BDV BIG DATA VALUE ASSOCIATION

AI-READY DATA PRODUCTS

BDV BIG DATA VALUE ASSOCIATION

AI-ready Data Products

List of authors (in alphabetical order):

- Clara María Pezuela Robles (ITI)
- Claudio De Majo (Gradiant)
- Daniel Alonso (BDVA)
- Ed Curry (Insight)
- Elena Simperl (KCL)
- Gabriella Laatikainen (VTT)
- Guven Fidan (Artech)
- Ioannis Chrysakis (Netcompany)
- Joan Giner Miguelez (BSC)
- Kuldar Aas (Ministry of Justice and Digital Affairs of Estonia)
- Neil Majithia (ODI)
- Pierluigi Plebani (Politecnico di Milano)
- Thomas Carey-Wilson (ODI)
- Thomas Hütter (SCCH)
- Tuomo Tuikka (VTT)

List of reviewers (in alphabetical order):

- Didier Navez (Dawex)
- Shane O'Seasnáin (VITO)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The AI Continent Action Plan[1], published by the European Commission in April 2025 highlights under its "Data for AI" pillar that "access to reliable and well-organised data is essential if the EU is to unlock the full potential of AI". The European Data Union Strategy in November 2025[2], (with the subtitle "Unlocking data for AI") details the need of "scaling up access to quality data for AI and innovation".

"Data for AI" has long been a key area of focus for BDVA community, and therefore was central to the BDVA responses and feedback to these EC initiatives, published as "Towards a European AI-Data Value Ecosystem"[3], and "Data at the core of Europe's Digital Strategy"[4].

To make these plans a reality, the BDVA community has identified an urgent need for the traditional concept of data product (focused on packaging and sharing datasets for general use) to evolve in order to meet the specialised requirements of AI, a new paradigm that in BDVA we refer to as "**AI-ready Data Products**"[5].

The present paper builds on past and ongoing discussions and activities within the BDVA community around the paradigm of "Data for AI" and how this can be embedded in the data product approach. The paper provides the path to follow to redefine the paradigm of data products in a way that AI practitioners can fully harness the power and value of data within evolving data ecosystems. In doing so, it aims to position these data ecosystems as strong catalysts for AI innovation by delivering industry-ready solutions tailored to the complexities and unique requirements of advanced AI applications.

To the best of our knowledge, ***this is the first attempt to systematically evolve the concept of a data product so that it accommodates the needs, constraints, and practices of AI systems and AI practitioners***. The work of the community has produced the following key needs:

- **Lifecycle:** The stages of the lifecycle need to be revisited to incorporate AI-ready elements with a particular focus on extended metadata requirements.

[1] https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan
[2] https://digital-strategy.ec.europa.eu/en/library/data-union-strategy-unlocking-data-ai
[3] https://bdva.eu/news/towards-a-european-ai-data-value-ecosystem/
[4] https://bdva.eu/news/data-at-the-core-of-europes-digital-strategy/
[5] https://bdva.eu/blog/ai-ready-data-products/

- **Technical:** The quality of data is central to its use by AI and techniques for preparing data for AI applications, and metadata models for describing data suited for AI consumption. There is a clean need to support both static data products to dynamic AI-ready Data Products, incorporating information from different stages of the data pipeline and explaining how these can be aligned with AI workflows.

- **Governance and Compliance:** Need to be enhanced to include the ethical dimension of Data Products, existing regulations and standards, and specific licensing and contractual considerations for AI models and AI-ready Data Products.

- **Readiness framework:** To drive adoption a readiness framework is needed to capture the broader requirements identified for AI-ready Data Products.

The convergence of data management best practices, modern data product principles, and AI-specific requirements is highly promising to enable "AI-ready" Data Products. The concept requires further exploration, knowledge exchange, co-creation, and validation across disciplines, industrial players, and other relevant stakeholders. Therefore, we plan to continue this work within BDVA community (including a new version of this document in 2026 that collects all these new developments), but also to extend the dialogue to external stakeholders, other associations, standardisation bodies, policymakers, industry actors and AI communities.

The BDVA community aspires to transform the initial ideas of this paper into concrete standards, tools, and methodologies that make AI-ready Data Products a reality in various industries and ecosystems, ultimately contributing to the successful implementation of the European Data Union Strategy through truly integrated AI–data value ecosystem that link infrastructure, data, AI, talent, regulation and innovation.

I.

INTRODUCTION

CONTEXT AND MOTIVATION

The term "data product" was originally coined in the seminal work of Zhamak Dehghani in 2019[6], where she presented the data mesh paradigm and, as a key aspect, applied the product thinking to datasets to make them easily discoverable, addressable, trustworthy, interoperable, secure and usable. Since then, the concept of a data product has been embraced by many organisations to streamline data reuse by both internal and external consumers across various use cases, reducing costs and saving time.

According to the Gartner Hype Cycle for Data Management in 2024, data products appear to be still on the rise but close to the peak of expectations. While there is no agreed-upon definition of data product, the concept of packaging datasets along with all the relevant elements identified by an organisation to facilitate their discovery, exchange, and consumption by others clearly supports data sharing and transactions. As an example, the European Committee for Standardization (CEN) identifies data product as a key element in a Trusted Data Transaction[7], defined therein as a "data sharing unit, packaging data and metadata, and any associated licence terms". This is why the data product concept has also been adopted by designers and implementers of data spaces, the instruments identified by the European Commission in their Data Strategy to break data silos in Europe and foster cross-sector and cross-country data sharing in a trusted and efficient way.

However, the primary application of data is moving more and more from traditional data analytics to increasingly sophisticated AI workflows. These impose unique demands on data, including specific descriptions, new data quality dimensions and metrics, and tools to facilitate risk assessment in compliance with standards or regulations like the AI Act. The traditional concept of a data product (focused on packaging and sharing datasets for general use) must then evolve towards supporting specialised requirements of AI, a new paradigm that in BDVA we refer to as "AI-ready Data Products"[8].

Leveraging in previous discussions, insights from several dedicated workshops[9], and the knowledge of BDVA community, the present document intends to reflect to what extent the current concept of data product and how it is today implemented respond to the needs of AI practitioners and AI applications. Notice that how AI can support the data product lifecycle is out of the scope of this paper.

[6] https://martinfowler.com/articles/data-mesh-principles.html
[7] https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/2024/cwa18125_2024.pdf
[8] https://bdva.eu/blog/ai-ready-data-products/
[9] https://bdva.eu/news/bdva-workshop-ai-ready-data-preparing-data-for-ai-impact/

Then, and referring back to the original definition of a data product (discoverable, addressable, trustworthy, interoperable, secure, and usable), we should be able to answer questions such as:

- Is the product described in a way that AI/ML communities can easily understand and assess?
- Does the product expose APIs and access points specifically designed for AI integration?
- Can the product be trusted when used to train AI/ML models?
- Is the data product able to interoperate with other AI/ML assets (e.g., via MCP or similar protocols)?
- Does the product ensure appropriate security and governance measures?
- Is the data properly prepared and formatted for direct use in AI models?
- Is the data product scalable enough to handle substantial AI-driven workloads?

And more generic questions, like:

- Should the data product incorporate specific components to better support its use by AI communities?
- How can we assess the readiness of a data product for effective use by AI practitioners?

This document addresses all these questions and tries to provide the path to follow to redefine the paradigm of data products in a way that AI practitioners can fully harness the power and value of data within evolving data ecosystems. In doing so, it aims to position these data ecosystems as strong catalysts for AI innovation by delivering industry-ready solutions tailored to the complexities and unique requirements of advanced AI applications.

Therefore, this paper is intended for a broad range of stakeholders engaged in the design, development, and use of data products in AI-driven contexts. First, it addresses AI practitioners who require reliable, well-prepared, and interoperable data assets to accelerate AI model development, training and deployment. It also affects data scientists and data product designers and owners, who are responsible for building and maintaining data products that can meet the requirements of AI ecosystems. In addition, the document is relevant for policymakers and standardisation bodies who shape the frameworks and regulations that govern trustworthy and compliant use of AI-ready data. Finally, industry should reflect on how data products can unlock new AI-driven opportunities in specific sectors.

The paper is structured as follows:

- **Section 2** presents three main aspects about data products: (i) components and (ii) stages of the lifecycle, as well as (iii) product owners and providers, and revisits them to incorporate AI-ready elements that are further explored in subsequent sections.

- **Section 3** explores the technical dimensions of an AI-ready Data Product. It begins with the quality of data in view of its use by AI, discusses techniques for preparing data for AI applications, and presents alternative metadata models for describing data suited for AI consumption. The section also examines the evolution from static data products to dynamic AI-ready Data Products, incorporating information from different stages of the data pipeline and explaining how these can be aligned with AI workflows.

- **Section 4** focuses on governance and compliance aspects, including the ethical dimension of Data Products, existing regulations and standards, and specific licensing and contractual considerations for AI models and AI-ready Data Products.

- **Section 5** advocates for the development of a readiness framework to assess all the aspects discussed throughout the document. It introduces the ODI readiness framework as a promising reference and proposes ways to extend it to encompass the broader requirements identified for AI-ready Data Products.

- Finally, the paper concludes in **Section 6** with a set of conclusions and recommendations, outlining next steps to further the evolution of the AI-ready Data Product concept.

II.

FOUNDATIONS

CONTEXT AND

MOTIVATION

Following the general introduction to data products in Section 1, this section focuses on two of their core dimensions: (i) lifecycle of a data product and (ii) elements of a data product. We explore how both aspects should be revisited, tailored and extended to better serve the purposes of AI needs and practitioners. This section also reflects on (iii) the impact that this new paradigm has on data product owners and providers, paving the way for future discussions in this regard. The aim of the section is to identify the key elements that will be further explored in the subsequent parts of the document.

2.1. Lifecycle of AI-ready data products

Like any other product, a data product follows a lifecycle consisting of different stages that ensure it remains aligned with the objectives of its provider, addresses the needs of its consumers, adapts to internal and external conditions, and can be effectively managed over time. What distinguishes this specific lifecycle is that every stage is intrinsically shaped by the data perspective.

There is currently no universally accepted standard defining the stages of a data product lifecycle, and terminology may vary across organisations and frameworks. However, after reviewing various sources and literature[10][11] [24], the following list presents a consolidated approach that captures the core stages in a comprehensive manner:

- **Inception**: this stage involves the identification of needs, objectives, and expected value, according to the business requirements and expectations of the provider

[10] https://www.ibm.com/think/topics/data-product
[11] https://kpmg.com/us/en/articles/2025/data-product-lifecycle.html

- **Design**: definition of the data product's structure and architecture, decision on what elements will compose it, and identification of relevant data sources
- **Development**: implementation of all the components of the data product
- **Packaging**: process of bundling all components into a coherent and usable product, including metadata and any associated license terms.
- **Publication**: make the product available to potential consumers by the publication of its metadata and associated license terms through a catalogue, marketplace, connector or similar.
- **Governance**: establishing mechanisms to manage properly the data product, according to rights of data holder, internal policies and external regulations
- **Consumption**: allow consumers to easily access, understand and use the data product for its intended purpose through specific APIs
- **Monitoring and maintenance**: continuous tracking of quality, usage, and performance metrics, addressing issues as they arise
- **Evolution and scale up**: iterative improving and adaptation to new conditions, requirements, changing landscape, new technologies, data sources, etc ...
- **Retirement**: decommissioning the data product due to reasons like lack of usage, non-compliance, obsolescence, replacement ...

Note that not all the stages described are strictly sequential, as some operate across multiple dimensions and span several phases, for example, governance, monitoring, and maintenance. Likewise, evolution and scale-up are not stages in the strict sense, but rather continuous phases that may occur throughout the entire lifecycle.

The final goal of these steps is to make sure that the data product can be seen by the consumer as an element of value and, at the same time, improper use of data is avoided, to preserve this value also from the provider perspective. As a fundamental property of data product, data product ownership must be always ensured to: (i) clearly define the responsibility on the proper delivering of the data product to the consumer; (ii) preserve the improper use of data considering that data can leave the boundary of the organisation and full control is not always possible.

Based on the identified stages of the data product lifecycle, we propose the following adaptation to AI requirements and needs (Table 1, next page):

Table 1. AI-ready stages of data product lifecycle

AI ready stage	AI ready stage
Inception	·Consider specific AI-driven use cases that the data will support (model training, fine tuning, inference, …). ·Assess feasibility of AI-related data
Design	·Define the architecture of the AI ready Data Product, revisiting the usual components to ensure AI usability (data for AI, metadata, licences) ·Incorporating new AI-oriented components (see previous section)
Development	· Align data pipelines with AI workflows · Ensure compatibility with MLOps,
Packaging	·Package data in structures that are easy for AI workflows to consume ·Include AI-specific documentation ·Ensure metadata explicitly describes AI relevance
Publication	·Provide metadata in machine-readable formats suitable for AI ·Provide access points or APIs accessible by AI applications and AI agents (MCP)
Governance	·Extend governance to cover AI-specific issues: data ethics, fairness, bias monitoring, consent and privacy for AI use cases, copyright for model training ·Consider AI licenses ·Auditability and traceability mechanisms across the data and AI pipelines
Monitoring and maintenance	·Continuously monitor data quality, model performance degradation, bias evolution, and consumption patterns. ·Support MLOps.
Consumption	·Enable seamless use of the data product in AI workflows ·Expose an "AI friendly" API that defines the allowed interaction with the data product ·Consider the possibility of executing the AI workflow on the provider or consumer side
Evolution and scale up	·Adapt to new AI regulations ·Ensure scalability for increasing demand ·Support extension to new AI trends, synthetic data generation, …
Retirement	·Determine when the product is no longer fit for AI use due to outdated data, regulation changes, or better alternatives. ·Ensure responsible retirement. ·Guidance on impacts and mitigation for models trained with this data

2.2. Elements of an AI-ready data product

As mentioned in Section 1, there is no common agreement on the components of a data product. Because a data product is typically designed to serve the interests of the organisation that creates and provides it, its composition often reflects the organisation's specific objectives, aiming to be competitive, appealing, and easy to use. However, in an environment where data sharing is increasingly common, the ultimate use of a data product may no longer be known in advance by its provider. As a result, data products can range from very simple approaches (such as in CEN Trusted Data Transactions, consisting basically of data, metadata, and licenses), to more elaborate packages that include elements like dashboards, templates, services, or even AI models. For this reason, it is fundamental – as also suggested in the seminal work of Dehghani – to mediate the interaction with a data product by means of an API. This section reflects on the different components of a data product, moving from basic to more extended versions, and concluding with the additional elements required for an AI-ready data product.

2.2.1. Basic data product

According to CEN Trusted Data Transaction Workshop Agreement CWA (and also adopted by CEN CENELEC Joint Technical Committee JTC25 "Data, Dataspaces, Cloud and Edge"[12]), a data product can be defined as a data sharing unit, packaging:

- data
- metadata
- any associated license terms

In order to facilitate discoverability, understandability and ability to transact the data product, the descriptive information associated with the data product would include:

- specific purposes the data product is intended for
- terms of usage
- legal terms
- commercial terms
- price, if any
- consent and authorisations

[12] https://standards.cencenelec.eu/ords/f?p=205:7::::FSP_ORG_ID:3485479&cs=1FB461F74403B24687DFD3DDC632A58E6

2.2.2. Extended data product

As mentioned in the introduction, the initial concept of data product comes from the data mesh architecture, where it was conceived as "a logical unit that contains all components to process and store domain data for analytical or data-intensive use cases and makes them available to other teams via output ports"[13]. In this sense, the data product should incorporate, on top of the basic data product, the following components (see Figure 2):

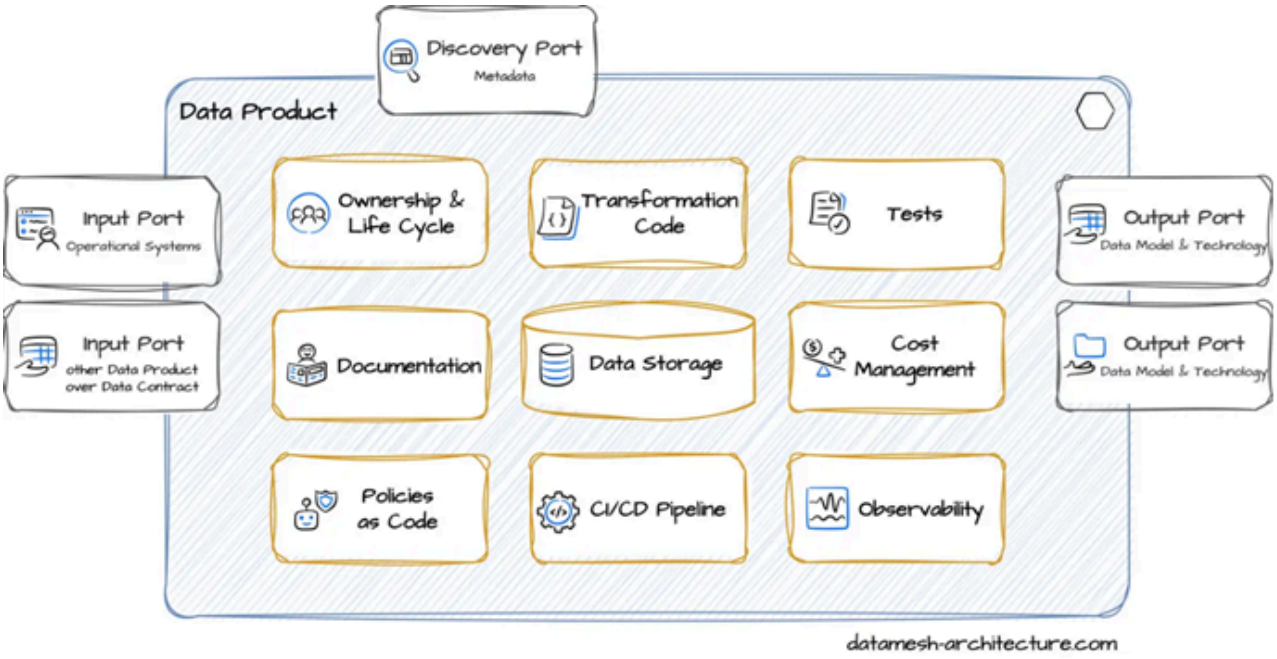

Figure 1. Data Product Internal Components (datamesh-architecture.com)

A data product is complemented with the **data service interface**, a collection of operations accessible through an interface (API) that provides access to the data or data processing functions offered by the data product[14]. These services act directly over specific datasets in the data product, performing some specific action over them in order to facilitate their access and use, such as data selection, extraction, combination, packaging, processing, delivery, interpretation and reuse[15]. Although the API offers a complete set of possible actions, data governance can define different levels of authorisations to different consumers, based on internal policies, norms, and regulations.

[13] https://www.datamesh-architecture.com/
[14] https://www.w3.org/ns/dcat#DataService
[15] https://dssc.eu/space/BVE2/1071257170/Value+creation+services#2.1.2-Data-handling-services

Another approach is provided by the **Open Data Product Specification**[16]. According to this, a data product should consider technical (infrastructure and access), business (pricing and plans), legal (licensing and IPR) and ethical (privacy) aspects, and, based on this, consider the following attributes and elements:

- Contract
- Details, including business details in different languages
- Pricing plans
- SLA
- Data quality
- Data access
- Payment gateways, to define how transactions are handled when pricing plans require financial exchanges
- License
- Data holder

Finally, in the Open Science context, although the term data product is not directly mentioned, **RO-Crate**[17]offers a machine-understandable format used to describe digital objects with metadata to ensure the FAIR properties (Findability, Accessibility, Interoperability, and Reusability). The metadata set includes information about the provider, links to external resources/identifiers, and information about storage and licensing.

2.2.3. AI-ready data product

Considering the product lifecycle described in Section 2.1, the introduction of AI workflows brings a set of distinct requirements that traditional data products were not originally designed to address. These requirements arise from the way AI systems (particularly large language models, retrieval-augmented systems, multimodal models, and autonomous agents) interpret and consume data, and include, among others, machine-interpretable semantic descriptions, pre-chunked, modular, and hierarchical data representations, support for diverse and AI-relevant data types (geometric, spatial, time series) and scalable and optimised data for model training. Building on these requirements and the traditional elements described in Section 2.2.2, an AI-ready Data Product should incorporate the following capabilities and extensions:

[16] https://opendataproducts.org/v4.0/#specification-aims-and-aspects
[17] (https://w3id.org/ro/crate/1.2)

- **AI ready datasets**, that include well curated datasets with enough quality to be used for its intended purpose (training, fine tuning, inference and evaluation), considering scale-appropriate preparation, de-duplication, and the potential use of synthetic data for augmentation.
- **AI ready metadata description**, enriched with machine-interpretable semantics aligned with vocabularies and knowledge structures widely adopted in AI communities
- Data pipeline alignment with AI workflows, capable of producing pre-chunked representations fitting within the context window, and enabling seamless integration into AI development cycles
- **AI oriented ports / interfaces**, including access mechanisms that expose data in formats, APIs, and protocols optimised for AI systems and agents. These may include integration with Model Context Protocol (MCP) and adaptation to other architectures (e.g., RAG-based or graph-based GraphRAG retrieval systems).
- **Integration with AI ecosystems**, enabling compatibility with AI development tools, model interaction frameworks, and agentic systems to facilitate seamless consumption of the data product.
- **Lineage tracking compatible with MLOps**, ensuring data traceability across preprocessing, training, deployment, ...
- **AI-specific licenses**, going beyond traditional data licences and considering specifically designed for AI scenarios
- **Ethics, fairness, and responsible AI**, allowing the embedding of these aspects into the data product.

This composition of elements is shown in Figure 2. It is important to note that, as with the traditional concept of a Data Product, the specific components and structure of an AI-ready Data Product will ultimately depend on the provider's strategy and the client's intended use. Therefore, the proposed composition is not meant to be definitive, but to highlight the importance of a new approach that incorporates AI driven elements not previously considered.

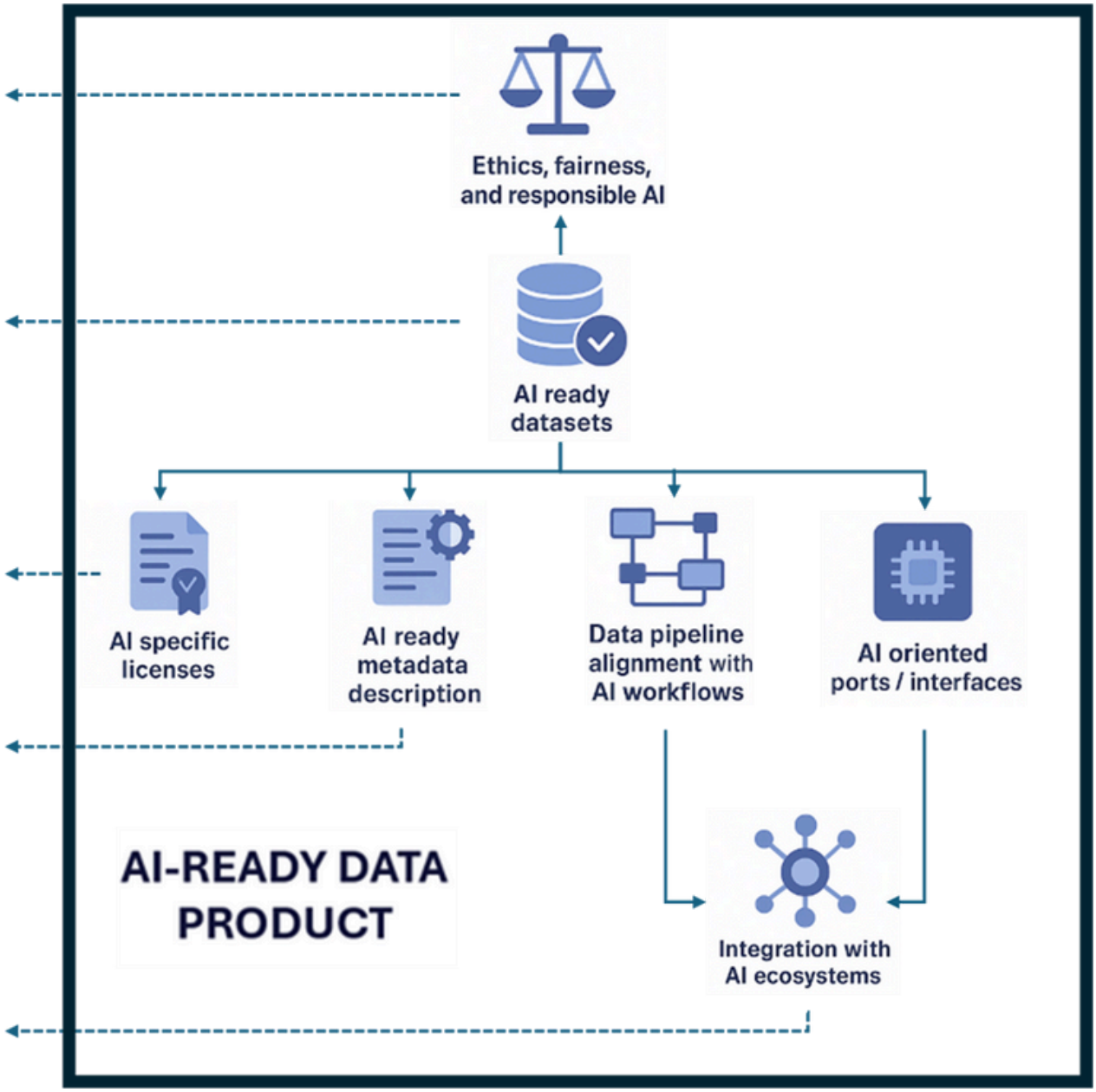

Figure 2. Elements of an AI-ready Data Product

2.3. AI-ready data product owners and providers

The shift this document proposes from traditional Data Products to AI-ready Data Products has also significant implications for data owners and providers. This evolution not only changes the technical requirements for all stages listed in section 2.1, but also affects required skills, governance responsibilities, operational workflows, and the tools needed to support the creation of data products optimised for AI use:

- Organisations will need practical guidance on how to implement the AI-ready data product lifecycle. Large organisations may rely on structured processes and dedicated teams, while SMEs may need lighter, more automated workflows. Different operational models will be relevant depending on organisational size, maturity, and sector.
- The need for new skills and competencies among data product owners and providers, including domain experts from across the organisation (e.g., Finance, HR, Operations) who are responsible for constructing and maintaining data products while understanding this new paradigm and. Beyond traditional data engineering, these roles now require awareness of AI-specific needs such as optimisation for AI consumption, robustness, trustworthiness, AI-driven data quality, and rich AI-ready metadata, all while remaining aligned with enterprise-wide standards..
- The adoption of new tools to produce and maintain AI-ready Data Products, that can include the use of AI agents and LLMs that extract and convert information into the standards and formats chosen by the company in support of these experts.
- Support (especially for SMEs) from existing initiatives, such as data intermediaries, data hubs, and sectoral data spaces by providing shared tools for standardisation, validation, anonymisation, and quality checks. These initiatives can help reduce the burden on providers and promote interoperability across organisations.

In summary, this transition presents substantial organisational, operational, and technical challenges, whose in-depth review is beyond the scope of this paper. Nevertheless, considering the aspects discussed here provides a foundation for identifying key elements that can inform future discussions and recommendations.

III. TECHNICAL BASIS FOR AI-READY DATA PRODUCTS

Building on the elements introduced at the end of the previous section, this one examines the key technical aspects of an AI-ready data product. We begin by revisiting the notion of data quality in the context of AI, emphasising current challenges, the state of the art, and future directions. Next, we introduce and explore techniques for preparing datasets to meet the quality required by AI systems. We then turn to the role of metadata, discussing the importance of defining models that are also adopted and understood by the AI community. Following this, we review APIs and access mechanisms specifically designed to interface with AI applications. Finally, we propose a shift from traditional static data products to dynamic AI-ready data products that continuously align the stages of the data pipeline with the steps of the AI workflow.

3.1. Data quality for AI

In its previous work “Elevating Data Quality: A Paradigm Shift for Data Spaces and AI Needs”[18] BDVA defines data quality as the “the extent to which data accurately represents the reality it seeks to capture”. This work also introduces the fit-for-purpose data quality to respond to specific goals of use cases and applications that will eventually use those data, and metrics defined and assessed by the data users instead of the providers. Data quality for AI and machine learning represents a relevant example of this fit-for-purpose paradigm. In fact, in the age of data-centric AI, the phrase "garbage in, garbage out" [1] has never been more relevant. Poor data quality undermines the reliability of AI systems, leading to various risks:

- *Hallucinations*: AI models trained on noisy or incomplete data may generate false or misleading outputs, resulting in wrong predictions and operational inefficiencies
- *Bias*: Skewed or unrepresentative data can propagate discrimination, raising serious ethical concerns and regulatory risks.

[18] https://bdva.eu/news/elevating-data-quality-a-paradigm-shift-for-data-spaces-and-ai-needs/

- *Data Shift*: Changes in data distributions over time (data drift) can lead to concept shift, where models no longer reflect the real-world phenomena they were trained to predict, again causing inaccurate results and increased costs. In industrial settings, these consequences translate into financial losses, flawed decision-making, and reputational damage [2].
- *Missing data*: on occasions the existing real data is uncomplete due to a bad capture or corrupted elements, so the provided information behind is discontinued and may lead into misinterpretations by AI models.

Finally, the identification of the gaps in data and generation of synthetic data to cope with them is becoming an emerging trend to overcome this risk.

Unlike traditional data systems, AI applications require a more nuanced approach to data quality. As mentioned, this aspect was initially explored in the BDVA discussion paper "Elevating Data Quality: A Paradigm Shift for Data Spaces and AI Needs", where fit-for purpose quality specifically applied to AI / ML and some ML oriented metrics were presented. This paper also explained how some European research & innovation projects are developing techniques to define metrics specifically tailored to intended user purposes.

In the case of AI, the suitability of data is not only judged by consumers but also by its **impact on model performance**. However, there is no universally accepted framework for evaluating data quality in AI contexts [3], including suitable quality dimensions and metrics. This lack of consensus complicates efforts to ensure data excellence [1].

Modern data quality (DQ) frameworks [4] emphasise the need for **continuous and automated evaluation** to meet the demands of AI-driven systems. State-of-the-art approaches must support robust measurement capabilities, scalable storage, automated analysis, and seamless integration into data pipelines. Unlike traditional DQ assessments, AI-ready data products require **model-aware quality checks**, where tasks such as outlier detection or missing value imputation are **tailored to the specific machine learning models** in use. Further guidelines for standards on data quality for AI have also been proposed by the ISO/IEC 5259 standard[19]. Tools like Great Expectations[20], Deequ[21], EvidentlyAI[22], Informatica[23], Experian[24], and Ataccama[25] have emerged to support these needs, offering features such as rule-based validation, statistical profiling, and drift detection. However, these tools often fall short in addressing the semantic and contextual nuances required for AI applications, highlighting the **need for more adaptive and intelligent DQ solutions.**

[19] https://www.iso.org/standard/81088.html
[20] https://greatexpectations.io/
[21] https://github.com/awslabs/deequ
[22] https://www.evidentlyai.com/
[23] https://www.informatica.com/products/data-quality.html
[24] https://www.edq.com/
[25] https://www.ataccama.com/platform/data-quality

As AI systems become increasingly data-centric, future research in data quality must evolve to **integrate tightly with AI model development workflows [5]**. Traditional data quality tools often fall short in addressing the dynamic, context-sensitive requirements of AI applications, offering limited support for semantic understanding, model-specific preprocessing, and adaptive quality metrics. A promising direction involves leveraging large language models (LLMs) to perform semantic evaluations and contextual checks, enabling **more intelligent and scalable assessments of data relevance and consistency**. Another key challenge is the **aggregation of heterogeneous data quality metrics**—from completeness and accuracy to fairness and drift detection—to reduce inspection fatigue and support holistic decision-making. Moreover, there is a pressing need to **revisit existing data quality dimensions** and introduce **new metrics** tailored to AI use cases, such as explainability, representational fidelity, and temporal stability. Addressing these challenges will be critical to ensuring that data products are not only technically robust but also ethically sound and operationally reliable in AI-driven environments.

3.2. Preparing AI-ready datasets

The previous section highlighted the great importance of specific data quality when using these data to feed AI models. Therefore, it is clear that the preparation of AI-ready datasets transcends traditional data collection and management techniques and should incorporate specific and sophisticated techniques that dramatically enhance model performance and reduce development cycles. Some of these techniques are:

- **Feature optimisation** involves strategic selection, transformation, and engineering of data attributes to maximise signal-to-noise ratios and expose underlying patterns that AI algorithms can efficiently leverage
- **Data enrichment** incorporates contextual information, domain knowledge, and cross-referencing with external sources to add semantic depth and relational understanding
- **Data augmentation** further extends dataset utility through synthetic data generation, controlled perturbations, and domain-specific transformations that improve model generalisation, address class imbalances, and enhance robustness against edge cases.
- **Synthetic data generation** complements (not replaces) real data when privacy, scarcity, or imbalance are constraints.

3.2.1. Feature optimisation

Feature optimisation is the foundation of developing AI-ready data sets because it directly influences models' ability to recognise useful patterns from raw data [6]. Feature optimisation involves careful selection, transformation, and engineering of attributes in a way that maximises the signal-to-noise ratio while reducing redundancy. Selection strategies ranging from filter methods by statistical correlation to wrapper and embedded that consider subsets of features during model training help reduce dimensionality and improve efficiency. Transformation also enhances data utility by scaling, normalisation, or discretisation, with more advanced approaches such as polynomial expansion or representation learning exposing underlying patterns otherwise not apparent. In text, image, or time-series spaces, such mappings may take the form of embeddings, spectral features, or wavelets that allow models to successfully learn domain-specific features. Finally, shared feature store deployment guarantees that engineered features are reused frequently across training and inference platforms, preventing duplication and ensuring feature lineage transparency and audibility.

3.2.2.Data enrichment

Data enrichment introduces external sources of knowledge, context signals, and semantic frameworks to the informational content of datasets [7]. This enables AI systems to be executed on data that is not just quantitatively sufficient but also qualitatively high. For example, datasets can be enriched with authoritative listings, weather or geospatial information, or cross-referenced against enterprise master data to make meaningful context. Semantic lifting techniques attribute properties to ontologies or taxonomies, creating machine-readable linkages that allow the information to be integrated with other resources and be better interpreted by algorithms. Knowledge graph creation is an extremely useful enrichment, allowing relational attributes to be extracted from raw tabular or text-based data. Temporal and spatial enrichment, such as adding lag features, moving averages, or geographic aggregations, allows models to identify temporal or spatial dependency. To maintain the enriched results trustworthy, each step of enrichment has to be followed by quality gates and provenance records so that the additional information is validated, reproducible, and fully transparent to downstream consumers.

3.2.3. Data augmentation

Data augmentation further increases the robustness and generalisation capacity of AI-ready datasets by synthetically increasing the size of training examples [8]. Computer vision processes data augmentation through geometric manipulation, colour perturbations, and more advanced methods like mixup or cutmix, which interleave or aggregate samples to expose models to novel patterns. In NLP, augmentation may involve synonym substitution, paraphrasing, or back-translation, all of which generate semantically equivalent but linguistically heterogeneous examples. Sensor and time-series data may be augmented using jittering, scaling, or time-warping to mimic real-world noise with temporal structure preservation. In table data, methods such as Synthetic Minority Oversampling (SMOTE) or class-conditional resampling are often utilised to mitigate imbalanced distributions that tend to bias model predictions. Above all, augmentation must be designed with domain constraints in mind to ensure the generated data remains realistic and label-conserving. As used wisely, augmentation reduces overfitting, improves fairness across underrepresented classes, and conditions models for edge cases that would otherwise yield brittle behaviour.

3.2.4. Synthetic data generation

Synthetic data generation is a meta-paradigm in AI dataset generation, particularly of high value whenever privacy, sparsity, or skewness limits the availability of the dataset in reality [9]. Modern generative models such as variational autoencoders, generative adversarial networks, diffusion models, and copula-based methods can generate high-fidelity synthetic data reproducing the statistical features of the original data without exposing any sensitive information. For graph- or sequential data, graph-based and autoregressive generators are becoming increasingly powerful. Besides utility, synthetic data also needs to meet high standards of privacy and regulatory compliance, especially when dealing with highly regulated domains such as finance and healthcare [10]. Differential privacy techniques, re-identification risk analysis, or k-anonymity provide safeguarding in such a way that nobody can be identified in the constructed datasets. Quality control is also crucial: synthetic data needs to be tested against real-world baselines on both statistical measures of similarity and task-specific performance metrics to establish its suitability for downstream use. Synthetic datasets need to come with proper documentation—such as generation method, privacy parameters, and intended use to instill accountability into their life cycle. When used responsibly, synthetic data enables risk-free experimentation, accelerates model development, and enhances the availability of training material without diluting trust and compliance.

3.2.5. When to use what

The choice of when to optimise, enrich, augment, or synthesise on the basis of features is mainly a function of the nature of the dataset and availability of the application domain. In the case of imbalanced labels, for example, synthetic oversampling or class-conditional data generation methods can neutralise bias and balance out with fairness metrics guaranteeing equitable outcomes between groups. When high-cardinality categorical features or sparse features exist in datasets, cautious encoding techniques and weight-learned embeddings offer a superior substitute to raw one-hot encodings. In regulated domains such as healthcare or finance, privacy-preserving synthetic data generation, augmented with differential privacy or robust provenance logs, becomes inevitable, allowing model construction without compromising compliance specifications. In the small data scenario, traditional augmentation practices, transfer learning, and enrichment from trusted external resources may provide the diversity necessary to achieve adequate performance. Such selections are not mutually exclusive; in practice, data sets optimized for AI usually employ several of these techniques in a layered configuration so that data products end up being robust, legally compliant, and tuned to actual AI processes.

3.3. Metadata description, management and tooling for AI

For AI-ready Data Products, metadata is essential to enable discoverability, interpretability, traceability, and automatic integration of datasets within machine learning (ML) and artificial intelligence (AI) workflows. Traditional metadata solutions, focused on descriptive and structural details, are insufficient to address the unique requirements of AI systems. Rather, AI-focused metadata also needs to record elements such as data provenance, model applicability, train/test splits, quality metrics, and algorithmic hazards.

In this context, newer standards have emerged that move beyond descriptive metadata to machine-actionable representations. **Croissant**[26] (from MLCommons) captures high-level ML dataset descriptions such as semantic types and learning tasks, combining metadata, resource file descriptions, and data structure into a single JSON-LD file[27]. Croissant makes datasets ML-ready, by enabling them to be directly loaded into ML frameworks and tools.

[26] https://mlcommons.org/working-groups/data/croissant/
[27] https://github.com/mlcommons/croissant

Similarly, **ML DCAT-AP**[28] extends the EU's DCAT-AP specification to describe ML datasets, models, evaluation metrics, and citations in line with EU semantic interoperability recommendations[29]. These formats are intended to enable compliance with the AI Act by embedding key information related to risk management and governance directly in metadata records. In the research domain, the Data Documentation Initiative Alliance (DDI Alliance) develops open metadata standards to enhance the quality, interoperability, and reusability of research data across social, economic, and health sciences. **DDI-CDI** (DDI Cross-Domain Integration) standard describes metadata for cataloguing and citation, with the fundamental purpose to describe data and process[30], and it is emerging as an option to enable data integration, discovery, and use by AI models in fields like social sciences and humanities.

Complementary enabling standards are also crucial. **PROV-O**[31] provides a W3C ontology for capturing provenance information (entities, activities, agents, and their relationships), which is vital for lineage, accountability, and auditability, mostly for the use of data to train AI models. **ODRL (Open Digital Rights Language)**[32], another W3C recommendation, allows expression of data usage rights and obligations, ensuring that licensing and access control policies can be attached to datasets in a machine-readable way. There are also interesting approaches complementing provenance and access rights for the industry, more focused on content certification. For instance, the **Coalition for Content Provenance and Authenticity** C2PA[33]) provides an open technical standard for publishers, creators and consumers to establish the origin and edits of digital content. On this regard, synthetic data generation presents new challenges in terms of data lineage. The Croissant Working Group is working in this direction, where it is very relevant to maintain the link between the real data that serves as a seed.

Beyond formal standards, the community has developed **documentation frameworks** that improve transparency and trust. **Model Cards** provide structured reporting on ML models, including intended uses, limitations, and performance across subgroups (Google Model Cards[34], Hugging Face[35]). **Datasheets for Datasets** offer a systematic way to document dataset motivation, composition, collection processes, and risks [11].

[28] https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/mldcat-ap
[29] https://semiceu.github.io/MLDCAT-AP/releases/2.0.0/
[30] https://ddialliance.org/ddi-cdi
[31] https://www.w3.org/TR/prov-o/
[32] https://www.w3.org/TR/odrl-model/
[33] https://c2pa.org/
[34] https://modelcards.withgoogle.com/
[35] https://huggingface.co/docs/hub/en/model-cards

In terms of management and tooling, **RO-Crate**, a community-driven specification, packages datasets and workflows with JSON-LD metadata, supporting reproducibility and interoperability in research contexts[36]. **Operational metadata tooling** is increasingly deployed to capture metadata automatically during the ML lifecycle. Platforms like **MLflow[37], Kubeflow Metadata[38]** and **TensorFlow Extended (TFX) Metadata[39]** track datasets, parameters, metrics, and artefacts throughout pipelines, making reproducibility and compliance auditable at scale.

Together, these standards, frameworks, and tools create a layered ecosystem for metadata in AI-ready data products: formal ontologies and vocabularies for data description (e.g., DCAT-AP, Croissant) complemented with provenance and access (PROV-O, ODRL, C2PA), structured documentation (Model Cards, Datasheets), and tools for management and operational pipeline integration (RO-Crate, MLflow, Kubeflow, TFX). This combination ensures that metadata is not only descriptive but actionable, bridging technical, legal, and ethical dimensions and supporting both interoperability in federated environments and accountability under the AI Act.

3.4. APIs and access

APIs are the mechanisms by which data products can be accessed by potential users, and eventually used in the intended applications. Traditional mechanisms include:

- REST APIs is the most commonly used way to facilitate data exchange between applications, and the most widely used standard for exposing data products
- GraphQL APIs is a more sophisticated method that exposes data models by using a single endpoint through which clients send requests. The API then accesses resource properties to get the client all the data they need from the query
- Streaming APIs are essential for real-time or near-real-time data products, and are the reverse of REST (where a request is made and a response is given) since here the server sends information to a client when an update is ready
- Semantic APIs add meaning and context to the data—using ontologies or knowledge graphs

[36] https://www.researchobject.org/ro-crate/
[37] https://mlflow.org/
[38] https://www.kubeflow.org/
[39] https://www.tensorflow.org/tfx/guide/mlmd

While these API types remain valid access points for AI applications, they might require some evolution to fully support AI workflows and models lifecycle. Additionally, with the rise of AI Agents, new opportunities emerge for accessing and interacting with data in more adaptive and intelligent ways. In this sense, AI agents would not replace traditional APIs, but they can enhance them by acting as orchestrators, understanding goals, providing context, and enabling more dynamic interactions.

In this context, an AI-Ready Data Product should integrate protocols like the Model Context Protocol[40] (MCP) or similar standards. Developed by Anthropic, MCP is an open protocol for connecting AI models (particularly large language models and agents) to external data sources, tools, and services in a structured way. MCP introduces features and design considerations that complement and extend traditional API types, enabling more seamless integration with AI-driven workflows

3.5. Data pipelines and AI workflows

In a traditional perspective, a data product can be conceived of as a static package: a pre-curated collection of data together with metadata and governance information. While this bundling is convenient for sharing and reuse in a generic sense, it is not sufficient for AI-readiness. Modern AI systems require data products to constantly evolve in response to model demand, regulatory compliance, and changes in the underlying data. This evolution redefines the notion of a data product from a static artefact to an operational living entity that, by packaging outputs from the different stages of the data pipeline, makes it easy to use and integrate into end-to-end AI workflows.

Data pipelines form the core of this approach. They scale and automate vital processes like ingestion from heterogeneous sources, data cleaning, feature engineering, labelling, and validation. Pipelines extend beyond preprocessing to provide version control, monitoring, and reproducibility, and data or schema changes are logically captured and auditable. This continuous and automated management of data pipelines reflects the principles of DataOps, which focus on data quality, collaboration, and governance across the full lifecycle of data products In AI cases, pipelines aren't technical plumbing within AI scenarios but enablers of high-quality, reliable data streams that directly impact model performance, fairness, and compliance.

[40] https://modelcontextprotocol.io/docs/getting-started/intro

AI workflows benefit from these pipelines by embedding them in the broader model training, testing, deployment, and monitoring lifecycle. This includes iterative loops where **feedback from deployed models is fed back into data collection and enrichment** to complete a continuous improvement loop. In this context, MLOps complements DataOps by extending operational principles to the AI model lifecycle, ensuring reproducibility, explainability, and compliance throughout model training, deployment, and monitoring." Pipelines and workflows together provide the operational underpinning in the scenario of AI-ready data products [12] [13] [15]. They make sure that data not only is technically ready but also is **contextually verified, legally up-to-date, and socially credible**. Bridging the gap between producers of data and consumers of models ensures early detection of drift, bias, or performance deterioration and their systematic correction.

Notably, **AI workflows** enable compliance-by-design: data minimisation checks, bias detection, explainability reporting, and secure audit logging can be incorporated as business-as-usual tasks. By this mechanism, workflows embed Responsible AI principles and oversight such as demanded by the EU AI Act. **The workflow orchestration** idea is the basis for making these abilities scalable. Workflow orchestration systems such as Apache Airflow, Kubeflow Pipelines, or Dagster provide a declarative method of defining, scheduling, and monitoring intricate pipelines. They allow modular tasks: data preprocessing, model training, hyperparameter tuning, and fairness testing—versioning and running them reproducibly in distributed environments. Orchestration not only guarantees scalability and reproducibility but traceability as well, since lineage information can be tracked automatically for auditability. With **orchestration**, data products may alter dynamically while being guaranteed quality, reproducibility, and compliance—quality traits required to incorporate them into real-world AI applications in areas such as agriculture, healthcare, mobility, and finance. Within emerging domains such as **data spaces**, these orchestrated pipelines and workflows are even more critical, as they enable trusted data sharing across organisations and borders, ensure interoperability through common standards, and provide verifiable audit trails that align with the governance frameworks of the European Data Strategy [14].

IV.

GOVERNANCE & COMPLIANCE OF AI-READY DATA PRODUCTS

This section addresses the non-technical dimensions of Data Products and explores how they should be revisited to respond to the needs and challenges introduced by AI. First, it highlights the importance of embedding ethical considerations arising from AI use and applications directly into the design and governance of Data Products. Particular attention is given to regulatory compliance, not only with frameworks specifically designed for AI (e.g., the EU AI Act) but also with conformity to standards related to AI risk management, transparency, and accountability. Finally, it examines the limitations of current data licensing approaches in the context of AI and proposes directions for developing AI-ready licensing models that better reflect the emerging realities of AI solutions.

4.1. Ethics dimension of AI-ready data products

This section focuses on ensuring that Data Products are AI-ethical (notice that data products are not only datasets, but can also incorporate AI models for which the ethics dimension is highly relevant). Alan Turing Institute defines AI ethics as "a set of values, principles, and techniques that employ widely accepted standards of 'right' and 'wrong' to guide moral conduct in the development and use of AI technologies" [16]. Further, AI Ethics can refer to a variety of attributes assigned to an AI system, such as trustworthiness, explainability and human-centricity [17] [18]. An overview of the most important criteria for AI ethical systems can be found in:

- IEEE Ethically aligned design[41]
- EU's Ethics guidelines for trustworthy AI[42]
- ISO/IEC 42001, an international standard that specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System within organisations [43]
- The global landscape of AI ethics guidelines[44].

[41] ead_v2.pdf
[42] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[43] ISO/IEC 42001:2023 - AI management systems
[44] Nature Machine Intelligence, 1(9):389–399, 2019. https://doi.org/10.1038/s42256-019-0088-2.

Based on these mentioned sources, some common elements of AI ethical systems are responsibility, accountability, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity. However, research shows that there is a significant gap between conceptual AI and its practical implementation, being one of the biggest challenges the lack of integration of AI ethical guidelines into software engineering practices. Thus, guidelines and recommendations are not enough to ensure that AI systems are ethical [19].

One of the key characteristics of ethical AI systems is explainability, which refers to the ability of humans to understand and trust the decisions made by AI models[45]. It is important to ensure the transparency of algorithmic decisions, so that ethical problems, such as biases in decision-making, can be identified and corrected. For instance, explainability tools, such as Local Interpretable Model-Agnostic Explanations (LIME) [20] can increase trust toward an AI system, by providing transparency and interpretability about their decisions.

There exist several AI Ethics tools, each of them having their own definition of what characteristics an ethical AI system should have: Microsoft's AI Fairness checklist[46], Google Responsible AI[47], Salesforce AI Ethics Model[48], Open Ethics[49].

Based on these inputs, the key ethical characteristics that need to be embedded in an AI-ready data products would be:

- **Explainability**: Data products should have metadata information that explains how AI decisions have been made to the users, indifferent of their technical expertise.
- **Fairness**: Unlike in the context of data spaces, fairness refers to the lack of bias and discrimination of AI models. Using a representative sample of data for training and regular assessments of AI models is a prerequisite to ensure fairness of AI-ready data products.
- **Transparency**: The openness and clarity related to the design of the AI system, the data sources, and decision-making processes. The metadata of the AI-ready data product should have information that ensures the transparency of how the data product was developed.

[45] https://www.ibm.com/think/topics/explainable-ai
[46] https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/
[47] https://cloud.google.com/responsible-ai
[48] https://www.salesforce.com/news/stories/salesforce-debuts-ai-ethics-model-how-ethical-practices-further-responsible-artificial-intelligence/
[49] https://openethics.ai/oemm/

- **Accountability**: this is related to the lineage, traceability and the information related to the data product owner. The metadata should describe the responsible entities of the development, publication, and management of the AI-ready data product.
- **Privacy**: AI-ready data products should adhere to the specific privacy regulations, and pay attention to the data minimisation, and secure storage of the data sources, and the data product during its whole lifecycle. The metadata should contain information whether any privacy-enhancing technology has been used.
- **Security**: Ensuring security, and protection against cyberattacks is very important also for AI-ready data products. This is related to ensuring the integrity of the data sources and AI models, and securing data products from unauthorised access, among others. To ensure the security of AI-ready data products, implementing specific encryption techniques and robust testing might be required.

4.2. Regulation and compliance

In the current age of intelligent systems and burgeoning development of increasingly performing AI models, ensuring regulatory compliance is an essential aspect to develop AI-ready Data Products, as witnessed by the growing amount of emerging legal frameworks, international standards and global policy principles. Navigating through these instruments can help better shape requirements for data transparency, traceability and ethical adoptions and use of datasets in AI systems.

Data Products must be governed by licensing agreements that are both legally robust and capable of adapting to specific demands and deployment mechanisms. These need to define usage rights, address data protection obligations through privacy-by-design principles and remove bias deriving from low data quality and discriminatory patterns. Equally important is the definition of frameworks clearly defining ownership issues and ensuring copyright law protection (see Section 4.3).

The **EU AI Act (Regulation (EU) 2024/1689)** attempts to address some of these issues, introducing a risk-based approach to AI regulation, which contains specific provisions pertinent to data products. In terms of data handling and governance, Article 10 insists on the need **to train, verify and test AI systems (especially high-risk systems) on relevant, representative, and error-free datasets**. Additionally, training, validation and testing datasets should:

- have the **appropriate statistical properties** (at the level of individual data sets or a combination thereof)
- consider the characteristics or elements particular to the **specific geographical, behavioral or functional** setting within which the high-risk AI system is intended to be used
- **appropriate safeguards for the fundamental rights** and freedoms of natural persons

Such a practice necessitates strong data governance procedures, such as annotation, enrichment, bias mitigation techniques, and documentation of data origin. Special categories of personal data may only be processed by providers under stringent controls to eliminate bias. Similarly, Article 50 on transparency obligations compels providers of generative and interactive systems to clearly label content produced by AI and notify users when interacting with it. This facilitates auditability and traceability by mandating machine-readable metadata and detection systems.

Besides, to ensure data accuracy and completeness, there is a need to implement appropriate risk management measures so that possible shortcomings are duly addressed. Noticeably, these requirements do not preclude the use of privacy-preserving techniques in the context of the development and testing of AI systems.

High-risk AI systems suggest the interest of defining domain-specific quality rules according to the nature of the existing features in the data sets, in the sense that different categories (geographical, behavioural, biometric, etc.) can have specific quality indicators, constraints and therefore treatments.

The **OECD AI Principles**[50], adopted in over 47 countries worldwide, provide another guide for ensuring that data governance procedures for AI-ready data products are in line with international standards for accountability, transparency, and fairness.

[50] https://www.oecd.org/en/topics/sub-issues/ai-principles.html

In particular, the principles promote privacy-preserving practices and bias mitigation strategies to ensure fairness and inclusivity, while also stressing the significance of documenting data sources and decision-making processes to promote transparency and explainability. Another essential principle is robustness and safety, which calls for AI systems and the data products that support them to be safe, robust, and monitored continuously over the course of their lives. In addition to providing technical and ethical guidance, the OECD AI Principles are strategically important for facilitating regulatory alignment and cross-border interoperability. The principles aid in bringing expectations regarding data quality, usage rights, and accountability into harmony as AI systems depend more on data from various jurisdictions. This is especially important for AI-ready data products that are meant to be deployed globally, as different legal frameworks can present serious difficulties. Developers and policymakers can promote inclusive innovation, lessen fragmentation, and enable responsible data sharing by incorporating the OECD's values into the design and licensing of data products.

In terms of compliance, ISO certifications are also important in establishing a structured and certifiable framework AI-ready data products governance. **ISO/IEC 42001[51] (AI management systems)**, the first artificial intelligence management systems (AIMS) standard in the world and mentioned in the previous section, provides businesses with a thorough framework for putting ethical AI practices into place throughout the AI systems' whole lifecycle. It offers a management system approach that incorporates risk assessment, policy development, and performance evaluation, with an emphasis on ethical oversight, transparency, and continuous improvement. It is especially pertinent for organisations creating or implementing AI-ready Data Products because of its **versatility across industries and organisational sizes**.

To help organisations identify and reduce risks like algorithmic bias, data quality problems, and operational vulnerabilities, **ISO/IEC 23894 (AI — Guidance on risk management)[52]** offers comprehensive guidance on risk management specific to AI. **ISO/IEC 5338 (AI system life cycle processes)[53]**, which outlines the development, deployment, monitoring, and retirement phases of AI systems, provides additional support for AI data governance. While adding AI-specific factors like model engineering, data sensitivity, and continuous validation, this standard expands upon well-established software engineering techniques.

[1] https://www.oecd.org/en/topics/sub-issues/ai-principles.html
[2] https://www.iso.org/standard/42001
[3] https://www.iso.org/standard/77304.html
[4] https://www.iso.org/standard/81118.html

ISO/IEC 42005 (AI system impact assessment[54]), on the other hand, focusses on impact assessments and helps organisations assess the potential effects of AI systems and the data products that underpin them on people, communities, and society as a whole. By mandating documentation of potential risks and benefits throughout the AI lifecycle, it encourages accountability and transparency.

4.3.Governance of AI-ready data products

As anticipated in Section 2.1, a key property of a Data Product is that its ownership must be always ensured. This guarantees both the accountability of the provider for its proper delivery to consumers and the prevention of inappropriate or unauthorised use of the data. When data is used for AI, it is not only about who owns the data product, but also about who controls how it is accessed, transformed, and reused within AI workflows.

Besides, while some existing standards and protocols to implement data provenance and traceability are presented when talking about metadata description of data products (section 3.3), some specific and additional aspects must be considered when applied to AI.

AI models are especially sensitive to specific features of training and input data, like quality, bias, completeness, representativeness, and others. Therefore, as also anticipated in Section 3.5. "Data pipelines and workflows", a detailed annotation of the source of the training data, preprocessing steps applied, different existing versions of the data, etc. is crucial to avoid bias and prevent incorrect model behaviour and low performance.

Besides, given the opaque nature of many AI systems (black boxes), data provenance and traceability are essential for knowing how inputs lead to outputs and understanding AI-based systems decisions (XAI). Therefore, in AI systems is key to know how data flows through preprocessing, model inference, postprocessing, … to enable reproducibility, compliance and identification of errors or bias.

Finally, decisions made by AI can have legal, financial, or ethical consequences. Therefore, it is crucial to keep auditable records of model predictions and other outputs, and the associated data used in the different phases of the model.

All these considerations must be integrated when designing an AI-ready Data Product, ensuring that relevant information is recorded in a standard and machine-readable format to support transparency, accountability, and trustworthy AI operations.

[54] https://www.iso.org/standard/42005

4.4. Data licensing and AI-specific contracts

This section serves as the contractual lynchpin for the governance framework of an AI-ready Data Product. While the preceding sections define principles and obligations, it is the data license or contract that renders them legally legible and enforceable. This section will explore the development of new legal instruments designed not merely to permit access but to strategically de-risk AI innovation and operationalise the very principles of responsible AI governance that define a modern data product. The foundational challenge in applying existing legal frameworks to AI is that they were engineered for a different technological and legal paradigm. Legacy licensing models, developed primarily for software or general creative content, are fundamentally ill-suited for the unique processes of AI development. Recent work[55] by the Open Data Institute and Duke University indicates that their application creates significant legal ambiguity, increases transaction costs, and ultimately hinders the responsible data collaboration necessary for AI innovation.

A core conceptual mismatch exists between the principles of copyright law, upon which many common licenses are built, and the mechanics of machine learning [21]. Traditional licenses, such as those from Creative Commons or various open-source software initiatives, are constructed around legal concepts of "reproduction," "distribution," and the creation of "derivative works" [56]. AI training, however, is a process of statistical learning and parameter adjustment, not direct reproduction of the training data in the final model. An AI model learns patterns, correlations, and weights from the data, but it does not typically retain copies of the complete original works in downstream stages of the AI data lifecycle[57]. This fundamental distinction makes the application of terms like "derivative work" or "adaptation" to a trained model highly contentious and legally uncertain across jurisdictions. Legacy licenses were designed to govern the relationship between a human creator and a human user of a creative work. They fail when the "user" is a machine learning algorithm and the "use" is an act of statistical abstraction rather than human consumption or adaptation [22]. This legal ambiguity manifests in several critical failure points, one of the most prominent being the interpretation of "non-commercial use." Research confirms widespread confusion over the definition of this term, particularly in scenarios where academic or non-profit research is later commercialised or contributes to a commercial product[58].

[55] https://theodi.hacdn.io/media/documents/Unlocking_data_collaboration.pdf
[56] https://creativecommons.org/using-cc-licensed-works-for-ai-training-2/
[57] https://theodi.org/insights/reports/understanding-data-governance-in-ai-a-lifecycle-perspective/
[58] https://theodi.hacdn.io/media/documents/Unlocking_data_collaboration.pdf

Licenses like Creative Commons Attribution-NonCommercial (CC-BY-NC)[59], while widely used for datasets, lack a clear, universally accepted definition of what constitutes "commercial" activity. This ambiguity creates significant compliance risks [23], acting as a major barrier to data sharing and deterring organisations, particularly smaller entities and startups that lack extensive legal resources to navigate such uncertainties.

Similarly, attribution requirements, which are a cornerstone of licenses like Creative Commons Attribution (CC-BY)[60], are often impractical to implement in an AI context. In a model trained on billions of data points, it is technically challenging to attribute every individual unit of data that contributed to the model's parameters or to a specific output generated by the model[61]. This increases the risk of non-compliance, undermining both the legal and ethical intent of the license. Creative Commons itself acknowledges these deep complexities[62], clarifying that its licenses are instruments of copyright and therefore only apply when copyright permission is required for a given use. In many jurisdictions, exceptions and limitations to copyright law may permit AI training without necessitating permission from the copyright holder, rendering the license terms inapplicable in those instances. This jurisdictional patchwork further compounds the uncertainty for developers building global AI systems.

4.4.1. Redefining rights and obligations for models and outputs

The most critical ambiguity left by traditional licenses concerns the legal status of the AI model itself and the outputs it generates. This uncertainty poses a significant risk to investment and innovation. Consequently, AI-specific contracts are being engineered to provide the legal clarity necessary to define and allocate these new forms of value. A central challenge is determining whether a trained AI model constitutes a "derivative work" of its training data. As highlighted in research on AI data-sharing practices, this is a primary source of legal uncertainty for developers. One interviewee noted that early attempts to use Creative Commons licenses failed precisely because "the number one question was 'what are the continuing rights or obligations as you pass through each machine learning phase?'". The risk is that a copyleft license[63] on the training data, such as CC-BY-SA, could be interpreted to mean that a multi-million-dollar proprietary model trained on that data must also be openly licensed under the same terms. This chilling effect has been a major impetus for the creation of new licensing models[64].

[59] https://creativecommons.org/licenses/by-nc/4.0/deed.en
[60] https://creativecommons.org/licenses/by/4.0/
[61] https://arxiv.org/html/2311.03731v2
[62] https://creativecommons.org/ai-and-the-commons/cc-signals/
[63] https://www.gnu.org/licenses/copyleft.en.html
[64] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4872366

To resolve this ambiguity, a new generation of data-centric licenses has introduced a crucial legal innovation: the explicit definition of AI models and their outputs as unrestricted "Results." This distinction serves the potential legal entanglement between the training data and the trained model, insulating the high-value model from the potential copyleft or restrictive terms of the often low-value (per-unit) input data. It is a strategic intervention designed to de-risk AI investment and encourage data sharing by providing legal certainty where copyright law is silent or ambiguous. Two key examples illustrate this approach:

- The **Linux Foundation's Computational Use of Data Agreement (C-UDA) v1.0**[65] defines a "Result" as "anything that you develop or improve from your use of Data that does not include more than a de minimis portion of the Data." Crucially, it explicitly states, "Artificial intelligence models trained on Data... are Results." The license then clarifies that it "does not impose any restriction with respect to the use, modification, or distribution of Results".
- The **Community Data License Agreement - Permissive v2.0 (CDLA-Permissive-2.0)**[66] adopts a similar strategy. It defines "Results" as "any outcome obtained by computational analysis of Data, including, for example, machine learning models and models' insights." It then unequivocally states that the agreement "does not impose any restrictions or obligations with respect to the use, modification, or sharing of Results".

Beyond the license grant, AI-specific contracts must also clearly allocate liability. Data providers, particularly those sharing data openly, seek to disclaim warranties about data quality and fitness for a particular purpose, and to limit their liability for how the data is used. This is evident in the strong "AS IS" and comprehensive limitation of liability clauses found in both the CDLA and C-UDA licenses. Conversely, data users, especially in commercial contexts, require a different set of assurances. They may seek contractual representations and warranties from the provider regarding the data's origin, its quality, and the provider's legal right to share it. Furthermore, they may demand indemnification from the provider against third-party intellectual property infringement claims that arise from the use of the training data. These negotiated terms are essential for building the trust required for high-stakes AI development.

[65] https://cdla.dev/computational-use-of-data-agreement-v1-0/
[66] https://www.newcastle.edu.au/library/teaching-and-research-support/copyright/open-licensing/other-licences/community-data-license-agreement-cdla-for-data

4.4.2. The emergence of AI-centric licensing paradigms

The contractual gap created by legacy frameworks is being filled by a rapidly evolving ecosystem of new licensing models. An AI-ready Data Product must be accompanied by a license that reflects a conscious choice of where it sits on this spectrum. Table 2 provides a comparative analysis of these emerging paradigms.

Table 2. Comparison of modern data and AI licensing models

	Key examples	Core philosophy	Treatment of AI models/outputs	Use case suitability
Permissive data licenses	CDLA-Permissive-2.0 [67], C-UDA 1.0 [68]	Maximise frictionless data sharing and computational use with minimal restrictions. Encourage innovation by de-risking model development.	Explicitly defined as unrestricted "Results," severing obligations from the source data.	Foundational model training, large-scale data analysis, and commercial AI development where legal certainty is paramount.
Copyleft data licenses	CDLA-Sharing-1.0 [69]	Ensure that improvements and modifications to the dataset itself remain open and are shared back to the community under the same terms.	"Results" (models/outputs) are unrestricted, but sharing of modified or enhanced data is subject to the copyleft provision.	Collaborative, community-driven data curation projects (eg, building a shared benchmark dataset).
Responsible AI Licenses (RAIL)	BigScience OpenRAIL-M [70], Stable Diffusion License[71], Llama 2 License[72]	Combine principles of open access with behavioural use restrictions to proactively mitigate potential harms and prevent misuse.	Use of the model and its outputs is explicitly restricted by ethical clauses (eg, no use for surveillance, disinformation, or in high-risk domains without safeguards).	Deployment of powerful generative models where harm mitigation is a primary concern for the licensor.
Proprietary/ custom agreements	Commercial API Licenses (eg, OpenAI, Google Cloud), Data Use Agreements (DUAs)	Grant specific, often limited and revocable, usage rights in exchange for a fee or as part of a service agreement. Control and monetisation are key.	Rights are explicitly defined and limited by the contract. Outputs may be owned by the user, but the vendor often retains rights to use input/output data for service improvement.	Commercial AI-as-a-Service platforms, enterprise solutions, and sharing of sensitive or high-value proprietary data.

[67] https://www.newcastle.edu.au/library/teaching-and-research-support/copyright/open-licensing/other-licences/community-data-license-agreement-cdla-for-data
[68] https://cdla.dev/computational-use-of-data-agreement-v1-0/
[69] https://docs.data.world/en/214274-common-license-types-for-datasets.html
[70] https://www.licenses.ai/blog/2022/8/26/bigscience-open-rail-m-license
[71] https://stability.ai/news/stable-diffusion-public-release
[72] https://ai.meta.com/llama/license/

4.4.3. Crafting the AI-ready data contract: essential provisions and a modular design

A truly AI-ready data product requires more than just a simple license grant; it necessitates a comprehensive contract that establishes a framework of trust, quality, and legal compliance between the data provider and user. Research on data collaboration practices has shown a strong demand for simplicity and standardisation to reduce transaction costs, alongside a concurrent need for flexibility to accommodate diverse data types, use cases, and legal regimes. The optimal solution is a **modular contractual framework**. This modular approach involves a set of standardised core terms that form the foundation of the agreement, combined with a menu of optional clauses or extensions that parties can select to tailor the contract to their specific needs. This balances the efficiency of standardisation with the necessity of customisation. Essential provisions in such a framework include:

- **Scope of use and restrictions**: The contract must clearly define the permitted uses of the data, for example, by specifying "Computational Use" as in the C-UDA. If the data is intended for use with a model governed by a RAIL, the associated behavioural use restrictions must be explicitly incorporated or referenced to ensure downstream compliance.
- **Representations and warranties**: To build trust, the data provider should make clear representations about the data's provenance, its known quality limitations, and their legal right to license the data for the intended purposes. These clauses provide a contractual basis for the data quality and provenance principles detailed in Sections 3.1 and 4.3 of this paper.
- **Intellectual property and indemnification**: The contract must unambiguously clarify ownership of the source data, any enhancements made to it, and the resulting AI models and outputs, as discussed in section 4.4.2. In commercial agreements, the data user will often require indemnification from the provider against third-party IP infringement claims related to the training data, shifting that risk back to the provider.
- **Confidentiality, privacy, and security**: For any non-public data, strong confidentiality clauses are paramount. The contract must mandate compliance with all applicable data protection laws (such as the GDPR) and specify concrete security requirements for data handling, storage, encryption, and eventual destruction.
- **Limitation of liability and disclaimers**: As seen in permissive open data licenses, providers will seek to disclaim all warranties and limit their liability to the greatest extent possible. In commercial contexts, these clauses are among the most heavily negotiated, with users pushing for greater accountability from the provider.

Central to the success of a modular framework is the development of standard definitions for key terms. Establishing clear, widely accepted definitions for concepts like "Data," "Result," "Non-Commercial Use," and "Computational Use" is critical. This provides a common legal language that can be used consistently across different agreements and modules, significantly increasing clarity and interoperability within the data ecosystem.

4.4.4. Licensing as a governance keystone

Ultimately, data licensing is not a mere legal formality but the central, operational mechanism for enforcing a data product's entire governance framework. The contract serves as the primary tool for operationalising regulations like the EU AI Act, with clauses mandating documentation on data provenance, bias mitigation, and quality assessments. For responsible AI, the license can be an instrument of enforcement, transforming ethical values into legal duties by incorporating behavioural use restrictions, as pioneered by RAIL licenses.

The future of AI data licensing lies in its deep integration with machine-readable technical standards, evolving the contract from a static document into a dynamic, socio-technical orchestrator. An advanced AI-ready data contract will connect its human-readable legal text to the data product's technical reality by requiring metadata to conform to standards like Croissant, which can embed rich information on licensing, provenance, and responsible AI considerations. In this model, the license acts as a central "pointer," legally activating and mandating compliance with other specialised governance artefacts such as technical metadata, ethical codes, and regulatory checklists. This positions the license as the bridge between the legal world of contracts and the technical world of data pipelines. In turn, this makes the concept of an "AI-ready data product" both legally and operationally viable.

V.

A FRAMEWORK FOR THE ASSESSMENT OF AI READY DATA PRODUCTS

As has been explored throughout this document, the concept of AI-ready Data Products represents a significant shift from the traditional notion of data products. On the one hand, it involves rethinking the conventional components of a data product (such as data, metadata, licences, and services) in terms of their suitability for AI applications. On the other hand, it introduces new elements specifically designed to address the requirements of AI practitioners. Overall, a systematic approach is needed to align the perspectives of data product providers and AI users, ensuring a shared understanding of what makes a data product fit for AI purposes. This leads to the need for an AI Readiness Framework for Data Products, capable of assessing and guiding their evolution toward effective use in AI-driven contexts.

In this way, there are already some frameworks that can be taken as starting point. The FAIR principles[73] framework, for instance, is the most renowned example of guidance for data practices, advocating that datasets should be "Findable, Accessible, Interoperable and Reusable" to enable data-led research and enterprise solutions. However, while FAIR provides an essential foundation for machine-actionable data management, the unique demands of sophisticated AI workflows necessitate a more granular and technically specific set of guidelines.Recent adaptations of FAIR (eg, FAIR-R[74] and the FAIR for AI[75] efforts) have begun to explore these limitations, highlighting the importance of structured labelling, provenance, and bias mitigation for AI purposes. However, these criteria often remain conceptual without concrete operational guidance.

In this section, we reflect on this need and provide a path to address it. First, we present the recently published **AI Readiness Framework by the Open Data Institute (ODI)**, a promising work that has been designed in this direction of targeting the practical aspects of a dataset's collection, preparation and publication to enhance its quality and utility for the AI ecosystem. Then, we propose to evolve this ODI framework to cover all the dimensions presented in the previous sections of this document. Finally, we outline three levels of readiness which, building upon the proposed framework, can serve as an initial basis for assessing the AI readiness of Data Products in the near future.

[73] https://www.go-fair.org/fair-principles/
[74] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5164337
[75] https://www.nature.com/articles/s41597-023-02298-6

5.1. ODI AI-readiness framework

This section presents a set of criteria for AI-readiness, relying on the framework for AI-ready data developed by the Open Data Institute (ODI)[76], which moves beyond high-level principles to provide actionable guidance for data publishers. The framework is used to define the practical characteristics of an AI-ready data product, structuring its recommendations across three core areas: (i) the intrinsic "Dataset Properties", (ii) the richness and utility of its associated "Metadata", and (iii) the robustness of the "Surrounding Infrastructure." This section offers a practical methodology for evaluating and enhancing the components of a data product to ensure they are optimised for AI applications.

The presented AI-ready data framework is positioned as a response to the unique demands of sophisticated AI workflows that necessitate a granular and technically specific set of guidelines, providing structured labelling, provenance, and bias mitigation for AI purposes. The framework addresses the limitations of high-level or domain-specific frameworks by providing some combination of holistic but actionable guidance that is specifically focused on the dataset, its metadata, and its supporting infrastructure. A summary of this framework is shown in Figure 3. The requirement for such a framework has emerged not merely as a technical preference but as a necessary response to a convergence of pressures. The technical demands of data-centric AI[77], which prioritise high-quality, iterative data management as a primary driver of model performance, have created a strong pull for more rigorous data practices.

Concurrently, the growth of regulatory requirements[78], like the EU AI Act[79], mandates auditable data governance (e.g., data provenance, lineage, etc.), establishing clear legal and ethical obligations. These technical and regulatory forces together expose the insufficiency of high-level principles alone, framing this AI-readiness framework as a crucial piece of governance infrastructure rather than a simple technical checklist.

[76] https://theodi.cdn.ngo/media/documents/A_framework_for_AI-ready_data.pdf
[77] https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence
[78] https://theodi.org/insights/reports/global-policy-observatory-for-data-centric-ai/
[79] https://artificialintelligenceact.eu/

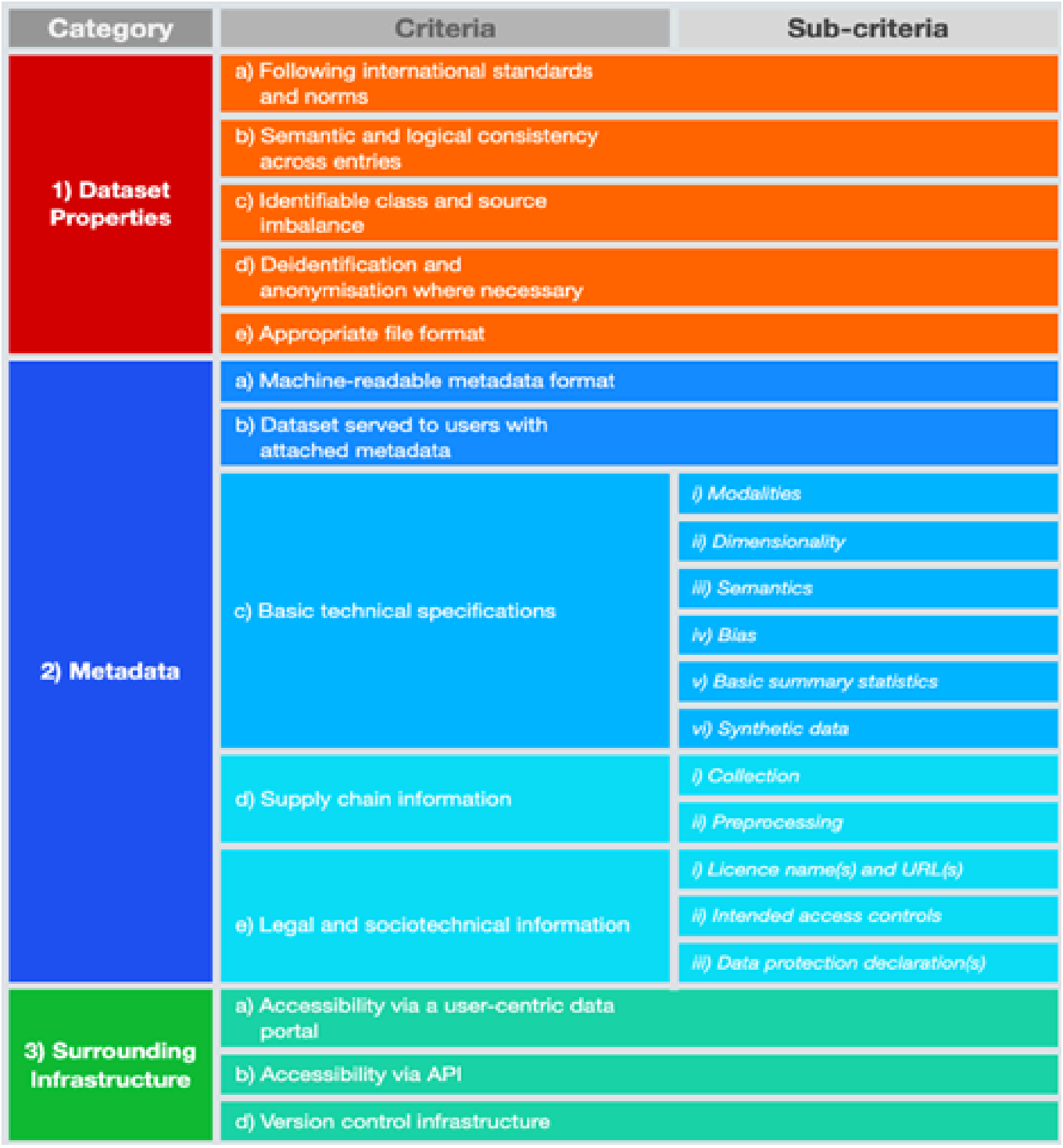

Figure 3. ODI AI-ready data framework

5.1.1. Dataset properties

According to this framework, these attributes determine a dataset's fundamental suitability for AI applications, focusing on its structure, content, and format:

- **Data values should follow established standards and conventions.** Adhering to recognised standards is essential for datasets to be consistent and interoperable, reducing the preprocessing overhead required by AI practitioners
- **Labels should be semantically and logically consistent.** To avoid ambiguity that can lead an AI model to learn fictional distinctions or suffer from endogeneity, a dataset publisher should ensure consistent and standardised data labels. In domain-specific cases, labels should adhere to internationally recognised vocabularies. This principle is core to the 'linked data' protocol, which ensures datasets are functional and shareable by conforming data labels to persistent uniform resource identifiers (URIs)
- **Class and source imbalance should be easy to identify.** A class imbalance in training data can severely skew AI model performance and introduce significant bias.While perfect balance is 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
- **Standard de-identification and anonymisation methods should be used where necessary.** Sensitive, personal information about individuals cannot be contained in shared datasets without appropriate steps to protect subjects' privacy, a consideration that must account for the greater 're-identification risk' where data is used to train AI
- **Datasets should be saved in appropriate file formats.** The choice of file format has direct implications for performance. While comma-separated value (.csv) files are widely used, enhancing them with metadata using the W3C's CSV on the Web (CSVW) standard can improve their interoperability. For large-scale AI training, formats like Apache Parquet are increasingly preferred due to their columnar storage, innate compression, and metadata handling capabilities, all of which are critical for AI model performance

5.1.2. Metadata

For AI-ready data products, metadata is not merely descriptive but operational; it provides the essential context for the accurate, responsible, and automated use of data in AI systems:

- **Metadata should be machine-readable.** To enable automated discovery, parsing, and integration into AI workflows, metadata must be stored and accessible in a structured, machine-readable format such as JSON-LD
- **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, but separating metadata from data can lead to the latter being disregarded, in turn leading to a subpar understanding of a dataset and its broader context (eg, collection practices).
- **Metadata should include basic technical specifications.** To accelerate exploratory data analysis, metadata should report on a dataset's modalities (such as text, image, video), dimensionality, semantics, potential biases, and basic summary statistics
- **Metadata should provide a holistic portrait of the entire dataset lifecycle.** To operationalise Responsible AI (RAI) principles and support regulatory demands for traceability, metadata must capture not only the methods of collection and preprocessing but also a broader range of supply chain information. Emerging standards like the Croissant-RAI extension[80] enable the documentation of data worker demographics and other social impact information, allowing practitioners to make more informed and responsible decisions when selecting datasets
- **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, alongside intended access controls and data protection declarations. This ensures AI practitioners have high legal confidence when using a dataset

The relationship between foundational principles, this operational framework, and regulatory compliance is made functional through technical metadata standards. High-level goals articulated in the aforementioned FAIR principles (such as providing rich metadata (F2), a clear usage license (R1.1), and detailed provenance (R1.2)) find their practical implementation in the specific criteria outlined here. Machine-readable metadata standards, particularly the emerging Croissant standard, provide the technical mechanism to enact these criteria by offering specific fields for ML semantics, provenance, and RAI information.

[80] https://docs.mlcommons.org/croissant/docs/croissant-rai-spec.html

This implementation is no longer optional but is driven by the regulatory imperative of legislation like the AI Act, which demands auditable records of data governance. In this way, metadata standards serve as the operational bridge, translating abstract principles into the concrete, auditable artefacts required for a compliant and trustworthy AI governance strategy.

5.1.3. Surrounding infrastructure

For a dataset to be AI-ready, it needs to be published within a surrounding infrastructure that is also AI-ready, supporting discovery, programmatic access, and robust versioning.

- **Datasets should be accessible via a user-centric data portal.** A sufficiently user-centric AI data portal facilitates not just discovery but also engagement, co-locating documentation and analytical tools to empower practitioners to assess and integrate datasets into their workflows.
- **Datasets accessible via AI-ready APIs as best practice.** Programmatic access is non-negotiable for modern AI development. An AI-ready API should employ aRESTful architecture, avoid artificial bottlenecks like pagination, which can hinder large-scale data ingestion, and facilitate the querying of subsets and splices of datasets.
- **Version control infrastructure as best practice.** A fit-for-purpose version control system is essential for reproducibility and for managing the dynamic nature of datasets in AI. Tools such as Data Version Control (DVC) [81] enable granular monitoring of data changes throughout the AI data lifecycle, ensuring transparent provenance tracking and facilitating the timely integration of new information for enhanced AI performance.

[81] https://dvc.org/

5.2. Evolution of an AI-readiness framework

The ODI readiness framework presented in the previous section provides a robust foundation for assessing how well data and metadata, and its associated infrastructure, are prepared to support AI applications. However, as discussed throughout this document, several dimensions of an AI-ready Data Product go beyond the current proposed framework.

Therefore, this section aims to explore those aspects that can complement the ODI framework to fully address the requirements of AI-ready Data Products. The following areas summarise how the ODI framework can be extended:

- **Data product perspective.** While the ODI framework addresses datasets, metadata, and infrastructure, it does not frame those elements together (with others) as a data product. To align with the data product perspective, it should be complemented with a product-oriented view that treats data as a packaged, shareable, and consumable asset, with considerations along all stages of the full lifecycle, from design and creation to delivery, maintenance, and retirement (as described in Section 2).
- **Dynamic data products.** The ODI framework implicitly assumes that data is a static asset to be published and accessed. However, AI-ready data products should work as dynamic, continuously evolving entities that stay aligned with AI workflows (as explained in Section 3). Therefore, the framework can be extended by incorporating mechanisms for continuous updates, version evolution, and alignment with AI workflow stages
- **AI as primary consumer.** The ODI framework assumes human users download and using datasets. In modern data-centric AI, and as explained in Section 3.4, the primary consumer may be AI models, AI agents, or automated pipelines. This would require the adaptation of traditional APIs and their combination with protocols like MCP
- **Iterative integration with AI.** The ODI framework assumes a one directional flow. AI workflows, however, operate through iterative cycles of training, evaluation, deployment, monitoring, and retraining. As discussed in Section 3.5, this would require closer integration with operational AI practices (MLOps).
- **Other AI specific aspects.** Although the ODI framework addresses generic data quality and provenance, it does not account for AI specific requirements.

5.3. Levels of readiness

Building upon the previous sections, we introduce a progressive maturity model that categorises data products across multiple levels of sophistication. This model moves from basic AI-compatible to fully autonomous AI-optimised data products, serving as a practical tool for providers to assess the current state of their products and to identify a clear pathway for enhancement.

Table 3 shows how each of the readiness levels responds to the needs identified in the previous section of the document. Notice that this section is just a starting point that only introduces the readiness dimensions and qualitative assessment. Specific metrics for quantitative evaluation are out of the scope of this document and remain as future work.

Table 3. Levels of AI readiness of a data product

Criterion	Foundational	Intermediate	Advanced
Data product concept and lifecycle	Basic elements to support AI applications Stages of product lifecycle not adapted to AI	Incorporates all necessary elements to support AI Stages partially adapted	The data product incorporates all elements to fully support AI. Fully AI-oriented lifecycle; stages optimised for AI workflows
Metadata	Basic AI-relevant quality dimensions are captured. Use metadata standards such as ML-DCAT-AP or Croissant Workflow and additional AI properties limited.	Richer machine-readable metadata captures additional workflow properties, provenance, quality metrics, and others	Comprehensive, machine-readable metadata covering bias, fairness, representativeness, … as well as workflow integration and provenance
Access and use	Standard APIs, manual download Human users as primary consumers AI pipelines can consume data, but without automated integration	AI models and pipelines can access data via standard APIs (REST, GraphQL), supporting some advanced features	AI pipelines and agents are primary consumers AI-native APIs (potentially using protocols like MCP) support dynamic interaction and automated integration
Dynamic integration with AI	The product is mostly static, with updates applied manually. No feedback from AI workflows. Data flow is one-directional.	Updates partially automated, enabling partial alignment with AI workflows. Limited feedback loops. Pipelines with minor adjustments to dataset versions	Continuously evolving data product with real-time (or near-real-time) updates Synchronised with AI workflows Full feedback loops with MLOps. Data updates or enrichment based on model performance.
Provenance	Basic tracking (e.g., W3C PROV-O)	Complemented with advanced features for AI, coming from Croissant WG or ML-DCAT-AP	Fully automated, auditable, and machine-readable lineage tracking throughout the lifecycle. Lineage tracking compatible with MLOps, ensuring data traceability across preprocessing, training, deployment, …
Licensing	Standard data licenses, with no AI-specific clauses	AI-specific clauses introduced	AI-ready licenses fully defining permissible AI uses, derivative outputs, and redistribution conditions
Ethical considerations	Not explicitly addressed	Basic ethical guidance or warnings embedded	Embedded, enforced, and monitored automatically; aligns with responsible AI principles
Compliance	Meets basic regulatory requirements (e.g. EU AI Act)	Meets AI Act and aligns with additional AI-relevant standards	Fully aligned with AI Act, ethical guidelines, and other applicable regulatory frameworks

VI. Conclusions, Future Work and Next Steps

The present paper builds on past and ongoing discussions and activities within the BDVA community around the paradigm of "data for AI" and how this can be embedded in the data product approach. To the best of our knowledge, this is the first attempt to systematically evolve the concept of a data product so that it accommodates the needs, constraints, and practices of AI systems and AI practitioners.

Throughout the preparation and writing process, it became clear the complexity of bringing together aspects of data management and data sharing (widely covered in data spaces and data ecosystems), the data product paradigm (emerging from data mesh and modern data architectures) and the AI dimension. Bringing these perspectives together would result in the great potential of the so-called AI-ready Data Products, but at the same time it shows the complexity of defining them in a comprehensive and actionable manner.

For this reason, this paper should be seen as a starting point rather than a definitive document. The topic requires further exploration, knowledge exchange, co-creation, and validation across disciplines and stakeholders. Therefore, we plan to continue this work within BDVA community, but also to extend the dialogue to external stakeholders, other associations, standardisation bodies, policymakers, industry actors and AI communities. Future work should focus on the following key directions:

- **Standardised concept of AI-ready data product**, moving towards having a shared understanding of what an "AI-ready Data Product". This involves developing a normalised set of fundamental components and clearly defined lifecycle stages that are unique to AI workflows and their evolutionary process. A shared standard will provide consistency and clarity when different organisations are sharing or designing AI-ready data products.

- **Assessment by AI practitioners**, including the evaluation of the concept, how the new aspects considered are relevant and what is possibly missing. Involve AI practitioners (e.g., data scientists and ML engineers) in assessing the proposed AI-ready data product concept. Their real-world assessment is critical in determining whether the newly introduced aspects are relevant and useful in real-world AI development contexts. Feedback from these practitioners will tell us what is working, what needs to be changed, and whether anything important is missing from the current concept.
- **Integration with data spaces and other data ecosystems**, since the data product concept relies at the core of their offering, and assessment on how the AI dimension fits in their design, architecture, governance and business. Investigate how the notion of AI-ready data products can be embedded within existing data-sharing infrastructures. Most data spaces and data ecosystems already rely on the notion of data products, so there is a need to identify how AI-specific requirements would be met within their design, architecture, governance, and business models. Additionally, extending the data product to the supply chain by crossing organisational boundaries entails the removal of some labels and features, in order to hide proprietary or private information, where data spaces can also support AI-ready Data Products. All this will ensure that AI-ready Data Products can thrive within broader data ecosystems without disrupting existing practices.
- **Evolution of existing AI readiness frameworks** (ODI, FAIR, …), to align with the goals described in this paper. This includes the definition of specific metrics for the readiness of the different features in the three levels of the proposed framework. Map and extend current data/AI readiness frameworks to cover the AI-ready data product perspective. For example, the Open Data Institute's AI-ready data framework and the FAIR data principles need to be remapped and adapted to include the goals and characteristics outlined in this paper. Updating these widely adopted frameworks will bridge the gap between high-level data readiness guidelines and the specific needs of AI-ready data products
- **Validation in industry sectors**, to test the concept with industry stakeholders in those sectors where the concept of data product is present (e.g., manufacturing) to assess feasibility, scalability and eventually business value. Collaborate with industry partners to deploy and test AI-ready data products to ascertain their practical feasibility and scalability. Validation in real-world environments is required so as to determine the practical business value of AI-ready data products as well as to advance the concept based on lessons learned in operational environments.

- **Guidance and recommendations to AI-ready Data Products owners and providers**, including clear advice for organisations on how to best implement an AI-ready data product production flow, with a discussion on how different operational solutions might be relevant for different types of organisations (large vs SMEs; tech start-ups vs traditional businesses etc). This would include investigation into existing tools for Data Products and benchmarking study to define whether they are valid for AI or should be expanded.
- **Exploration of collaborative business models around AI-ready Data Products**: investigate how AI-ready Data Products can enable and support collective business models between and within industries and organisations. This involves establishing incentive regimes, value-sharing mechanisms, and governance arrangements that favour data and AI asset co-creation, particularly in data spaces. Establishing sustainable, trust-based mechanisms for collective data stewardship and AI development will be required to ensure AI-ready data products generate collective economic and societal value.

In summary, the convergence of data management best practices, modern data product principles, and AI-specific requirements is highly promising to enable "AI-ready" Data Products. This is, nevertheless, an emerging area that requires additional research and collaboration. The BDVA community intends to transform the initial ideas of this paper into concrete standards, tools, and methodologies that make AI-ready Data Products a reality in various industries and ecosystems.

VII. List of relevant standards and specifications

Listed according to the order of sections in the document they relate to:

- **Open Data Product Specification** (https://opendataproducts.org/v4.0/#open-data-product-specification) provides a structured approach to creating, managing, and utilising data products
- **ISO/IEC 5259 "Data quality for analytics and machine learning (ML)"** series https://www.iso.org/standard/81088.html, including Part 1: Overview, terminology, and examples, Part 2: Data quality measures, Part 3: Data quality management requirements and guidelines, Part 4: Data quality process framework and Part 5: Data quality governance framework
- **MLDCAT-AP**, https://semiceu.github.io/MLDCAT-AP/releases/2.0.0/, is an application profile that extends DCAT-AP in the field of machine learning. It aims to describe machine learning models, together with their datasets, quality measured on the datasets and citing papers
- **Croissant**, https://mlcommons.org/working-groups/data/croissant/, standardises how ML datasets are described to make them easily discoverable and usable across tools and platforms
- **RFC-WAI0-002-R1 — AI Discovery Metadata Standard**, https://www.witchbornsystems.org/rfc/RFC-WAI0-002-R1_ai.pdf, introduces machine-readable metadata for AI endpoints, including provenance, licensing, and audit policies.
- **Model Context Protocol (MCP)**, https://modelcontextprotocol.io/docs/getting-started/intro, is an open-source standard for connecting AI applications to external systems, including data sources
- **Agent2Agent Protocol (A2A)**, https://a2a-protocol.org/, a standard for communication and collaboration between AI agents, supporting distributed agentic AI systems.
- **IEEE Synthetic Data Industry Connections Activity**, https://standards.ieee.org/industry-connections/activities/synthetic-data/, an initiative to develop privacy and accuracy standards for synthetic data.
- **Great Expectations Framework**, https://towardsdatascience.com/how-to-validate-the-quality-of-your-synthetic-data-34503eba6da/, an open standard for data quality validation, used in combination with synthetic data generation tools.

- **ISO/IEC 42001 "AI management systems"**, https://www.iso.org/standard/42001, requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organisations.
- **ISO/IEC 23894 (AI / Guidance on risk management)**, https://www.iso.org/standard/77304.html, offers comprehensive guidance on risk management specific to AI.
- **AI Risk Management Framework (AI RMF 1.0)**, https://www.nist.gov/artificial-intelligence/ai-standards, developed by NIST to guide trustworthy and responsible AI development and use
- **ISO/IEC 5338 (AI system life cycle processes)**, https://www.iso.org/standard/81118.html, outlines the development, deployment, monitoring, and retirement phases of AI systems, provides additional support for AI data governance
- **ISO/IEC 42005 (AI system impact assessment)**, https://www.iso.org/standard/42005, provides guidance for organisations conducting AI system impact assessments.
- **AI Auditing Framework by The IIA**, https://www.theiia.org/globalassets/site/content/tools/professional/aiframework-sept-2024-update.pdf, provides guidance for internal auditors to assess AI systems across governance, risk, and compliance dimensions
- **UNESCO Recommendation on the Ethics of Artificial Intelligence**, https://www.unesco.org/en/artificial-intelligence/recommendation-ethics, the first global standard on AI ethics adopted by 194 member states

VIII.

BIBLIOGRAPHY

[1] Aroyo, L., Lease, M., Paritosh, P., & Schaekermann, M. (2022). Data excellence for ai: why should you care? Interactions, 29(2), 66-69.
[2] Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. Communications of the ACM, 41(2), 79-82.
[3] Engemann, K. (2014). Measuring data quality for ongoing improvement: a data quality assessment framework. Benchmarking: An International Journal, 21(3), 481-482.
[4] Ehrlinger, L., Haunschmid, V., Palazzini, D., & Lettner, C. (2019). A DaQL to monitor data quality in machine learning applications. In Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30 (pp. 227-237). Springer International Publishing.
[5] Jarrahi, M. H., Memariani, A., & Guha, S. (2023). The principles of data-centric AI. Communications of the ACM, 66(8), 84-92.
[6] Amiriebrahimabadi, M., Mansouri, N. A comprehensive survey of feature selection techniques based on whale optimisation algorithm. Multimed Tools Appl 83, 47775–47846 (2024). https://doi.org/10.1007/s11042-023-17329-y
[7] Özcan, F., Lei, C., Quamar, A., & Efthymiou, V. (2021). Semantic Enrichment of Data for AI Applications. In International Workshop on Data Management for End-to-End Machine Learning (DEEM'21). ACM. https://doi.org/10.1145/3462462.3468881
[8] Mumuni, A., Mumuni, F., & Mumuni, A. (2022). Data augmentation: A comprehensive survey of modern methods. Information Fusion, 87, 36-58] https://doi.org/10.1016/j.array.2022.100258
[9] Bao, E., Xiao, X., Zhao, J., Zhang, X., & Ding, B. (2021). Synthetic Data Generation with Differential Privacy via Bayesian Networks (PrivBayes). Journal of Privacy and Confidentiality, 11(3] https://doi.org/10.29012/jpc.776
[10] Big Data Value Association. (2025). Synthetic data in healthcare: Benefits and opportunities, technological, clinical, regulatory gaps & challenges, ways ahead for impact maximisation. Big Data Value Association. https://bdva.eu/news/synthetic-data-in-healthcare/
[11] https://dl.acm.org/doi/10.1145/3458723
[13] Kundavaram, V. N. K. (2025). Optimising Data Pipelines for Generative AI Workflows: Challenges and Best Practices. IJSAT-International Journal on Science and Technology, 16(1). https://doi.org/10.71097/IJSAT.v16.i1.1527

[14] Chrysakis, I., Agorogiannis, E., Tsampanaki, N., Vourtzoumis, M., Chondrodima, E., Theodoridis, Y., ... & Doulkeridis, C. (2025, March). Multi-Partner Project: Green. Dat. AI: A Data Spaces Architecture for Enhancing Green AI Services. In 2025 Design, Automation & Test in Europe Conference (DATE) (pp. 1-7). IEEE. https://doi.org/DATE64628.2025.10992729
[15] Confalonieri R, Kutz O, Calvanese D, et al. Data journeys: Explaining AI workflows through abstraction. Semantic Web. 2023;15(4):1057-1083. doi:10.3233/SW-233407
[16] D. Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute, 2019.
[17] S. Thiebes, S. Lins, and A. Sunyaev. Trustworthy artificial intelligence. Electronic Markets, 31:447–464, 2021. https://doi.org/10.1007/s12525-020- 00441-4.
[18] E. Kazim and A. S. Koshiyama. A high-level overview of AI ethics. Patterns, 2(9), 2021. https://doi.org/10.1016/j.patter.2021.100314.
[19] Jantunen, M. (2025). Integration of Artificial Intelligence Ethics into Software Engineering Processes: Challenges, Concerns, and Opportunities
[20] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. SIGKDD
[21] The Evolving Role of Copyright Law in the Age of AI-Generated Works, J. Hutson, https://doi.org/10.21202/jdtl.2024.43
[22] Mammen, Christian and Collyer, Michael and Dolin, Ron A. and Gangjee, Dev S. and Melham, Tom and Mustaklem, Maggie and Sundaralingam, Pireeni and Wang, Vincent, Creativity, Artificial Intelligence, and the Requirement of Human Authors and Inventors in Copyright and Patent Law (July 05, 2024). Available at SSRN: https://ssrn.com/abstract=4892973 or http://dx.doi.org/10.2139/ssrn.4892973
[23] Longpre, S., Mahari, R., Chen, A. et al. A large-scale audit of dataset licensing and attribution in AI. Nat Mach Intell 6, 975–987 (2024). https://doi.org/10.1038/s42256-024-00878-8
[24] Abel Goedegebuure, Indika Kumara, Stefan Driessen, Willem-Jan Van Den Heuvel, Geert Monsieur, Damian Andrew Tamburri, and Dario Di Nucci. 2024. Data Mesh: A Systematic Gray Literature Review. ACM Comput. Surv. 57, 1, Article 11 (January 2025), 36 pages. https://doi.org/10.1145/3687301

BDV BIG DATA VALUE ASSOCIATION

BDVA is an industry-driven international not-for-profit organisation with 250 members all over Europe and a well-balanced composition of large, small and medium-sized industries, start-ups as well as research and user organisations. Our mission and objectives are:

- To boost Data and AI research, development and innovation for European competitiveness, societal wellbeing and sustainable progress
- To develop the innovation ecosystem that enables and accelerates the data-driven and AI-enabled digital transformation of the economy and society, with European values and focus but global impact and ambitions.
- To foster excellence in European Data and AI research, in science and business.
- To anticipate, lead and keep up with the dynamic change that Data and AI brings to business and society.

BDVA enables existing regional multi-partner cooperation, to collaborate at the European level through the provision of tools and know-how to support the cocreation, development and experimentation of pan-European data-driven and AI applications and services and know-how exchange. Through BDVA, its Task Forces and labelled hubs (i-Spaces), our members build new collaborations, co-create new projects, share knowledge and jointly develop guidelines, frameworks and strategic roadmaps for industry and policymakers.
Together with our members and our collaboration partners, we advance all related areas connected to the data economy such as data spaces, data privacy, industrial and ethical AI, generative AI, business models, standardisation, skills, computing and many others. BDVA is contributing to all these discussions, having significant impact, developing relevant collaborations and with a very well-established community of members that are at the core of the European data and AI ecosystems!
BDVA is a private member of the EuroHPC Joint Undertaking and it is a founder member of the AI, Data and Robotics Partnership. BDVA has developed a strong and growing cooperation with Gaia-X, IDSA and FIWARE through the Data Spaces Business Alliance (DSBA) and collaborates with many data and industry-driven AI national initiatives and other European communities.
In October 2024 the BDVA community celebrates the Association's 10th anniversary, that keeps on growing in breadth and in depth thanks to our members, our team and our collaboration partners. Join us in the celebration!

BDVA is open to new members! Visit BDVA.EU to learn more about members and activities. You can contact us anytime at info@bdva.eu.

BDV BIG DATA VALUE ASSOCIATION

BDVA
Data, AI and Robotics (DAIRO) aisbl
Avenue des Arts, 56
1000 Bruxelles
Belgium

BDVA.eu
info@bdva.eu