaration(s) | |
| 3) Surrounding Infrastructure | a) Accessibility via a user-centric data portal | | Datasets should be sourced from a user-centric data portal like the European Data Portal. |
| | b) Accessibility via API | | RESTful API architectures are industry standard, mainly when exact dataset use cases vary or remain undefined. |
| | c) Version control infrastructure | | Dataset Version Control enables tracking a dataset's entire lifecycle, including post-publication. |

**FIGURE 1** Overview of the AI-readiness framework.

Infrastructure. These components group related requirements but are not independent checkpoints—each reinforces the others. For instance, having an excellent technical infrastructure (like an API or portal) will not make a dataset AI-ready if the data are poorly structured or lacks documentation. Conversely, a pristine dataset in CSV format is useless if no one can find it or trust its provenance. Therefore, the framework's three parts should be seen as a triad of practices that ensure a dataset can be effectively used in AI. Importantly, our interviews with experts reinforced this point: Actual AI readiness comes from aligning data content, metadata, and infrastructure cohesively. Publishers are encouraged to take a "big picture" view of their data practices, rather than checking off isolated boxes.

Another strongly emerging theme is the value of closing the gap between data publishers and data users. Often, those who collect and publish data (e.g., a government open data portal or a research institution) operate at arm's length from the AI practitioners who ultimately use that data. Our framework emphasizes fostering a dialogue between these groups. Publishers can iteratively improve their datasets to better serve AI applications by engaging with the AI community, such as soliciting feedback on what formats or fields are needed or observing how the data are utilized in AI projects. This feedback loop makes AI-readiness an iterative process that can be progressively refined. In practical terms, the framework's recommendations encourage providing clear channels for users to report issues or request enhancements, and being transparent about dataset limitations so that users can account for them in modeling.

In the following sections, we outline the core components of the ODI's AI-ready data framework. For each element (Dataset, Metadata, and Infrastructure), we describe the key considerations and actionable practices that make a dataset AI-ready (and related to FAIR or other frameworks from Table 1). We then illustrate how these components come together in two real-world examples. By adopting these practices, data publishers can significantly lower the barrier for AI developers to find, access, understand, and trust their data, ultimately accelerating innovation and ensuring that AI systems are built on a solid foundation of quality data.

## Dataset properties

The first component of the framework focuses on the properties of the dataset. This concerns the raw material: The data that AI models interact with. To be AI-ready, a dataset should have specific characteristics that optimize it for ML tasks and avoid common pitfalls. Key aspects of the Dataset pillar include:

**Following international standards and norms**: AI-ready datasets follow well-established data standards, schemas, or ontologies relevant to their domain. Using standard formats and codes (e.g., adhering to international standards like ISO date formats, or domain-specific standards for scientific data) makes data more interoperable immediately. Data consumers and AI tools can more easily interpret data that uses shared conventions. Using standardized encodings and reference codes (for instance, a protein dataset using UniProt IDs for proteins, or a bibliographic dataset using ISBN/ISSN for publications) also means models can readily link or integrate the data with other sources. Conformance to standards reduces friction and errors when ML pipelines process data later, in the context of FAIR principles allowing data to be Interoperable (especially according to FAIR principle I1).

**Semantic and logical consistency**: High-quality datasets maintain semantic and logical consistency throughout their contents. This means that terms and labels are used consistently and unambiguously throughout the data. For example, if a dataset uses the term "USA" in one part and "United States" in another to denote country, or mixes numerical codes with text labels inconsistently, an AI model could misinterpret categories or fail to reconcile identical values. Consistency also extends to how data relationships are represented; if hierarchies or linked records exist, these should be noncontradictory and described clearly. Ensuring internal consistency of this kind helps ML systems to recognize patterns in the data more easily by making them explicit. Assuring this aspect may involve validation rules or checks during data preparation, such as verifying that all entries conform to best practice formats like uniform resource identifiers (URIs) (Berners-Lee, Fielding, and Masinter 2005) for data labels and Resource Description Framework (RDF) for expressing data relationships. Alternatively, as raised by FAIR principle I2, data should follow a vocabulary that is itself Findable, Accessible, Interoperable, and Reusable.

**Identifiable class and source imbalance**: Datasets intended for AI must accurately reflect real-world populations or phenomena, or at least document their limitations. In practice, most datasets have some degree of imbalance or bias. For instance, a collection of images might have many more examples of one category than another, or a dataset of economic indicators might heavily feature certain regions of the world over others (in medical

contexts, Schwabe et al. (2024) also raise timeliness as a source of imbalance that leads to bias). AI-ready data practices call for making these imbalances (called class imbalances) explicit and, where possible, mitigating their effects. For example, a dataset should ideally come with basic summary statistics and breakdowns (how many records of each class, distributions of key variables, etc.) so that an AI data publisher can immediately gauge if there is a class imbalance or sampling bias. If certain groups or categories are underrepresented, this should be identified, and the data publisher may take corrective action, such as augmenting the dataset or flagging the issue for AI engineers. In cases where data are intentionally limited (a dataset focuses on a specific demographic or period), that context should be conveyed so that any AI model's applicability is not overstated beyond the data's scope. Not all biases can be eliminated, but knowing they exist allows AI data publishers to compensate, adjust models, or avoid deployment in inappropriate contexts.

**De-identification and anonymization**: The responsible collection of data is crucial for the properties of datasets involving human or sensitive data. AI-ready data should be vetted for privacy, confidentiality, and ethical considerations. This might mean that personal identifiers are removed or anonymized where appropriate (e.g., in a health dataset, patient names and IDs would be stripped or encoded). If a dataset contains personal data, publishers should ensure they have the necessary rights to share it and comply with regulations such as the General Data Protection Regulation (GDPR). Even beyond legal compliance, ethical best practices include obtaining informed consent if data were collected from individuals, or at least being transparent about data sources and potential harms. In an AI context, privacy safeguards are essential not just for legal reasons but also to enable broader use of the data. If data are adequately anonymized or aggregated, it can often be shared more freely, allowing more researchers to benefit without risking individual privacy (Curzon et al. 2021). For data with no personal aspect (for instance, a dataset of scientific measurements or machine logs), this criterion may be less about privacy and more about the ethical use of data. Ultimately, an AI-ready dataset should not inadvertently expose sensitive information, and its publishers should communicate any steps taken to protect the individuals represented in the data.

**AI-ready file format**: Lastly, the practicality of using the dataset in AI workflows depends on the formats and packaging of the data. AI-ready datasets are provided in formats that facilitate ML processing and in a manner that can handle the scale required. For example, large datasets might be offered in a chunked or partitioned format (such as a collection of compressed files or a cloud bucket of images) that data pipelines can stream or download selectively. Tabular data might be provided in Excel or PDF format (human-friendly) and CSV (GOV.UK 2023) or Parquet format (machine-friendly) (Vohra 2016). In some cases, offering multiple formats is ideal: One rich, original format for completeness and fidelity, and another simplified format for quick ingestion into common ML frameworks. Additionally, considerations such as using compression and ensuring reasonable file sizes fall under this category. Of course, if a single file is large (as is so often the case with AI datasets), it may be difficult for local systems to process and display it efficiently. Providing data in a database or via an API can be a solution to this—similarly, recent work on Model Context Protocol architectures may facilitate better data management and access for modern AI workflows (Anthropic 2024). In summary, AI datasets should be packaged for convenience, allowing any AI data publishers to spend minimal effort converting or restructuring the files before model training. Choosing modern, widely supported file formats (and providing documentation about the format) goes a long way toward that goal. Note, in this criterion, two interlinked considerations that must be made. First, domain dictates format: Data should be published in a format that is well-established in the topical domain it sits in, respecting industry standards and what users are accustomed to. Second, and relatedly, file formats are intrinsically linked with research software used to either build the dataset or use it. Intentional design rather than ad hoc publishing means data publishers can take in these considerations and produce data products that are AI-ready in the unique contexts they are in.

This pillar is about making the data as clean, structured, and informative as possible at the source, so downstream AI projects do not have to do heavy lifting to make the data usable.

## Metadata

Even the best raw data can be misused or underutilized without context. That is why the second core component of the framework is Metadata, that is, the information about the data. Comprehensive and well-structured

metadata turns a dataset from a mere collection of numbers or text into a self-describing resource that others can understand, evaluate, and trust. Key practices for the metadata component include the following:

**Machine-readable metadata format**: AI-ready data come with metadata in standard, machine-readable formats. This could be as simple as a CSV file accompanied by a JSON or XML metadata file, or a more domain-specific schema, such as an RDF description for a linked dataset. The crucial point is that metadata should be parseable by software, not just human-readable prose. Established metadata standards or vocabularies, such as MLCommons' Croissant (Akhtar et al. 2024), can facilitate this, supporting the use, reuse, and discoverability of datasets. For example, publishing metadata in JSON-LD or RDF enables search engines and AI agents to discover the relationships and attributes of the dataset automatically. Machine-readable metadata may include the dataset title, description, keywords, column definitions, units of measurement, and other relevant details. Predictably, structuring this information means tools can ingest it; imagine an AI system that can automatically decide if a dataset is suitable for a task by reading its metadata profile.

**Metadata attached to the dataset**: It is essential to serve metadata alongside the data for users—whether human data scientists or AI agents—to have appropriate awareness of the context surrounding a dataset and guide their work accordingly. In practice, this means that when someone accesses or downloads the dataset, they should immediately have access to its documentation and descriptive information. Many open data repositories accomplish this by providing a data catalogue page (including all the metadata), embedding markup in the HTML of that page, and bundling a README or schema file with downloads of the dataset. AI-ready data takes it a step further; wherever the data are stored (portal, API, etc.), the metadata is either embedded or directly linked, unlike in some repositories, such as the Hugging Face datasets API, which requires a separate query method for users to retrieve metadata. For instance, each API endpoint for data could also return metadata about the fields. Or if data are provided as files, including a metadata JSON and a data dictionary ensures that a data scientist does not have to hunt for a separate report or contact the provider just to understand what each column means. In short, the delivery of the dataset always includes its metadata as an attached or easily accessible component. This practice saves time and reduces errors, because users will not mistakenly use a field without knowing its definition or misinterpret codes in the data. FAIR principle A2 provides an important addition to this criterion: Metadata should be accessible even when data it represent are not. This is essential for archiving, provenance, and reproducibility of research, especially in cases where data access is limited (or could, in the future, be limited).

**Basic technical specifications**: High-quality metadata should capture the technical specifics and context that AI practitioners need to know. This means the metadata should record details such as the dataset's structure and size (eg, number of records, number of features, data modality, such as "this is a time-series dataset" or "this is geospatial data with these coordinate reference systems"). It should also include summary statistics or properties that give a snapshot of the dataset's contents. For example, minima, maxima, averages of numeric fields, or counts of categories, as this helps quickly assess suitability. Critically, metadata is where important flags and notes should live. If the dataset has known biases or is missing specific information segments, this should be explicitly noted (e.g., "This survey under-sampled rural areas; the data is primarily urban"). Suppose data in the set has been synthetically generated or augmented (which is increasingly common in AI training sets). In that case, the metadata should indicate which parts are synthetic or how the synthetic data were created. Additionally, metadata can outline quality assurance steps, such as error-checking methods or validation splits. These technical details serve as a user manual for the dataset, enabling an AI engineer to quickly understand how the data may behave during training or what preprocessing might be required. All of these attributes can or will be described through the Croissant vocabulary in its v1.0 (Benjelloun et al. 2024) or upcoming v1.1 (Benjelloun et al. 2025) releases. The more a dataset "describes itself," the less guesswork for the user, which ultimately means faster and more reliable AI development. Reflecting FAIR, these subcriteria specifically tie in with principle R1, requiring meta(data) to be richly described with a plurality of accurate and relevant attributes.

**Supply chain information**: Trust in data often stems from knowing its origin and how it has evolved. AI-ready datasets, therefore, supply rich provenance information and version tracking as part of their metadata, like in Croissant. Provenance informa-

tion includes the data's origin (who collected it, using what methods or instruments, and when). It might outline the data generation process. For example, "these sensor readings were collected by X agency every 10 minutes via satellites and then cleaned with Y method." Clear provenance, like that in the CommonCorpus dataset published by Pleias (PleIAs and Hugging Face 2024), enables users to assess the reliability and potential biases of the data source. Version history is equally important, especially for datasets updated regularly or corrected over time. If a dataset is static (one-off), at least a publication date and any revisions should be noted. If it is dynamic, the metadata should indicate the release or update schedule (e.g., "monthly snapshots"), and ideally provide unique identifiers or version numbers for each release. This way, an AI practitioner can ensure they use the most recent version or can reproduce results by using the same version later on. In more advanced implementations, datasets may have changelogs that highlight changes between versions, such as new records added, errors removed, and other notable updates. Providing this within the metadata (or via links in metadata to a version control system) means that anyone using the data for AI can be aware of changes that might affect model performance. It also supports broader governance, as regulators or auditors can track whether data used in an AI system have been collected, created, or altered, particularly by outsourced entities (Casilli et al. 2024), which is essential for accountability. FAIR isolates the importance of provenance information in the context of reusability with principle R1.2.

**Legal and socio-technical information**: Lastly, metadata should convey the legal and social context for the dataset's use. An AI-ready dataset clearly states the dataset's license or terms of use in its metadata (Carey-Wilson et al. 2025). Rather than burying licensing in fine print or on a separate webpage, the metadata might include a field like "License: CC BY 4.0" and a link to the full license text. This is important because AI developers need to know if they can legally use the data for commercial purposes, such as creating derivative works, especially when training models that might themselves be distributed. Clarity here prevents misuse and encourages (re-)use as per FAIR principle R1.1; an open license signals to the community that the dataare available for innovation. Beyond licensing, metadata can note other relevant societal or policy information: for example, if the data have any usage restrictions ("not to be used for credit decisions" or "only valid for research purposes"), or if there are ethical considerations ("participants consented for non-profit research use"). It is also helpful to include references to any governance frameworks or data protection measures applicable (e.g., "An ethics board reviewed this dataset" or "Data is anonymised as per GDPR guidelines"). Embedding this socio-technical information in metadata ensures that AI practitioners and any automated systems processing the dataset are immediately alerted to obligations and context. The Croissant RAI extension (Jain et al. 2024) recommends noting things like preferred data access protocols and known data protection notes in the metadata, as does the Responsible AI Licenses initiative (RAIL Initiative 2025). All of these help paint a comprehensive picture around the data, allowing it to be responsibly integrated into AI projects.

Overall, the Metadata component of the framework is designed to make the dataset self-describing, transparent, and easy to evaluate. Good metadata tells anyone who encounters the dataset what it contains, how it can be used, and why it can be trusted (or what limits to heed). This rich contextual layer is indispensable for AI readiness; it lets human practitioners and even AI-driven data catalogues automatically reason about the suitability of data, ensuring that models are built on a well-understood foundation.

## Surrounding infrastructure

The third component of the AI-ready data framework looks beyond the data and its documentation to the environment in which the data are made available. Even high-quality data with excellent metadata can falter if the mechanisms to discover, access, and maintain data are not AI-friendly. The surrounding infrastructure encompasses the tools, platforms, and services that connect users (or AI systems) with the dataset. An AI-ready infrastructure ensures that data are released and actively supported in a way that aligns with AI workflows. Key elements of an AI-ready data infrastructure include the following:

**Accessibility via a user-centric data portal**: A well-established approach for publishing AI-ready data is via a well-designed data portal or platform that puts user needs front and center. "User-centric" means that the portal makes it as easy as possible to find relevant datasets and interact with them (Costa, Walker, and Simperl 2020). Specifically, an AI-ready data portal will provide flexible search and browsing capabilities (possibly with

conversational, LLM-based elements), enabling users to quickly locate the data they need through various means. It will also provide tools to preview or visualize the data directly on the platform. For example, a portal might allow a user to view a snippet of a dataset, plot a few fields, or generate basic statistics without requiring a download. In the case of complex data, such as 3D molecular structures or geospatial maps, the portal may include specialized viewers or integration with notebooks to dynamically explore the data. The portal should also present all the dataset's metadata and documentation alongside the download options; essentially co-locating data and its description, so one does not have to search elsewhere for how to use it. A good portal will also consider accessibility and interoperability as per the 10 principles outlined by Costa, Walker, and Simperl (2020): following web standards, ensuring the site is navigable (possibly even providing a multilingual interface targeting global users), and adopting interoperability standards for data exchange. data.europa, for instance, is often cited as a model that incorporates many of these principles. It provides interactive tools to evaluate data quality and relevance. Ultimately, a user-friendly portal lowers the barrier to entry: A data scientist can transition from discovery to initial analysis in minutes, and an AI agent (such as a crawler or automated model trainer) can programmatically retrieve information through standard interfaces and well-structured pages. The Research Data Alliance (RDA) introduces Data Discovery Paradigms (RDA Data Discovery Paradigms Interest Group 2024; Wu et al. 2019) that provide clear principles for a user-centric research repository to take.

**Accessibility via an API**: AI-ready data should be available through both machine-friendly access points and human-friendly portals. The most common form of this is an API (Application Programming Interface) that allows developers to query or download data via code. From an AI perspective, an API can be transformative. Instead of manually downloading and updating a file periodically, a practitioner can write a script to fetch the latest data or pull just the needed subset for training. Well-known AI datasets, such as those hosted by platforms like Kaggle or large text and image datasets, provide a simple Python API call or command-line tool to access the data. Data publishers should aim to provide ease of access by offering a RESTful API or other modern data access services for their datasets (Fielding 2000). An API for an AI-ready dataset should ideally allow for flexible queries (for instance, filtering by date range, retrieving specific fields, or streaming data in chunks) so that users can obtain precisely what they need without unnecessary bandwidth consumption. Performance considerations include avoiding restrictive pagination or very low rate limits, which can hinder AI usage when large volumes are required quickly. Interview feedback from AI engineers suggests that truly AI-friendly APIs avoid needless complexity: A straightforward REST API returning JSON or CSV (or even direct binary data streams for extensive data) is preferred, as it can be easily integrated with data science libraries. In some cutting-edge scenarios, publishers leverage emerging data access frameworks, such as data spaces or cloud data warehouses, to enhance their capabilities. A "data space" is a federated infrastructure where data publishers and consumers connect through standardized protocols, allowing data to be accessed in situ (e.g., via secure connectors) rather than copied over (Open Data Institute 2024b). For instance, a language dataset might be accessible through a shared data space that multiple NLP researchers can tap into, ensuring they always have access to the latest version. Whether via traditional APIs or newer data sharing models, the principle is the same: Make it effortless for AI tools to talk to the data. Users should be able to write a one-liner or use a standard library (such as "Pandas" in Python or a HuggingFace Datasets loader) to import the dataset directly into their analysis or training pipeline.[1]The RDA Artificial Intelligence and Data Visitation (AIDV) Working Group advocates for "data visitation" (RDA Artificial Intelligence and Data Visitation Working Group 2025) models—where algorithms travel to the data rather than downloading it—requiring infrastructure that supports secure processing environments and embedded ethical review protocols. An AI-ready infrastructure must therefore operationalize these community standards, ensuring that APIs and portals are not only accessible to human researchers but are also compliant with emerging automated governance workflows.

**Version control infrastructure**: The infrastructure surrounding a dataset should support the fact that data are often a living asset; it may evolve, be corrected, or change over time. For AI systems, keeping track of these changes is vital (imagine a model that needs retraining on the latest data, or an audit trail to see what data version informed a particular model decision). Thus, AI-ready

infrastructure includes robust version control and update infrastructure for datasets (Sajid 2024). In practice, every time a dataset is updated or modified, there should be a straightforward way to identify that update, such as a new version number, timestamp, or unique identifier. The infrastructure might automatically maintain older versions or their diffs, so one can roll back or compare changes. Many modern data platforms are adopting tools analogous to software version control, such as technologies like DVC (Data Version Control), which track data file changes like how Git tracks code changes (European Commission 2025). Using such tools, data publishers can expose a history of changes: for example, "20 new records added on 2025-05-01; 5 records updated on 2025-06-01 to correct errors." While the needs of the AI ecosystem may be more complex, this basic level of transparency is beneficial for AI practitioners, enabling models to be updated or checked against evolving data. Additionally, if the dataset is updated in real-time or frequently, the infrastructure may support streaming updates or notifications, such as webhooks or feeds that announce new data availability. These features ensure thst the pipeline can continuously incorporate fresh data for AI applications that require the most current data, such as an AI system analyzing social media trends or real-time satellite imagery. Newly generated or continuously updated datasets also play a crucial role in the benchmarking of AI performance. In summary, treating data with the same discipline as software (i.e., managing versions, documenting changes, and enabling updates) is a hallmark of an AI-ready data infrastructure. It gives both confidence and flexibility: confidence that one can reproduce old results if needed (using the same data version) and flexibility to seamlessly integrate improvements or new information into AI models as the data evolves.

These three elements (user-centric portals, easy programmatic access, and strong version/update support) form the backbone of an AI-ready data infrastructure. In essence, they ensure that once a dataset exists (with suitable properties and metadata), it can be found, accessed, and kept in sync with the needs of AI developers. A publisher might achieve this via a single platform that encompasses all (for instance, a well-designed open data portal that offers web UI, API, and tracks versions), or through a combination of tools (maybe hosting data on a repository like GitHub or DataVerse for versioning, while providing an API endpoint for queries, etc.). The exact implementation can vary, but the goals remain consistent. An essential outcome of building such infrastructure is that it fosters a community and ecosystem around the data, allowing users to engage, provide feedback, and contribute to improvements (e.g., suggesting data corrections through the portal or adding notebooks and examples).

Before moving on, it is worth noting that investing in infrastructure pays dividends beyond AI. It generally improves data accessibility for all kinds of users. However, it is particularly crucial for AI because of the scale and speed at which AI projects operate. AI researchers may need to train across millions of data points or rapidly experiment with different slices of data; a clunky download process or unclear update policy can significantly slow this process down. Conversely, a smooth infrastructure can accelerate AI development; consider how well-packaged image datasets, with easy-to-use download APIs and clear documentation, have accelerated computer vision research. The same can be true for any field of data. Thus, building an AI-ready infrastructure is about creating the pipelines that let data flow efficiently from the publisher's hands into the AI applications that can derive value from it.

## USING THE FRAMEWORK: THREE EXAMPLES

### Evaluating the AI-readiness of Re-LAION-2B

LAION-5B, unveiled in 2022 by the nonprofit LAION, broke ground as the first openly licensed web-scale corpus of 5.85 billion image–text pairs scraped from Common Crawl (LAION 2022). Roughly 2.3 billion entries were English while the remainder spanned more than a hundred languages, giving the dataset unmatched coverage for multilingual retrieval and zero-shot transfer studies (Schuhmann et al. 2022). Each row consisted of the image–text pair and an array of measurements, including embedding vectors for the pair, scores in NSFW and aesthetic classifiers, and image metadata like height and width. These measurements would allow researchers to programmatically slice high-quality or safe subsets of the dataset and build custom training corpora for their models. Such functionality meant that LAION-5B enabled the training of Stable Diffusion (Wiggers 2022), arguably the first mainstream image generation AI model.

The dataset was later found to contain toxic and illegal content (Birhane et al. 2023; Birhane, Prabhu, and Kahembwe 2021), prompting its takedown and in-house analysis by LAION. LAION released Re-LAION-5B (LAION 2024) in August 2024, a full re-crawl that blocked certain URL

matches, applied a stronger nude-content model, and removed duplicates in the dataset via perceptual hashes and CLIP embeddings. The cleanup trimmed several hundred million unsafe or near-identical pairs, leaving ≈5 billion high-confidence examples and ships reproducible Docker logs for end-to-end auditing.

Re-LAION augments each image–text pair with ISO-639-3 language tags, OCR scores, and caption-length statistics, further enriching the feature space for data-centric analysis. Licensing moved to Apache-2.0, eliminating share-alike constraints, and the dataset is now hosted on Hugging Face (rather, a collection of 2B versions (Hugging Face 2025) is hosted instead of a 5B dataset).

LAION has already contributed plenty to the AI world, but the feature-rich, assuredly safe, and (most importantly) open nature of Re-LAION means the dataset could be used to expand the capabilities of image-native or multimodal generative AI in the future. However, its full potential will only be reached if it is published in an AI-ready way.

Here, we use the AI-readiness framework to evaluate Re-LAION-2B (specifically, laion/relaion2B-multi-research Hugging Face 2025). We check technical documentation, published literature, LAION blog posts, and the dataset itself for evidence of adherence to each requirement. Annotations describe the extent of any adherence, coding whether a requirement was met, partially met, or not met.

## Dataset properties

**International standards and norms—met**: Samples are annotated with three-letter ISO 639 codes for their language. All image embeddings are generated with CLIP, a generic image embeddings model.

**Semantic and logical consistency—met**: All annotations are made with the same, consistent methodology. Embeddings are generated with CLIP, while p_unsafe measurements are generated with LAION's open source NSFW detection model.

**Identifiable class and source imbalance—met**: The feature-rich nature of the dataset facilitates exploration of imbalances via histogram, although note that no histograms are served to users in the data card or elsewhere. Source imbalance is identifiable via the URL on each row of the dataset.

**De-identification and anonymization—not met**: Re-LAION's safety focus was mostly focused on CSAM. Privacy problems regarding faces, licence plates, and personally identifiable information are not explicitly addressed, making the dataset's use for AI a legal difficulty.

**Appropriate file format—met**: On Hugging Face, all datasets are converted to Apache Parquet shards.

## Metadata

**Machine-readable format—partially met**: Via publishing on Hugging Face, the dataset automatically has a basic Croissant metadata record. However, there is no concerted effort to publish more in-depth, machine-readable metadata elsewhere.

**Dataset served with attached metadata—not met**: The Hugging Face repository does not have a data card, nor a README.md file.

**Basic technical specifications—partially met**: Each data point has a host of technical metadata that contributes to the overall feature space of the dataset. However, due to the lack of metadata overall, basic technical specifications about the overall dataset are not available.

**Supply chain information—partially met**: LAION blog posts and published literature contain full methodologies about data collection and annotation. This fits with the organization's general approach to openness. This is not in the actual metadata that accompanies the data, though.

**Legal and sociotechnical information—met**: The dataset's licence is stated on its Hugging Face page, within keywords. Information is also available on the LAION website.

## Surrounding infrastructure

**User-centric portal—met**: The landing page on laion.ai directs to the Hugging Face repository. Hugging Face is a platform that has been built from the ground up as a user-centric portal for AI practitioners, meaning it provides all the functionality necessary for a data portal.

**API accessibility—met**: Usage with Hugging Face is simple via the Hugging Face Datasets API. It should be noted that, with this API, metadata is not served alongside the dataset upon query: In fact, calls to different APIs are necessary for both datasets and their metadata to be brought into a developer's programming environment.

**Version control infrastructure—met**: Hugging Face has built-in versioning functionality based on git.

### Overall assessment

Re-LAION-2B meets a lot of AI-readiness requirements solely from being published on Hugging Face, a testament to the ease by which data can be made available and suitable for AI practitioners. However, on metadata, the dataset falls short, not supplying enough information for practitioners to make responsible, well-evidenced decisions about usage. As a result, the dataset is not AI-ready.

## Evaluating the AI-readiness of a dataset on the Protein Data Bank

Proteins, fundamental molecules for all known life, are composed of long chains of amino acids, their sequences determined by DNA. The interactions among these amino acids cause the macromolecule to twist, fold, and coil into distinct three-dimensional configurations, each suited to its biological function.

Understanding the relationship between protein structure and function is central to bioinformatics. This knowledge is crucial for the design and synthesis of functional proteins, which hold promise for medical therapies. The Protein Data Bank (PDB) serves as a globally accessible, open repository, archiving nearly every protein sequence and molecular structure elucidated over the past five decades. Its contributions to biomedical advancements have been significant, notably serving as training data for the Nobel Prize-winning AI model, AlphaFold.

As an example from the PDB, consider the dataset for the human insulin protein. This mmCIF dataset provides granular information, including the following: the protein's classification; its organism of origin (e.g., Homo sapiens); the expression system utilized in its experimental sequencing (e.g., E. coli); the fundamental amino acid sequences of the protein; the outcomes of the sequencing experiment and external validation metrics; the protein's mutational status relative to its original, wild-type form; and relevant associated literature references.

Following the methodology of the preceding example, we assessed the AI-readiness of this dataset by examining the dataset itself, its presentation on the PDB data portal, and its accompanying technical documentation.

### Dataset properties

**International standards and norms—met**: Strict adherence to the widely recognized mmCIF standard for macromolecular crystallographic data (PDBx/mmCIF dictionary v5.398) is observed. The dataset employs standardized biochemical conventions (e.g., three-letter amino acid codes like Ala for alanine), and UniProt accession codes are used for macromolecule identification.

**Semantic and logical consistency—met**: Labels and nomenclature consistently follow IUPAC/IUBMB biochemical standards, ensuring semantic clarity across all dataset entries. Furthermore, uniform definitions and notations are applied for biochemical modifications and structural elements.

**Identifiable class and source imbalance—partially met**: Protein sources and expression systems are explicitly documented. While class imbalance (e.g., the prevalence of proteins versus other molecular entities like water molecules) is implicitly discernible, its explicit quantification or summarization in the metadata would be beneficial.

**De-identification and anonymization—N/A**: As the dataset intrinsically lacks identifiable or sensitive personal data, concerns regarding anonymization are not applicable.

**Appropriate file format—partially met**: Data is provided in the mmCIF file format, which is optimized for molecular data and internationally accepted. However, AI-readiness could be enhanced by offering supplementary formats (e.g., providing the amino acid sequence in a CSV alongside the mmCIF dataset) to streamline integration with various AI workflows.

### Metadata

**Machine-readable format—partially met**: The mmCIF standard is structured, inherently machine-readable, and well-suited for bioinformatics applications. Nevertheless, incorporating AI-specific metadata formats (such as JSON-LD) could further improve interoperability and utility within broader AI contexts.

**Dataset served with attached metadata—met**: Data and its associated metadata are integrated within the PDB files, ensuring that metadata is delivered alongside data entries.

**Basic technical specifications—met**: The dataset documents dimensionality and provides essential summary statistics (resolution, R-factor). Moreover, refined experimental structure information (e.g., method: x-ray diffraction and resolution details) is explicitly included.

**Supply chain information—partially met**: Data collection methods and refinement processes (e.g., via software such as Phenix and DENZO) are documented. However, the metadata could be further enhanced by providing more explicit traceability for preprocessing details beyond experimental conditions. This information is typically found in referenced literature, a standard practice in the biomedical field, but less ideal for AI practitioners without specialized subject matter expertise.

**Legal and sociotechnical information—partially met**: Although the dataset's usage license is not explicitly detailed within the metadata, it is clearly specified on the official PDB website.

### Surrounding infrastructure

**User-centric portal—met**: The dataset is readily accessible through a dedicated page on a transparent and user-friendly data portal. This portal facilitates straightforward searching, downloading, visualization (including in 3D), and detailed data exploration, such as grouping structures and performing pairwise structure alignment.

**API accessibility—met**: The dataset is accessible via a robust, standard API provided by the PDB, enabling efficient programmatic access.

**Version control infrastructure—met**: The PDB website supports auditing and revisions, with updates automatically recorded under "audit revision history" on the dataset's web page and within its metadata.

### Overall assessment

In conclusion, the PDB, as exemplified by the human insulin dataset, demonstrates AI-readiness in its adherence to established international standards and its data accessibility. However, there remain opportunities for improvement, particularly in diversifying file formats, enriching metadata clarity and scope, and explicitly documenting legal information within the metadata itself. Based on this evaluation, the PDB (specifically, the human insulin example) is deemed AI-ready.

## Evaluating the AI-readiness of ERA5

In the domain of climate science and meteorology, the ERA5 dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) represents the state-of-the-art in global atmospheric reanalysis. Released operationally in 2019, ERA5 provides hourly estimates of a vast array of atmospheric, land, and oceanic climate variables from 1940 to the present day (Hersbach et al. 2020).

The dataset is generated by assimilating historical observations into a numerical weather prediction model (IFS Cycle 41r2), creating a coherent, physics-consistent record of the Earth's atmosphere on a 31-km global grid (Hersbach et al. 2020). ERA5 has recently become the foundational bedrock for the "AI revolution" in weather forecasting; it served as the primary training corpus for breakthrough models such as NVIDIA's FourCastNet (Pathak et al. 2022), Huawei's Pangu-Weather (Bi et al. 2023), and Google DeepMind's GraphCast (Lam et al. 2023).

Its pivotal role in enabling these systems makes it an ideal candidate for evaluating "strong" AI-readiness in a complex scientific domain. We assess ERA5 below using our framework.

### Dataset properties

**International standards and norms—met**: The dataset adheres strictly to World Meteorological Organization (WMO) standards. Variables are encoded using standard GRIB and NetCDF conventions, and parameters (such as geopotential height or specific humidity) map to globally recognized meteorological definitions and units (World Meteorological Organization 2019).

**Semantic and logical consistency—met**: As a reanalysis product, ERA5 is generated by a physical model that enforces internal consistency across variables (e.g., ensuring wind speed and pressure fields are physically consistent with one another). This "physics-enforced" consistency is a key reason for its success in training AI emulators (Rasp et al. 2020).

**Identifiable class and source imbalance—met**: While the observational inputs (satellite data, weather stations) vary over time and geography, the output is "regridded" to a complete, gap-free global grid. The "imbalance" of input observations is well-documented in the accompanying quality assurance reports, which detail the changing observational network from 1940 to the present (Hersbach et al. 2020).

**De-identification and anonymization—N/A**: The dataset contains physical measurements of the Earth system and does not involve personally identifiable information.

**Appropriate file format—met**: Data are natively available in GRIB and NetCDF formats, which

are standard in the domain. Crucially, to support cloud-native AI workflows, "Analysis-Ready, Cloud Optimized" (ARCO) versions of ERA5 (stored in Zarr format) have recently been made available, specifically designed to streamline ingestion into ML pipelines (Rasp et al. 2024; Stern et al. 2022).

### Metadata

**Machine-readable format—met**: The Climate Data Store (CDS) provides metadata in a structured, machine-readable format accessible via API. Furthermore, the GRIB and NetCDF files themselves contain rich self-describing headers (defining coordinate systems, units, and time steps) that software libraries like `xarray` can parse automatically (Hoyer and Hamman 2017).

**Dataset served with attached metadata—met**: Metadata regarding variable definitions, units, and temporal validity is embedded directly within the data files, ensuring that the data never become "orphaned" from its context.

**Basic technical specifications—met**: The dataset is exhaustively documented with regards to dimensionality (hourly intervals, 137 vertical levels, 31-km horizontal resolution), spatial coverage, and unit specifications (Hersbach et al. 2020).

**Supply chain information—met**: The provenance is transparent: The specific version of the generating model (IFS Cy41r2) and the data assimilation methodology (4D-Var) are public knowledge. The distinction between the consolidated "ERA5" product and the preliminary "ERA5T" (near real-time) product is explicitly managed and labeled in the metadata (ECMWF 2020).

**Legal and sociotechnical information—met**: The data are released under the Copernicus License, which functions similarly to a Creative Commons Attribution (CC-BY) license. This open licensing is clearly stated on the portal and allows for commercial use, a critical enabler for the thriving ecosystem of private AI weather startups (European Commission 2019).

### Surrounding infrastructure

**User-centric portal—met**: The CDS acts as a highly user-centric portal. It provides web-based tools for browsing, subsetting, and plotting data without downloading, alongside comprehensive documentation and a "toolbox" for server-side processing (Raoult et al. 2017).

**API accessibility—met**: The dataset is supported by the `cdsapi`, a robust Python client that allows users to programmatically request and retrieve specific slices of data. This API is the standard entry point for AI researchers and is widely integrated into open-source data loaders (Raoult et al. 2017).

**Version control infrastructure—met**: The infrastructure supports versioning and updates. When errors are discovered (such as stratospheric temperature biases in the 2000–2006 period), corrected versions (e.g., ERA5.1) are released and clearly distinguished in the catalogue, allowing AI practitioners to choose the appropriate version for training (Simmons et al. 2020).

### Overall assessment

ERA5 exemplifies a dataset that is "AI-ready by design," even if it predates the current generative AI boom. Its adherence to strict scientific standards, combined with modern infrastructure (the CDS API and ARCO formats), removes almost all friction for AI developers. The ability for researchers to write a simple script to fetch petabytes of consistently formatted, physics-checked training data have directly enabled the creation of models like GraphCast. Consequently, **ERA5 is deemed AI-ready**.

## Note: Maturity versus AI-readiness

The PDB has been an established data repository for half a century yet exceeds in our measurement of its AI-readiness, while the much more recent Re-LAION-2B, released with the explicit intention of supporting AI, does not (Table 2). Perhaps temporal maturity is the reason: The PDB's structure, contents, metadata, and infrastructure have evolved over time with the help of domain experts that use it and engage in dialogue with the repository developers, in turn solidifying it as an invaluable resource that will support innovation for decades more. The young Re-LAION-2B can offer no comparison here. However, the strong AI-readiness of ERA5, despite its comparatively recent release, suggests this is not the deciding factor. Temporal maturity should not be conflated with AI-readiness; a newly published dataset is not inevitably "weak," nor is an older, more mature dataset automatically "strong" (nor vice versa).

**TABLE 2** Summary of AI-readiness evaluations across the three case study datasets. Ratings indicate the extent to which each dataset adheres to the framework criteria.

| Criteria | Re-LAION-2B | PDB (human insulin) | ERA5 |
|---|---|---|---|
| **1. Dataset properties** | | | |
| Int. standards & norms | Met | Met | Met |
| Semantic consistency | Met | Met | Met |
| Class/Source imbalance | Met | Partially met | Met |
| De-identification | Not met | N/A | N/A |
| Appropriate file format | Met | Partially met | Met |
| **2. Metadata** | | | |
| Machine-readable format | Partially met | Partially met | Met |
| Attached metadata | Not met | Met | Met |
| Basic tech. specifications | Partially met | Met | Met |
| Supply chain info. | Partially met | Partially met | Met |
| Legal & sociotechnical | Met | Partially met | Met |
| **3. Surrounding infrastructure** | | | |
| User-centric portal | Met | Met | Met |
| API accessibility | Met | Met | Met |
| Version control infra. | Met | Met | Met |
| **Overall assessment** | **Not AI-ready** | **AI-ready** | **AI-ready** |

## PROPOSITIONS FOR THE AI-READY FUTURE OF THE OPEN DATA ECOSYSTEM

We conclude by discussing the further steps that should be taken for the entire data ecosystem to be made AI-ready in order to realize its true potential in supporting an innovative future.

First, metadata practices must be standardized to support data discoverability and interoperability. When datasets are combined, robust provenance and lineage metadata are essential for tracing data elements to their origins. For instance, this can substantially help to understand which licences apply to different parts of a dataset. What's more, machine-readable metadata can make licensing terms more immediately understandable, though the ecosystem currently struggles to represent licensing information accurately in metadata (Carey-Wilson et al. 2025). The ethics of web-scraping and adherence to terms of service also demand attention. Scraping web data for AI training has led to legal disputes, particularly when copyrighted material or data violating the terms of service is used. While robots.txt files offer a technical restriction, not all web scrapers comply (Longpre et al. 2024). The concentration of market power among large entities (Gans 2024) is also leading to closed-door data-sharing agreements with opaque terms (Bestall 2024). This trend limits open access and requires countermeasures.[2]

Second, for this reason, collaborative governance and community engagement in the ecosystem around AI datasets must be furthered (Massey et al. 2024). Critical AI datasets require clear governance, not an absence of it, with institutional norms currently shaping its content, structure, and access (Fumega 2024). Data repositories and platforms play a key role in guiding users towards responsible practices, such as providing comprehensive metadata (Costa, Walker, and Simperl 2020). This responsibility extends from data publishers to platform providers and data consumers. For this reason, feedback loops between data publishers and AI practitioners must be strengthened. Infrastructure should support these mechanisms to enable iterative dataset improvements based on the AI community's evolving needs. Finally, the ecosystem must promote equitable access and mitigate data monopolies as noted earlier. The move towards private data-sharing agreements, driven by legal concerns and market power, could stifle innovation by restricting access to essential datasets. Active promotion of open and nondiscriminatory data access (OECD 2025) is necessary, aligning with initiatives like the EU Data Act.

Third, data infrastructure must mature to handle AI's scalability and usability requirements. The demand for large, diverse, and high-quality datasets (Samborska 2025) has increased the importance of open data portals, repositories, and APIs. These platforms, including government open data sites, data science platforms like Kaggle, and model/data hubs like Hugging Face, are central to the AI data ecosystem (Open Data Institute 2024a). To better serve AI, this infrastructure must evolve. Platforms need to expand their support for diverse data types, such as

video and audio, and AI-optimized formats like Apache Parquet. Parquet supports efficient columnar storage and complex data relationships, which are critical for AI model training. Such advancements complement existing efforts by scientific data portals (as well as governmental open data initiatives) to provide data in semantic formats like RDF. Discoverability and semantic richness also need improvement; AI systems require rich, standardized metadata to automatically find, understand, and integrate data.

Addressing these points requires cooperation between data publishers, research institutions, businesses, standards bodies, and policymakers. The goal is to transform the entire open data ecosystem so that data are AI-ready by design, promoting responsible innovation.

## CONCLUSION

As AI continues to advance, the importance of foundational data only grows. The case studies demonstrate that making data AI-ready can dramatically increase its value and impact. Our framework for AI-ready data provides a roadmap for data publishers to elevate their datasets across three key fronts: the quality of the data, the richness of its metadata, and the support of the infrastructure surrounding it. Data publishers can ensure that their data are usable and trustworthy for AI applications by taking a holistic approach that addresses all these components.

One of the clear lessons is that AI-readiness does not happen by accident or just via a one-time effort. It is a deliberate process and, often, an ongoing commitment. It means thinking about the end-user (or end-model) at every step of the data lifecycle: from adopting standards at collection time, to documenting context when publishing, to maintaining channels for access and feedback after release. It also means breaking down silos—technical teams, documentation teams, and platform teams need to work in concert. Encouragingly, many of the practices we have discussed are also simply data best practices. They lead to better data management and higher data quality in general, not only for AI scenarios. In fact, by making a dataset AI-ready, publishers inevitably make it more useful for all kinds of analysis, not just ML.

It is also important to acknowledge that AI-readiness is a spectrum, not a binary state. A few datasets will be perfect for every criterion, and that is okay. The goal is to push towards greater readiness. Perhaps a team begins by enhancing metadata and adding a license file this quarter; next quarter, they focus on setting up an API, and so on. Even modest improvements, such as providing a data schema or an example of a Jupyter notebook with the dataset, can lower the barrier for AI practitioners. Over time, these enhancements compound. Moreover, data publishers can gain valuable insights into what to prioritize by engaging with the AI community and enabling a continuous dialogue that can inform their iteration and updates of their datasets.

In conclusion, AI-ready data practices enable datasets to reach their full potential. In an era where AI is hungry for high-quality data, providing such data can catalyze new research discoveries, power innovative applications, and inform better decision-making. Organizations like the ODI are helping translate lofty principles into actionable checklists and guidance that anyone can apply. The frameworks will undoubtedly evolve. For instance, specific domains, such as healthcare or climate data, may impose additional requirements, including ethical considerations or domain-specific metadata standards. However, the core idea remains: If we want robust, fair, and effective AI, we must begin with strong, well-prepared, and well-documented data.

## ONGOING WORK

In December 2025, the ODI produced a v2 of the Framework for AI-ready data (Massey, Simperl, and Majithia 2025), introducing "governance" as a fourth category, informed by feedback from users. Criteria within this category include "governance policy-as-code," "documented roles and responsibilities," "publicly identifiable points of contact," and "clear data access processes."

The framework was adapted and applied for research on the AI-readiness of data held by local governments in the United Kingdom (Simperl et al. 2025).

### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

### ORCID

*Neil Majithia* https://orcid.org/0009-0008-3969-2514

### ENDNOTES

[1] Accessibility criteria here are distinctly different from those present in the FAIR principles.

[2] We note that an agentic AI version of the worldwide web will expound certain aspects of dataset discoverability. Agents autonomously searching for datasets, downloading them, and using them for analysis will require strong infrastructural standards regarding metadata to ensure accuracy, trustworthiness, and responsibility.

### REFERENCES

Afzal, Shazia, Rajmohan C, Manish Kesarwani, Sameep Mehta, and Hima Patel. 2020. "Data Readiness Report." https://doi.org/10.48550/arXiv.2010.07213.

Akhtar, Mubashara, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, et al. 2024. "Croissant: A Metadata Format for ML-Ready Datasets." In Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning, SIGMOD/PODS'24, 1–6. ACM. https://doi.org/10.1145/3650203.3663326.

Anthropic. 2024. "Introducing the Model Context Protocol." https://www.anthropic.com/news/model-context-protocol.

Benjelloun, Omar, Elena Simperl, Pierre Marcenac, Pierre Ruyssen, Costanza Conforti, Michael Kuchnik, Jos van der Velde, et al. 2024. "Croissant Format Specification 1.0." MLCommons. https://docs.mlcommons.org/croissant/docs/croissant-spec-1.0.html.

Benjelloun, Omar, Elena Simperl, Pierre Marcenac, Pierre Ruyssen, Costanza Conforti, Michael Kuchnik, Jos van der Velde, et al. 2025. "Croissant Format Specification 1.1." MLCommons. https://docs.mlcommons.org/croissant/docs/croissant-spec-1.1.html.

Berners-Lee, Tim, Roy T. Fielding, and Larry M. Masinter. 2005. "Uniform Resource Identifier (URI): Generic Syntax." Request for Comments RFC 3986, Internet Engineering Task Force. https://datatracker.ietf.org/doc/rfc3986/.

Bestall, Alex. 2024. "With a Data Licensing Framework in Play, Rights Holders Can Embrace AI." https://variety.com/vip/data-licensing-framework-rights-holders-can-embrace-ai-1236035265/.

Bi, Kaifeng, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. "Accurate Medium-Range Global Weather Forecasting With 3D Neural Networks." *Nature* 619(7970): 533–38. https://doi.org/10.1038/s41586-023-06185-3.

Birhane, Abeba, Sanghyun Han, Vishnu Boddeti, and Sasha Luccioni. 2023. "Into the Laion's den: Investigating Hate in Multimodal Datasets." *Advances in neural information processing systems* 36: 21268–84. https://dl.acm.org/doi/10.5555/3666122.3667052.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." Preprint, available on arXiv as arXiv:2110.01963. https://doi.org/10.48550/arXiv.2110.01963.

Carey-Wilson, Thomas, Gefion Thuermer, Elena Simperl, and Lee Tiedrich. 2025. "Unlocking Data Collaboration: A Study on Data Sharing Practices and Developing Standard Data Licence Terms to Promote Access and Social Good." https://doi.org/10.61557/XPWM7241/.

Casilli, Antonio A., Paola Tubaro, Maxime Cornet, Clément Le Ludec, Juana Torres-Cierpe, and Matheus Viana Braz. 2025. "Global Inequalities in the Production of Artificial Intelligence: A Four-Country Study on Data Work." In The Handbook of Digital Labor (eds J.L. Qiu, S. Yeo and R. Maxwell). https://doi.org/10.1002/9781119981831.ch12.

Castelijns, Laurens A., Yuri Maas, and Joaquin Vanschoren. 2020. "The ABC of Data: A Classifying Framework for Data Readiness." In Machine Learning and Knowledge Discovery in Databases, edited by Peggy Cellier and Kurt Driessens, 3–16. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-43823-4_1.

Clark, Timothy, Harry Caufield, Jillian A. Parker, Sadnan Al Manir, Edilberto Amorim, James Eddy, Nayoon Gim, et al. 2024. "AI-Readiness for Biomedical Data: Bridge2AI Recommendations." *bioRxiv*. https://doi.org/10.1101/2024.10.23.619844.

Costa, Eric, Johanna Walker, and Elena Simperl. 2020. "Sustainability of (Open) Data Portal Infrastructures: Open Data Portal Assessment Using User-Oriented Metrics." European Data Portal. https://data.europa.eu/sites/default/files/sustainability-data-portal-infrastructure_5_open-data-portal-assessment.pdf.

Curzon, James, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib. 2021. "Privacy and Artificial Intelligence." *IEEE Transactions on Artificial Intelligence* 2(2): 96–108. https://doi.org/10.1109/TAI.2021.3088084.

ECMWF. 2020. "ERA5: Data Documentation." European Centre for Medium-Range Weather Forecasts. https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation.

European Commission. 2019. "Licence to Use Copernicus Products." https://cds.climate.copernicus.eu/api/v2/terms/static/licence-to-use-copernicus-products.pdf (accessed June 6, 2025).

European Commission. 2025. "Data Act Explained | Shaping Europe's Digital Future." https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained.

Fielding, Roy Thomas. 2000. "Architectural Styles and the Design of Network-Based Software Architectures." https://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm.

Fumega, Silvana. 2024. "The Global Landscape of Data Governance." https://www.cigionline.org/articles/the-global-landscape-of-data-governance/.

Gans, Joshua S. 2024. "Market Power in Artificial Intelligence." Working Paper 32270, National Bureau of Economic Research. https://doi.org/10.3386/w32270.

GOV.UK. 2023. "Using Metadata to Describe CSV Data." https://www.gov.uk/government/publications/recommended-open-standards-for-government/using-metadata-to-describe-csv-data.

Hersbach, Hans, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, et al. 2020. "The ERA5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146(730): 1999–2049. https://doi.org/10.1002/qj.3803.

Hiniduma, Kaveen, Suren Byna, Jean Luca Bez, and Ravi Madduri. 2025. "AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI." In Proceedings of the 36th International Conference on Scientific and Statistical Database Management, pp. 1–12. 2024. https://doi.org/10.1145/3676288.3676296.

Hoyer, Stephan, and Joseph Hamman. 2017. "xarray: N-D labeled arrays and datasets in Python." *Journal of Open Research Software* 5(1): 10. https://doi.org/10.5334/jors.148.

Hugging Face. 2025. "laion/relaion2B-multi-research." Dataset. https://huggingface.co/datasets/laion/relaion2B-multi-research.

Jain, Nitisha, Mubashara Akhtar, Joan Giner-Miguelez, Rajat Shinde, Joaquin Vanschoren, Steffen Vogler, Sujata Goswami, et al. 2024. "A Standardized Machine-Readable Dataset Documentation Format for Responsible AI." https://doi.org/10.48550/arXiv.2407.16883.

Koesten, Laura, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. 2020. "Dataset reuse: Toward translating principles to practice." *Patterns* 1(8): 100136.

LAION. 2022. "LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets." https://laion.ai/blog/laion-5b/ .

LAION. 2024. "Releasing Re-LAION 5B: Transparent Iteration on LAION-5B With Additional Safety Fixes." https://laion.ai/blog/relaion-5b.

Lam, Remi, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Zach Eaton-Rosen, et al. 2023. "Learning Skillful Medium-Range Global Weather Forecasting." *Science* 382(6677): 1416–21. https://doi.org/10.1126/science.adi2336.

Longpre, Shayne, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, et al. 2024. "Consent in Crisis: The Rapid Decline of the AI Data Commons." *Advances in Neural Information Processing Systems* 37: 108042–87, https://doi.org/10.52202/079017-3431.

Massey, Joe, Elena Simperl, and Neil Majithia. 2025. "Bringing Governance Into the ODI Framework for AI-Ready Data." The Open Data Institute. https://theodi.org/news-and-events/blog/bringing-governance-into-the-odi-framework-for-ai-ready-data/.

Massey, Joe, Poorvi Yerrapureddy, Vinay Narayan, Elena Simperl, Astha Kapoor, and Matthew Barber. 2024. "What Makes Participatory Data Initiatives Successful?" https://doi.org/10.61557/RPDN8062.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54(6): 1–35. https://doi.org/10.1145/3457607.

Niven, Timothy, and Hung-Yu Kao. 2019. "Probing Neural Network Comprehension of Natural Language Arguments." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 4658–64. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1459.

OECD. 2025. "Enhancing Access to and Sharing of Data in the Age of Artificial Intelligence." OECD.

Open Data Institute. 2024a. "Policy Intervention 4: Ensuring Broad Access to Data for Training AI Models." https://theodi.org/news-and-events/blog/policy-intervention-4-ensuring-broad-access-to-data-for-training-ai-models/.

Open Data Institute. 2024b. "What Are Data Spaces and What Do They Do?" https://theodi.org/news-and-events/blog/what-are-data-spaces-and-what-do-they-do/.

Pathak, Jaideep, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, et al. 2022. "FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators." *Preprint, available on arXiv as arXiv:2202.11214*. https://doi.org/10.48550/arXiv.2202.11214.

PleIAs and Hugging Face. 2024. "PleIAs/common_corpus." Dataset. https://huggingface.co/datasets/PleIAs/common_corpus.

RAIL Initiative. 2025. "Responsible AI Licenses (RAIL)." https://www.licenses.ai.

Raoult, Baudouin, Cedrick Bergeron, Angel Alivia, Pierre-Antoine Draymond, and Eoin Gleeson. 2017. "Climate Service Develops User-Friendly Data Store." *ECMWF Newsletter* 151: 22–27. https://www.ecmwf.int/en/elibrary/80024-newsletter-no-151-spring-2017.

Rasp, Stephan, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. "WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting." *Journal of Advances in Modeling Earth Systems* 12(11): e2020MS002203. https://doi.org/10.1029/2020MS002203.

Rasp, Stephan, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Jason Vigh, et al. 2024. "WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models." *Preprint, available on arXiv as arXiv:2308.15560*. https://doi.org/10.48550/arXiv.2308.15560.

RDA Artificial Intelligence and Data Visitation Working Group. 2025. "Global Adaptation of the AI Bill of Rights for Research." https://www.rd-alliance.org/groups/artificial-intelligence-and-data-visitation-aidv-wg/activity/.

RDA Data Discovery Paradigms Interest Group. 2024. "Ten Principles to Improve Dataset Discoverability." Supporting Output. https://doi.org/10.15497/RDA00101.

Sajid, Haziqa. 2024. "Best Practices for Data Versioning for Building Successful ML Models." https://encord.com/blog/data-versioning/.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. ""Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI." In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15. Yokohama, Japan: ACM. https://doi.org/10.1145/3411764.3445518.

Samborska, Veronika. 2025. "Scaling Up: How Increasing Inputs has Made Artificial Intelligence More Capable." Our World in Data. https://ourworldindata.org/scaling-up-ai.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, et al. 2022. "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models." *Advances in neural information processing systems* 35: 25278-94. https://dl.acm.org/doi/10.5555/3600270.3602103.

Schwabe, Daniel, Katinka Becker, Martin Seyferth, Andreas Klaß, and Tobias Schaeffter. 2024. "The METRIC-Framework for Assessing Data Quality for Trustworthy AI in Medicine: A Systematic Review." *npj Digital Medicine* 7(1): 1–30. https://doi.org/10.1038/s41746-024-01196-4.

Simmons, Adrian, Cornel Soci, Julien Nicolas, Bill Bell, Paul Berrisford, Rossana Dragani, Johannes Flemming, et al. 2020. "Global stratospheric temperature bias and other stratospheric aspects of ERA5 and ERA5.1." 859, ECMWF. https://doi.org/10.21957/rcx12w3u.

Simperl, Elena, Neil Majithia, Thomas Carey-Wilson, Anabell Kuldmaa, and Priit Liivak. 2025. "Insights from UK Councils on Standards, Readiness and Reform to Modernise Public Data for AI." Report, The Open Data Institute. https://theodi.org/insights/reports/insights-from-uk-councils-on-standards-readiness-and-reform-to-modernise-public-data-for-ai/.

Stern, Charles, Ryan Abernathey, Joseph Hamman, Rachel Wegener, Chiara Lepore, Sean Harkins, and Alexander Merose. 2022. "Pangeo Forge: Crowdsourcing Analysis-Ready, Cloud Optimized Data Production." *Frontiers in Climate* 3: 782909. https://doi.org/10.3389/fclim.2021.782909.

U.S. Department of Commerce. 2025. "Generative Artificial Intelligence and Open Data: Guidelines and Best Practices." https://www.commerce.gov/news/blog/2025/01/generative-artificial-intelligence-and-open-data-guidelines-and-best-practices.

Verhulst, Stefaan, Andrew Zahuranec, and Hannah Chafetz. 2025. "Moving Toward the FAIR-R Principles: Advancing AI-Ready Data." Preprint, available at SSRN. https://doi.org/10.2139/ssrn.5164337.

Vohra, Deepak. 2016. *Apache Parquet*, 325–35. Berkeley, CA: Apress.

Wiggers, Kyle. 2022. "A Startup Wants to Democratize the Tech Behind DALL-E 2, Consequences Be Damned." https://techcrunch.com/2022/08/12/a-startup-wants-to-democratize-the-tech-behind-dall-e-2-consequences-be-damned/.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3(1): 160018. https://doi.org/10.1038/sdata.2016.18.

World Meteorological Organization. 2019. *Manual on Codes: International Codes*, vol. I.2. Geneva: World Meteorological Organization. https://community.wmo.int/about-manual-codes-volume-i2.

Wu, Mingfang, Fotis Psomopoulos, Siri Jodha Khalsa, and Anita de Waard. 2019. "Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories." *Data Science Journal* 18: 3. https://doi.org/10.5334/dsj-2019-003.

**How to cite this article:** Majithia, N., T. Carey-Wilson, E. Simperl, and N. Shadbolt 2026. "An actionable framework for AI-ready data." *AI Magazine* e70054. https://doi.org/10.1002/aaai.70054

## AUTHOR BIOGRAPHIES

**Neil Majithia** is a researcher at the Open Data Institute (ODI). With a background in Economics, he has contributed to a range of research programmes at the ODI although mainly focused on Data-centric AI, exploring the interactions between government and AI in the context of data.

**Thomas Carey-Wilson** is a researcher at the Open Data Institute (ODI). He explores data governance practices throughout the AI data lifecycle.

**Elena Simperl** is a professor of Computer Science at King's College London, where she co-directs the King's Institute for Artificial Intelligence. She is also the director of Research at the Open Data Institute, a fellow of the British Computer Society and the Royal Society of Arts, and a Hans Fischer Senior Fellow. Elena's work is at the intersection between AI and social computing. She features in the top 100 most influential scholars in knowledge engineering of the last decade and the Women in AI 2000 ranking. Elena co-chairs the Croissant working group in ML Commons, developing an open standard to improve data portability, discovery and use in AI. She is the president of the Semantic Web Science Association.

**Nigel Shadbolt** began his academic career in 1978 with a double first in Psychology and Philosophy at the University of Newcastle. His interests in logic, computation, and cognition led him to the Department of Artificial Intelligence at the University of Edinburgh where he obtained his PhD. Since then, he has filled a range of roles in both the public and private sectors, including as the Allan Standen Professor of Intelligent Systems at the University of Nottingham, professor of Artificial Intelligence at the University of Southampton, information advisor to the UK Government, and as chairman of the Open Data Institute, which he co-founded with Sir Tim Berners-Lee. He is a fellow of the Royal Society, the Royal Academy of Engineering, and the British Computer Society. He became the Principal of Jesus College in 2015 and is a professor in the Department of Computer Science where he leads the Human Centered Computing Group. At Oxford, he has focused his research on human centered AI in a wide range of applications. Most recently, he was asked to lead the setting up of the Oxford Institute of Ethics in AI.