A framework for AI-ready data

ODI Research
ADVANCING TRUST IN DATA

May 2025

Contents

About

This report, authored by Neil Majithia, Thomas Carey-Wilson and Elena Simperl, was researched and produced by the Open Data Institute (ODI) and published in May 2025. If you would like to share feedback or get in touch, contact the research team at research@theodi.org.

We would like to thank all participants in this research, including interviewees and others who provided valuable input.

Introduction

Data is the foundation of AI. Poor quality data drives up costs and can lead to hidden problems for AI models, especially in complex fields like healthcare and policy,[1] while biased data negatively affects the performance of AI models[2] and uninspected evaluation datasets can lead to false positives or overestimates of model accuracy.[3] For data publishers to realise 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.

Across the research in the ODI's data-centric AI programme, we have come to define the AI-readiness of a dataset over four components: its **technical optimisation for machine learning**; its **overall quality and adherence to standards**; its **legal compliance**; and its **responsible collection**. These components encompass a dataset's ethical and regulatory compliance as well as its ease of use for both technical practitioners (for example, AI engineers and researchers) and supporting actors in the ecosystem (for example, responsible AI (RAI) teams and governance and regulatory bodies).

[1] Sambasivan et al. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. doi: 10.1145/3411764.3445518
[2] Mehrabi et al. (2019). A Survey on Bias and Fairness in Machine Learning. doi: 10.48550/arXiv.1908.09635
[3] Niven & Kao. (2019). Probing Neural Network Comprehension of Natural Language Arguments. doi: 10.18653/v1/P19-1459

The concept of AI-readiness has grown in prominence but generally remains too vaguely defined and complex to implement practically. Existing frameworks, such as AIDRIN,[4] Bridge2AI[5] or FAIR-R[6], provide high-level recommendations across the four components of AI-readiness but often lack the actionable specificity required by data publishers, who are rarely AI specialists themselves. In contrast, initiatives in the field of RAI[7] provide specific recommendations for the latter two components of AI-readiness, but lack the holistic view that a data publisher needs when taking steps towards AI-ready data practices.

This report proposes an AI-readiness framework for data that addresses these limitations. By engaging AI practitioners and domain experts and combining their insights with the ODI's expertise, we outline the concrete steps that data publishers and repository providers can take to achieve genuine, holistic AI-readiness. Our framework promotes a 'by design' approach to AI-readiness, targeting the practical aspects of a dataset's collection, preparation and publication to enhance its quality and utility for the AI ecosystem. Significantly, our approach does not extend to evaluating organisational or institutional AI-readiness; instead, it focuses on data, metadata, and the infrastructure surrounding a published dataset and how each component can be made AI-ready. The methodology for our framework design is detailed in Appendix I.

[4] Hiniduma et al. (2024). AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI. doi: 10.1145/3676288.3676296
[5] Clark et al. (2024). AI-readiness for Biomedical Data:Bridge2AI Recommendations. doi: 10.1101/2024.10.23.619844
[6] Verhulst, Zahuranec & Chafetz. (2025). Moving Toward the FAIR-R principles: Advancing AI-Ready Data. doi: 10.2139/ssrn.5164337
[7] Pushkarna, Zaldivar & Kjartansson. (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. doi: 10.1145/3531146.3533231

Beyond FAIR and other data readiness frameworks

The FAIR principles framework[8] is the most renowned example of guidance for data practices. The principles refer to data, metadata, and surrounding infrastructure[9] and advocate that, for the global digital ecosystem to fully enable data-led research and enterprise solutions, datasets should be Findable, Accessible, Interoperable and Reusable (FAIR). However, despite FAIR's broad adoption, our discussions with AI practitioners revealed a consensus that the principles lack the technical specificity needed for practical AI-readiness guidance. FAIR Implementation Profiles (FIPs)[10] do seek to add that detail to FAIR guidance, although published FIPs are generally constrained to life sciences and rarely contain recommendations for dataset AI-readiness[11].

Recent adaptations of the FAIR principles have started to explore these limitations. Verhulst, Zahuranec & Chafetz (2025) proposed expanding FAIR with a dimension focused on AI-readiness (FAIR-R), with the authors highlighting the importance of structured and accurately-labelled datasets[12] for AI purposes. FAIR-R also entails provenance, transparency, bias mitigation, detailed annotation, and governance practices for AI datasets. Still, these criteria remain conceptual in the paper, offering high-level discussion topics without concrete operational guidance for data publishers. Similarly, Huerta et al. (2022)[13] co-designed AI-specific adaptations for the FAIR principles based on conversations between international FAIR initiatives; again, however, these lack granular recommendations that data publishers could use.

In their fundamental analysis of a range of data quality frameworks, Priestley, O'Donnell & Simperl (2023)[14] conclude that, while traditional assessments can help identify data quality issues, they frequently stop short of providing concrete operational guidance. Other AI-readiness frameworks also face these limitations: for example, Hiniduma et al. (2024)[15] combine traditional data quality metrics

[8] Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. doi: 10.1038/sdata.2016.18
[9] GO FAIR. n.d. FAIR Principles.
[10] GO FAIR, n.d. FAIR Implementation Profile.
[11] FIPs can be explored with the FAIR Connect dashboard.
[12] Verhulst, Zahuranec & Chafetz. (2025). Moving Toward the FAIR-R principles: Advancing AI-Ready Data. doi: 10.2139/ssrn.5164337
[13] Huerta et al. (2023). FAIR for AI: An interdisciplinary and international community building perspective. doi: 10.1038/s41597-023-02298-6
[14] Priestley, O'Donnell & Simperl. (2023). A Survey of Data Quality Requirements That Matter in ML Development Pipelines. doi: 10.1145/3592616
[15] Hiniduma et al. (2024). AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI. doi: 10.1145/3676288.3676296

(such as completeness or duplication) with AI-specific ones (such as fairness), but remain an evaluative tool rather than actionable for dataset design or publication.

Similarly, the Afzal et al. (2020)[16] work mainly offers conceptual guidance, outlining data characteristics and quality transformations without specific instructions, as does the Castelijns, Maas & Vanschoren (2020) ABC framework[17].

Dataset AI-readiness frameworks are often designed specifically for single domains and subject matters. For example, the Bridge2AI task force[18] offers actionable, practical guidelines for biomedical datasets[19] to be made AI-ready, with helpful advice for data publishers, albeit focusing on the environment around a dataset rather than its specific properties. Schwabe et al. (2024)[20] are also in the field of medical data. However, their METRIC framework is diametrically opposite to Bridge2AI, centring on intrinsic data quality metrics found across medical literature rather than taking a holistic view of data practices. Finally, a report by the U.S. Department of Commerce[21] provides specific action items for making open government data holistically AI-ready. Though these recommendations are strong and well evidenced, they are limited to their context in government rather than their application to enterprise or scientific data.

Academic and civil society organisations have written the above frameworks, but various enterprise AI-readiness frameworks also exist. Generally, these are concerned with organisational preparedness for using AI, but most contain a 'data governance' or 'data management' component that is relevant for our purposes. For example, Accenture's AI Maturity Explorer[22] and Deloitte's AI-readiness & Management framework[23] ask operational questions about the use of data, including whether a 'data and AI' culture has been embedded in their organisation and how well data science and ML teams work together. Deloitte's framework does mention that high-quality, labelled data is essential, but crucial details relating to the role of metadata, RAI and infrastructure are missing.

---

[16] Afzal et al. (2020). Data Readiness Report. doi: 10.48550/arXiv.2010.07213
[17] Castelijns, Maas & Vanschoren. (2020). The ABC of Data: A Classifying Framework for Data Readiness. doi: 10.1007/978-3-030-43823-4_1
[18] A programme set up by the National Institutes of Health in the U.S. Department of Health and Human Services.
[19] Clark et al. (2024). AI-readiness for Biomedical Data:Bridge2AI Recommendations. doi: 10.1101/2024.10.23.619844
[20] Schwabe et al. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. doi: 10.1038/s41746-024-01196-4
[21] AI and Open Government Data Assets Working Group. (2025). Generative Artificial Intelligence and Open Data: Guidelines and Best Practices.
[22] Accenture. n.d. What is AI Maturity.
[23] Deloitte. (2024). AI Readiness & Management Framework (aiRMF).

We should note that enterprise data maturity frameworks exist (such as JISC's Data Maturity Framework[24], DAMA's DMBOK Framework[25] and Gartner's Data Governance Maturity Model[26]) and could be reference points for designing and evaluating dataset AI-readiness. The frameworks help measure core data maturity and quality (featuring essential criteria like metadata management and data storage). However, they lack a focus on AI, and their consideration of issues like data quality is limited to how it can support business needs (such as 'defining data structures and relationships to support business processes and objectives').

Our work builds on the frameworks discussed in this section, addressing their limitations and ensuring that our recommendations for dataset AI-readiness are granular but take a holistic perspective, providing actionable guidance for data publishers and repository owners.

Our framework for AI-ready data: a snapshot

The following page contains a visualisation of our framework, with more detail in the next section.

[24] JISC. n.d. Data maturity framework.
[25] Atlan. (2024). DAMA-DMBOK Framework: All You Need To Know in 2025.
[26] Atlan. (2024). Gartner Data Governance Maturity Model: What It Is, How It Works.

Category	Criteria	Sub-criteria	Guidance and Examples
1) Dataset Properties	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.

Our framework for AI-ready data: in-depth

This section breaks down the specific requirements that define our AI-readiness framework. These requirements detail the dataset properties, metadata and surrounding infrastructure necessary to support AI practitioners' usage. Contextual explanations and references for further exploration accompany each criterion.

Preconditions

Before making their data AI-ready, a publisher should ensure that their datasets meet some general preconditions[27] [28]. Data quality (such as **cleanliness**, **completeness** and **validity**[29]) is paramount and requires thorough cleaning to eliminate anomalies, errors, duplicates and missing entries to avoid errata leading to biased or imprecise analyses[30]. Equally important is providing **metadata** (information that clarifies a dataset's origins, type, creator, intended uses, and more) to enable better and more efficient decision-making[31]. Finally, datasets should be **shared as openly as possible** (being mindful of potential legal or privacy constraints) to promote equitable access and drive innovation in the AI ecosystem[32]. Each of these preconditions is a step towards FAIRness, but also provides the foundation of our AI-readiness requirements.

[27] House of Commons Committee of Public Accounts. (2025). Use of AI in Government.
[28] Nagle, Redman & Sammon. (2017). Only 3% of Companies' Data Meets Basic Quality Standards.
[29] Government Data Quality Hub. (2021). Meet the data quality dimensions.
[30] Castelijns, Maas & Vanschoren. (2020). The ABC of Data: A Classifying Framework for Data Readiness. doi: 10.1007/978-3-030-43823-4_1
[31] Atlan. (2024). What is Metadata? Examples, Benefits & Use Cases in 2025.
[32] Open Data Institute. (2023). Data-centric AI.

1) Dataset properties

Several inherent attributes determine a dataset's suitability for AI applications.

a) Data values should follow established standards and conventions – Adhering to recognised standards and conventions (for example, countries' ISO3 codes for geographical datasets, and the ISO 8601 format for timing information) is essential for datasets to be consistent, interoperable, and high-quality for AI applications.

b) Labels should be semantically and logically consistent – To avoid ambiguity, a dataset publisher should ensure consistent and standardised data labels. Without labelling consistency, an example medical dataset in which some adverse electrocardiogram measurements are labelled 'cardiac arrest' and others labelled 'heart attack' would create ambiguity, leading an AI model to learn fictional distinctions between the two synonymous terms, or suffer from endogeneity. Annotators should therefore follow a set protocol, with labels validated by others or with the help of labelling toolkits[33].

Importantly, labelling consistency provides more than error and bias reduction. The 'linked data' protocol ensures that datasets are made functional, interoperable and shareable, with the core underlying principle that data labels conform to persistent uniform resource identifiers (URIs)[34]. URIs allow labels to follow standard ontologies and vocabularies while facilitating data representation via the Resource Description Framework (RDF)[35], enabling relationships between data labels to be easily expressed but semantically light. URIs and RDF are regarded as the fourth criterion of the five-star schema that can be used to grade the quality of open data[36].

c) Class and source imbalance should be easy to identify – A class imbalance in training data can severely skew AI model performance. Aggregated datasets that contain information from different sources similarly cause bias if one source is represented more than the others. While perfect class and source balance are aspirational, datasets should at least make imbalances easier to identify; for example, an aggregate dataset should have a column that provides provenance information for each data entry[37].

d) Standard de-identification and anonymisation methods should be used where necessary. Sensitive, personal information about individuals cannot be

[33] Examples include the Giglou et al. (2025). OntoAligner Python toolkit and the COMA 3.0 system.
[34] Network Working Group. (2005). Uniform Resource Identifier (URI): Generic Syntax.
[35] RDF Working Group. (2014). Resource Description Framework (RDF).
[36] Ontotext. n.d. What Is Five-Star Linked Open Data?
[37] For example, the CommonCorpus dataset published by Pleias.

contained in shared or public datasets without appropriate steps to protect subjects' privacy and safety. Importantly, this should consider the greater 're-identification risk' where data is used to train AI[38].

e) Datasets should be saved in appropriate file formats – Some file formats for structured data are more useful for AI contexts than others. Excel spreadsheets (.xlsx) are considered the worst due to their size, binary format, and reliance on proprietary software. In comparison, comma-separated value (.csv) files are widely used because of their flexibility. Enhancing CSV files with metadata using the W3C's CSV on the Web (CSVW) standard improves their interoperability and machine-readability[39]. CSVW allows for the inclusion of metadata that describes the structure and semantics of the data, facilitating better integration and reuse across different systems. However, the interviews confirmed that (the currently widely used) Apache Parquet files[40] are arguably the most AI-ready file formats due to several unique properties, including their 'columnar storage', metadata attachments, multimodal capabilities and innate compression. Parquet files are being adopted by the two largest AI data repositories, Hugging Face and Kaggle. For more relational datasets, such as knowledge graphs, the previously mentioned RDF file format is also recommended[41].

2) Metadata

Metadata attached to a dataset can provide users with all kinds of contextual information. Specific subsets of this information are critical for AI contexts to enable accurate, responsible use of the dataset. Metadata is saved in text-based files according to a chosen vocabulary or ontology such as the Croissant metadata standard[42].

a) Metadata should be machine-readable – Metadata should be stored and accessible so that programmatic workflows and automated systems can load and parse it effectively. In practice, this entails machine-readability, requiring metadata to be written into a format such as JSON-LD (via Croissant[43]). This enhances the metadata's communicability, interoperability and discoverability online with tools such as Google Dataset Search[44].

---

[38] Curzon et al. (2021). Privacy and Artificial Intelligence. doi: 10.1109/TAI.2021.3088084
[39] Government Digital Service. (2023). Using metadata to describe CSV data.
[40] Apache. n.d. Apache Parquet
[41] RDF Working Group. (2014). Resource Description Framework (RDF).
[42] Majithia, Carey-Wilson & Simperl. (2024). Transforming AI data governance with Croissant: a new standard for ML metadata.
[43] ibid.
[44] Google Dataset Search.

b) Metadata should be attached to a dataset – A dataset should be provided to a user with its corresponding metadata attached. This requirement is often overlooked; for instance, datasets accessed through the Hugging Face datasets API[45] are devoid of attached metadata, demanding a separate query method for users to get contextual information about a dataset. This example of metadata separation, compared to the way metadata is strictly attached to GEOTIFF files[46], presents an inconvenience at best and, at worst, discourages metadata usage and fosters bad AI data practices.

c) Metadata should include basic technical specifications – Interviewees mentioned that reporting on a dataset's modalities, dimensionality, semantics, bias and basic summary statistics made their exploratory analysis more efficient when choosing whether to use it for their AI work. Each attribute can be described with the Croissant vocabulary in its current v1.1 or upcoming v2 release. Furthermore, if any dataset component has been synthetically generated or machine-annotated, it must be reported in the metadata. Interviewees made clear that when using a dataset, synthetic components, and the assumptions used to create them, need to be expertly evaluated, meaning synthetic generation must be reported in metadata. Many global regulations, such as Canada's Artificial Intelligence and Data Act (AIDA)[47] and the EU AI Act,[48] concur.

d) Metadata should provide a holistic portrait of the entire dataset lifecycle and operational context – To operationalise Responsible AI (RAI) principles - as outlined in, for instance the RAI Croissant extension[49]- metadata must capture not only the methods of collection and preprocessing (including the roles of any outsourced entities and their competent authorities) but also a broader range of use cases. For instance, capturing data worker demographics and other social impact information, as applicable. Some indication of inequalities exists in the target populations involved in business process outsourcing (BPO) of dataset collection and labelling tasks[50]. Given this, including data worker demographic information, such as country of origin, average income bracket and/or legal entities involved in the supply chain, could help researchers accumulate more granular evidence when spotlighting these issues. This could also help practitioners make more informed decisions when selecting datasets. This more detailed metadata enables users to discern

[45] Hugging Face Datasets.
[46] Geospatial World. (2009). GeoTIFF – A standard image file format for GIS applications.
[47] Muhammad & Yow. (2023). Demystifying Canada's Artificial Intelligence and Data Act (AIDA): The good, the bad and the unclear elements. doi: 10.1109/CCECE58730.2023.10288878
[48] EU AI Act.
[49] Akhtar et al. (2024). Croissant RAI Specification.
[50] Casilli et al. (2024). Global Inequalities in the Production of Artificial Intelligence: A Four-Country Study on Data Work. doi: 10.48550/arXiv.2410.14230

potential biases and ensure that compliance requirements are integrated throughout the AI data lifecycle. Therefore, this metadata can support dynamic audit and inspection processes, supporting more effective and responsible data governance.

e) **Metadata should include other legal and sociotechnical information** – Publishers should provide the name of a dataset's licence and a URL link to its permissions and restrictions within the metadata. Practitioners should ensure that any licensing metadata clearly defines and distinguishes between key terms (for example, delineating 'model outputs' vs 'model results' and specifying what 'non-commercial' use entails in practice). Other important information, again outlined in the RAI Croissant extension[51], includes intended access protocols and data protection declarations, which should similarly be made available.

3) Surrounding infrastructure

For a dataset to be AI-ready, it needs to be published within a surrounding infrastructure that is also AI-ready.

a) Datasets should be accessible via a user-centric data portal – A sufficiently user-centric AI data portal facilitates user engagement with datasets. Design should draw on the ten principles outlined in Costa, Walker & Simperl (2020)[52] (including: organising datasets by usage, investing in discoverability, publishing comprehensive metadata, adopting interoperability standards, co-locating documentation and analytical tools, and ensuring accessibility) to be robustly suitable for AI. For example, portals like the European Data Portal[53] actively support automated and manual workflows by providing interactive tools for exploring, evaluating and validating data relevance and quality alongside a toolset for data aggregation. This empowers practitioners to assess and integrate massive, continuously updated datasets into AI systems, thereby maintaining a responsive and adequate infrastructure in rapidly-evolving analytic contexts.

b) Datasets accessible via AI-ready APIs are best practice – Large, well-known AI datasets are easily accessible for data scientists and practitioners via their API infrastructure, enabling users to access and download a dataset with a single line of code. However, a genuinely AI-ready data infrastructure should not solely depend on API access; it must also incorporate standardised protocols suited for AI practitioners' usage. Research on user-centric data publishing indicates that such approaches

[51] Akhtar et al. (2024). Croissant RAI Specification.
[52] Costa, Walker & Simperl. (2020). Sustainability of (Open) Data Portal Infrastructures: Open Data Portal Assessment Using User-Oriented Metrics.
[53] data.europa. (2023). A European approach to Artificial Intelligence and the role of open data.

enhance data discoverability and usability, empowering practitioners to assess whether a dataset meets their needs quickly[54] [55]. Similarly, data access technologies like model context protocol (MCP) can be designed to enhance AI agents' usage rather than practitioners' [56].

When discussing what makes an API AI-ready, interviewees claimed it should employ a RESTful architecture (which is particularly important if we don't know the use cases of a dataset, given the architecture's flexibility), avoid pagination (which could artificially bottleneck their work), and facilitate the querying of subsets and splices of datasets (which is functional in AI contexts, where datasets are often vast). Alternatively, datasets can be made accessible on data spaces[57], which are standard organisational interfaces enabling secure, governance-compliant data sharing[58].

Data spaces are increasingly recognised in academic and industry literature as critical to achieving AI-readiness. These platforms employ data space connectors to transport data from publisher to user, facilitating data 'transactions' while preserving data sovereignty and integrity[59]. Such infrastructures address the challenges associated with massive, rapidly updating datasets by ensuring reliable, real-time data integration. In this way, they support effective machine learning and inference processes, making them particularly beneficial for AI applications. This benefit is demonstrated and further explored in the EU CEDAR project[60].

c) **Version control infrastructure is best practice** – A fit-for-purpose version control is essential for AI applications. For instance, generative AI systems rely on continuously updated, massive training corpora or real-time online data to ensure their responses remain current and relevant. Tools such as Git or Data Version Control (DVC)[61] enable granular monitoring of data changes throughout the AI data lifecycle. This ensures transparent provenance tracking and facilitates the timely integration of new information for enhanced AI performance.

[54] Koesten et al. (2021). Talking datasets – Understanding data sensemaking behaviours. doi: 10.1016/j.ijhcs.2020.102562
[55] Koesten et al. (2019). Collaborative Practices with Structured Data: Do Tools Support What Users Need? doi: 10.1145/3290605.3300330
[56] Anthropic. (2024). Introducing the Model Context Protocol.
[57] European Language Data Space. n.d. What is a Data Space?]
[58] Inverarity, Simperi & Massey. (2024). What are data spaces and what do they do?
[59] Data Spaces Support Centre. (2023). Data Spaces 101.
[60] CORDIS. n.d. Common European Data Spaces and Robust AI for Transparent Public Governance
[61] Data Version Control. n.d. Versioning Data and Models.

Using the framework: two examples

In this section, we illustrate the application of our AI-readiness framework by conducting structured assessments of two datasets: one from the Protein Data Bank (PDB) and the other from the Linked Open British National Bibliography.

To use the framework when assessing a dataset, we:

1. Download the dataset and open it with suitable software;
2. Evaluate it against the 'Dataset Properties' criteria of the framework;
3. Find metadata that corresponds to the dataset, typically accessible on the dataset's web page;
4. Evaluate the metadata against the 'Metadata' criteria of the framework;
5. Navigate the website on which the dataset is hosted, identifying documentation about it and its functionality;
6. Evaluate the website and its documentation against the 'Surrounding Infrastructure' criteria of the framework.

Evaluation was qualitative, examining whether, and to what extent, each criterion has been met. The evaluations below are colour-coded in **green**, **amber** or **red** according to whether each criterion is **met**, **partially met**, or **not met**.

Evaluation of the AI-readiness of datasets on the Protein Data Bank (PDB)

Proteins, crucial molecules for all life, consist of long amino acid chains whose sequence is encoded by DNA. The amino acids in the chain all interact, making the macromolecule twist, fold and wrap around itself, giving proteins unique 3D shapes suited to their tasks. Understanding the relationship between structure and function is fundamental to bioinformatics, enabling the targeted design and synthesis of functional proteins for potentially revolutionary medicines. The Protein Data Bank (PDB) provides a global, openly accessible repository of nearly every protein sequence and molecular structure discovered over the last 50 years. IT has significantly contributed to biomedical advancements, having been used as training data for the Nobel Prize-winning AI model AlphaFold.

Example dataset: Human insulin

An illustrative example from the PDB is the dataset for the human insulin protein, with each

entry providing highly granular data, including:

- The protein's classification, organism of origin (here, Homo sapiens) and the expression system used to sequence it in an experiment (here, E. coli);
- The protein's underlying amino acid sequences;
- The sequencing experiment results and external validation metrics;
- The protein's mutation status relative to the original protein;
- Associated literature references.

As described above, we assessed this dataset against the criteria outlined in the AI-readiness Framework:

1. Dataset Properties

a. **International standards and norms:** It adheres strictly to the widely recognised mmCIF standard for macromolecular crystallographic data (PDBx/mmCIF dictionary v5.398). It employs standardised biochemical conventions (for example, three-letter amino acid codes such as Ala for alanine). Furthermore, UniProt accession codes are used to identify macromolecules.

b. **Semantic and logical consistency:** Labels and nomenclature follow IUPAC/IUBMB biochemical standards, ensuring semantic clarity across dataset entries. Also, there are uniform definitions and notations for biochemical modifications and structures.

c. **Identifiable class and source imbalance:** Explicit identification of protein sources and expression systems is documented. Class imbalance (for example, proteins versus other molecular entities such as water molecules) is implicitly captured, but could benefit from explicit quantification or summarisation in metadata.

d. **De-identification and anonymisation:** The dataset inherently excludes identifiable or sensitive personal data; therefore, anonymisation concerns are not applicable.

e. **Appropriate file format:** The data is provided in the mmCIF file format, which is optimised for molecular data and internationally accepted. However, AI-readiness could be enhanced through additional supplementary formats (for example, providing the amino acid sequence in a CSV alongside the mmCIF dataset) to facilitate integration with AI workflows.

2. Metadata

a. **Machine-readable metadata format:** While the mmCIF standard is structured, inherently machine-readable, and suitable for bioinformatics, adopting AI-oriented metadata formats (such as JSON-LD) could further enhance interoperability and usability within broader AI contexts.

b. **Dataset served with attached metadata:** Data and metadata are combined within the PDB files, ensuring integral metadata delivery alongside data entries.

c. **Basic technical specifications:** Documents dimensionality (3D atomic coordinates) and basic summary statistics (resolution, R-factor). Furthermore, refined experimental structure information (for example, method: X-ray diffraction and resolution details) is explicitly noted.

d. **Supply chain information:** Data collection methods and refinement processes (via software such as Phenix and DENZO) are documented. However, the metadata could benefit from additional explicit traceability regarding preprocessing details beyond experimental conditions; this information would likely be presented in the referenced literature, which is standard in the biomedical field but not ideal for AI practitioners lacking in subject matter expertise.

e. **Legal and socio-technical information:** The dataset usage licence is not explicitly detailed within metadata, but is specified on the PDB website.

3. Surrounding Infrastructure

a. **Accessibility via a user-centric data portal:** The dataset is accessible via a page on a transparent, usable data portal, which allows easy search, download, visualisation (including in 3D) and detailed data exploration, such as grouping structures and pairwise structure alignment.

b. **Accessibility via API:** The dataset is accessible through a robust, standard API provided by the PDB, facilitating efficient programmatic access.

c. **Version control infrastructure:** The PDB website facilitates audits and revisions, with updates automatically registered under ‘audit revision history’ documented on a dataset’s web page on the PDB and in its metadata.

Overall assessment

Overall, the PDB is AI-ready in compliance with established international standards and data accessibility. However, it could further improve by enhancing file format diversity, metadata clarity and richness, and explicit legal documentation within metadata. Our overarching assessment is that this deems the PDB (at least, the human insulin example) to be AI-ready.

Evaluation of the AI-readiness of datasets on the Linked Open British National Bibliography (BNB)

The Linked Open BNB is The British Library's public knowledge graph that describes books and serials published or distributed in the United Kingdom and the Republic of Ireland since 1950, as well as forthcoming titles supplied through the Cataloguing-in-Publication (CIP) programme. The linked-data subset contains metadata for 'over 5 million' publications exposed as RDF dumps and via a SPARQL endpoint. Each record is richly typed (for example, 'dct: BibliographicResource'), identified by a persistent URI, and cross-linked to external authorities such as Virtual International Authority File (VIAF), International Standard Name Identifier (ISNI) and Library of Congress Subject Headings (LCSH).

Example dataset: Admiralty Notices Record (BNB ID 001092622)

A typical entry to a BNB dataset includes:

- A publication's canonical label – for example, Annual summary of Admiralty notices to mariners;
- Unique BNB identifier (blt:bnb) and owl:sameAs link;
- The publication's creator – such as Great Britain. Hydrographic Department;
- Work/manifestation typing, publication place, subject headings and a Primo permalink.

As described above, we assessed this dataset against the criteria outlined in the AI-readiness Framework:

1. Dataset properties

a. **International standards and norms:** The RDF serialisations use a range of vocabularies, such as DC-Terms, BIBO, ISBD, RDA, FOAF, SKOS, W3C Time and Geo. Relevant external standards (such as ISO 639-3 language, MARC country, VIAF/ISNI identifiers) are followed throughout.

b. **Semantic and logical consistency:** This data was formerly in an older format used by the British Library called 'Machine-Readable Cataloguing' (MARC). It has been converted into RDF, which applies authority control (every person, place or subject is matched to an official identity record) and creates owl:sameAs links (a technical label for a link between two values representing the same thing). However, work-level entities (books) are not fully deduplicated, and some legacy rdfs:literals persist, so a minority of resources violate strict normalisation rules.

c. **Identifiable class and source imbalance:** Vocabulary of Interlinked Datasets (VoID - a

small machine-readable factsheet that accompanies the data) statistics show ≈ 4.7 million monograph manifestations (individual book editions) versus ≈ 0.3 million serials (publications in parts over time, such as journals, magazines or annual reports). Unsurprisingly, English-language material dominates. This imbalance is acknowledged implicitly via those counts, but not discussed in an explicit bias statement.

d. **De-identification and anonymisation:** Only public bibliographic data (authors, publishers, subjects) is present; there is no personal or behavioural data, so GDPR risk is minimal.

e. **Appropriate file format:** Bulk data is supplied as RDF/XML, N-Triples and CSV (~320 MB per monthly uncompressed dump). No Parquet or JSON-LD derivatives are provided, so users targeting AI pipelines may have to convert the RDF graph to a tabular or embedding-friendly format for greater ease.

2. Metadata

a. **Machine-readable format:** Each release includes a VoID description. Each DOI landing page additionally carries a block of DataCite-JSON metadata, but there is no Croissant or JSON-LD metadata bundle.

b. **Dataset served with attached metadata:** Every dump includes the VoID file, and the SPARQL endpoint (a search box for data stored as RDF 'triples') returns the same descriptive triples with query results.

c. **Basic technical specifications:** Modality (structured graph); dimensionality (triple counts, distinct subject counts and dump size are declared in VoID); semantics (namespaces and property usage are documented; no SHACL shapes are published); bias (no explicit bias/representativeness statement); summary statistics (VoID provides counts, but not value-level distributions); synthetic data (none used or generated). No bias statement, synthetic-data flag or shape constraints are included.

d. **Supply-chain information:** The VoID describes the MARC-to-RDF conversion, Unicode normalisation and monthly refresh cycles. The Extensible Stylesheet Language Transformations (XSLT) is open source on The British Library's GitHub. DOIs of the form 10.22021/LODBNBx identify each dump. However, a formal change log or diff between releases is not provided.

e. **Legal and socio-technical information:** The data is released under CC0 1.0 Public-Domain Dedication with a non-binding attribution request. Licence terms are plain-text in VoID but not embedded in every record.

3. Surrounding infrastructure

a. **Accessibility via a user-centric data portal:** A legacy portal from data.gov.uk and a new beta interface provide search and entity pages. The migration is ongoing, and legacy download links on data.gov.uk are scheduled for withdrawal. However, the original (redundant) data.gov.uk page is still published, and neither portal contains very clear or sophisticated visualisation or exploration features.

b. **Accessibility via API:** An unauthenticated SPARQL endpoint is available but supports SPARQL 1.0 only. Some undocumented limits (≈ 60s timeout, 10,000-row cap) affect extensive extractions. No REST interface is offered.

c. **Version-control infrastructure:** Monthly dumps (until mid-2020) are timestamped (BNBLODS_YYYYMM) and assigned DOIs. Regular releases were paused during the portal migration. No machine-readable diffs or provenance traces between versions exist.

Overall assessment

In summary, the Linked Open BNB is partially AI-ready. It scores highly on standards compliance, open licensing and persistent identifiers. Still, it falls short, as there are no Parquet/JSON-LD derivatives, SPARQL 1.1 features, shape/quality metadata, explicit bias statement, or change logs.

Conclusion

This report presents a framework for data publishers to make their datasets AI-ready. The framework's design bridges the gap between the high-level principles found in literature and the lived experiences of AI professionals, who must invest substantial efforts in preparing the data to train and fine-tune foundational models. In this sense, we move beyond the currently prevalent mix of high-level and non-AI-specific criteria to provide concrete guidance for data publishers and repository providers.

We combined key findings from a literature review, expert interviews, and the ODI's collective experience to underscore the importance of selected attributes for data, metadata and supporting infrastructure to be AI-ready. We investigated each requirement and collected it into our resulting framework.

Asked for opinions on the framework's design, multiple interviewees agreed it should be sectioned according to dataset properties, metadata information and surrounding infrastructure. However, they clarified that none of the three components should be considered independent. Instead, each is intrinsically linked to the other, and a holistic approach to data practices ensures preconditions (such as quality and metadata) are met, in addition to AI-readiness.

Similarly, one interviewee remarked that there should be no boundary between publishers and users of a dataset, a conclusion we have found across the research in our DCAI programme. A constant dialogue can create a positive feedback loop between data providers and AI practitioners, allowing publishers to improve and iterate their datasets, metadata and infrastructure according to usage.

These two recommendations – taking a holistic approach to data practices and encouraging dialogue between data publishers and users – should be central to implementing our AI-readiness framework.

Our framework will be further developed and refined. It could be expanded to provide instructions for implementing its recommendations or to include subject matter-specific requirements, such as those for medical or geospatial datasets.

It is important to note that AI-ready data practices do not solely help in AI contexts. Instead, the recommendations in our framework present overall best practices for data publishers to follow to ensure that their datasets are high-quality, responsible and, above all, valid for the entire technological ecosystem. This allows datasets to reach their full potential, improving efficiency, facilitating better decision-making, and providing the foundation for innovation.

Appendix I: Methodology

To design our AI-readiness framework, we combined insights from:

1. A literature review, in which we investigated the academic and private sector attitudes and requirements for AI data.
2. Interviews with domain experts, including a data infrastructure expert at a large machine learning platform, a chief data officer at Sage Bionetworks, a compliance specialist at AI & Partners, an institute associate professor of computer science & engineering and data science at the Center for Responsible AI at New York University, and an AI researcher at King's College London.
3. Our collective experiences at the Open Data Institute (ODI). We have been involved in data across various subject matters and contexts for more than 12 years.

Appendix II: Interviewee acknowledgements

Neil Majithia, Thomas Carey-Wilson and Elena Simperl authored this report. The wider ODI team provided additional input, and we thank the following interviewees for the insights they offered:

- Prof. Julia Stoyanovich, Institute Associate Professor of Computer Science and Engineering, Associate Professor of Data Science, and Director of the Centre for Responsible AI – **New York University**.
- An anonymous interviewee – **a large machine learning platform**
- Susheel Varma, Chief Data Officer – **Sage Bionetworks**
- Sean Musch, CEO and Founder – **AI & Partners**
- Michael Borrelli, Director – **AI & Partners**
- Nicole Obretincheva, PhD candidate, Department of Informatics – **King's College London**