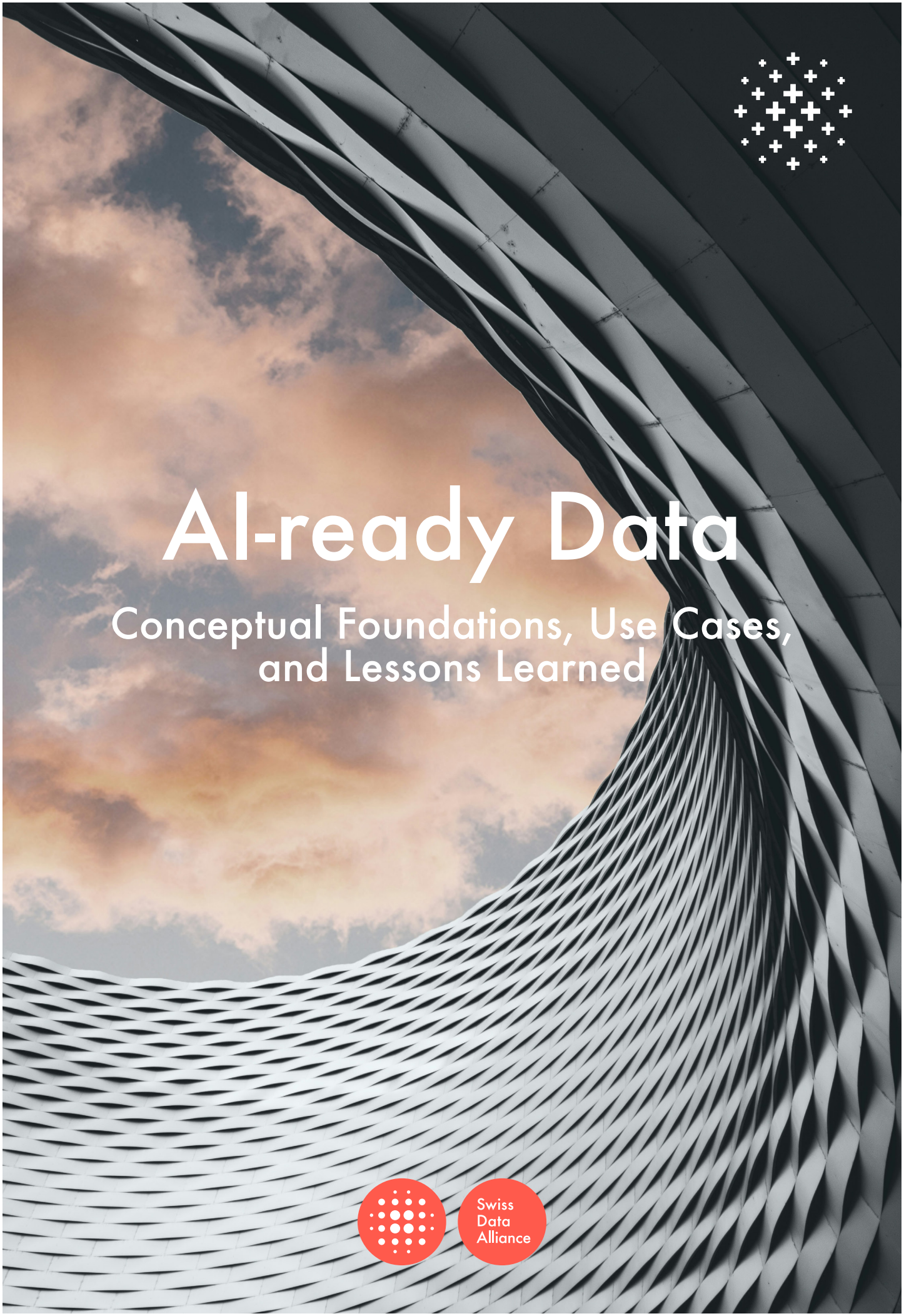

AI-ready Data
Conceptual Foundations, Use Cases,
and Lessons Learned
Swiss
Data
Alliance

White Paper



AI-ready Data: Conceptual Foundations, Use Cases, and Lessons Learned

AI Action Plan for Switzerland

CONTEXT

This white paper has been developed in the context of the AI Action Plan for Switzerland (AI-ActionPlan.ch) by the Working Group “AI-ready Data”.

AUTHORS AND CONTRIBUTORS

Christine Legner, Faculty of Business and Economics (HEC), University of Lausanne
Diego Kuonen, Statoo Consulting and Geneva School of Economics and Management (GSEM), University of Geneva
Jingyang Wang, Faculty of Business and Economics (HEC), University of Lausanne
André Golliez, Swiss Data Alliance

More than 80 members of the Working Group “AI-ready Data”, representing business, research, education and the public sector, have participated in workshops and contributed use cases and expert feedback to this white paper.

SUGGESTED CITATION

Legner, C., Kuonen, D., Wang, J. & Golliez, A. (2026). AI-ready Data: Conceptual Foundations, Use Cases, and Lessons Learned. White Paper of the Working Group “AI-ready Data” – AI Action Plan for Switzerland, Swiss Data Alliance (https://iris.unil.ch/handle/iris/285961).

Version 1.0 as of October 26, 2026.

Table of Contents

1.	Introduction	4
2.	Foundations: Why are Data Critical for AI?	6
3.	Key Concepts: What are AI-ready Data?	10
4.	Use Cases: What are Practical Challenges and how are Data Prepared for AI Systems?	15
5.	Learnings and Recommendations	36
6.	Next Steps on “AI-ready Data” in the AI Action Plan for Switzerland	41
Appendix: AI-ready Data – Literature Review		43

1. Introduction

Artificial Intelligence (AI) is rapidly transforming how organisations across business, research, education and the public sector operate, innovate and create value. Yet, the effective and trustworthy use of AI does not depend on algorithms, models and computing infrastructure alone. **AI cannot operate without data, and AI-ready data are an important foundation for trustworthy AI systems**. They help reduce risks associated with unreliable, biased or inappropriate data and improve the quality and reliability of AI outputs. Documented data quality, provenance, meaning and permitted uses also **help building trust** in AI systems among users and other stakeholders.

The importance of data is also reflected in the **practical challenges of AI projects**: more than 80% of the time spent on an AI project may be devoted to data[1], while 72% of organisations report data management as a key challenge in scaling AI[2]. Moreover, even among high-performing organisations, 70% report difficulties in developing processes for data governance and integrating data into AI models quickly.[3] The availability of data that are fit for their intended use is therefore fundamental for **developing, operating and improving AI systems** and for **helping build trust** in their use.

Against this background, the **AI Action Plan for Switzerland**[4] brings together concrete actions aimed at strengthening Switzerland's readiness for AI across five key areas. One of these is **AI-ready Data**[5], with the overarching objective of "*ensuring the availability of high-quality, appropriately documented and augmented datasets for training or running AI systems and strengthening their provision and use across business, research, education, and the public sector*".

An important action within this area is to "define and provide an assessment framework for AI-ready data". The action seeks to create a shared understanding of AI-ready data, specify how datasets can be assessed with regard to their "AI-readiness", and provide guidance for their provision and consumption. **This white paper provides the conceptual foundation** for this work.

Creating a common language can be seen as the heart of this white paper. A shared understanding of terms and a resulting common language enable the active exchange of experience and knowledge and simplify communication about ideas, projects and services. This is particularly important in the context of AI-ready data, where different communities – including data providers, data consumers, AI practitioners, organisations and policymakers – may approach the topic from different perspectives and with different requirements.

Building on this common language, the purpose of this white paper is to establish the rationale for and a **common understanding of AI-ready data, while illustrating the practical challenges and approaches** involved in making data AI-ready**.** It is structured in four parts:

1. Foundations: Why are data critical for AI?
2. Key concepts: What are AI-ready data?
3. Use cases: What are practical challenges and how are data prepared for AI systems?
4. Learning and recommendations: What are the learnings from the use cases and further actions?

To address these questions, the white paper draws on an extensive review of more than 100 sources, covering practitioner reports as well as academic publications, to derive a definition and key characteristics

[1] Davenport, T. H., Hoerl, R. W. & Redman, T. C. (2025). To Create Value with AI, Improve the Quality of Your Unstructured Data, Harvard Business Review (https://hbr.org/2025/05/to-create-value-with-ai-improve-the-quality-of-your-unstructured-data).

[2] Gartner (2025). Lack of AI-ready Data Puts AI Projects at Risk (https://www.gartner.com/en/newsroom/press-releases/2025-02-26-lack-of-ai-ready-data-puts-ai-projects-at-risk).

[3] Tavakoli, A., Harreis, H., Rowshankish, K. & Bogobowicz, M. (2024). Charting a path to the data- and AI-driven enterprise of 2030, McKinsey Quarterly (https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/charting-a-path-to-the-data-and-ai-driven-enterprise-of-2030).

[4] "AI Action Plan for Switzerland: Shaping our AI future", coordinated by digitalswitzerland in close collaboration with the Federal Office of Communications (https://ai-actionplan.ch/), for key areas and concrete actions see https://ai-actionplan.ch/topics.

[5] "AI-ready Data" (https://ai-actionplan.ch/topics/ai-ready-data).

of AI-ready data. This framework is then used to assess AI-readiness of data in nine productive use cases and provides insights into the underlying data work to prepare AI-ready data. By addressing the four questions, the white paper provides the **basis for subsequent work** on assessment approaches, practical guidance and the responsible provision and use of AI-ready data across Switzerland.

2. Foundations: Why are Data Critical for AI?

2.1. AI SYSTEMS AND CHARACTERISTICS

To clarify what it means for data to be "AI-ready", it is important to understand AI systems and their specific characteristics. Drawing on the terminology adopted by the Swiss Federal Administration, which in turn is based on that of the Organisation for Economic Co-operation and Development (OECD), an **AI system** is defined as:

> "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."[6]

While much of the current public debate focuses on generative or agentic AI[7], AI systems encompass a much broader range of applications. Examples include:

- a fraud detection system that uses historical transaction data to predict whether a new transaction is potentially fraudulent;
- a credit risk assessment system that estimates the probability of default or supports credit-related decisions;
- a predictive maintenance system that uses sensor and operational data to predict equipment failures or maintenance needs;
- a recommendation system that suggests products, services or content based on available data and observed patterns;
- an image recognition system that identifies or classifies objects, persons or features in images; and
- a large language model-based assistant[8], that generates new content such as text, images, audio, video or software code in response to user input;
- an open-source AI assistant that uses verified public data, laws and standards to provide citizens with trustworthy answers and guidance, with every answer traceable to authoritative sources.

These examples illustrate that AI systems can differ considerably in their purpose, underlying models, level of autonomy and degree of adaptiveness. For the purposes of this white paper, we distinguish broadly between traditional AI and generative AI systems:

- **Traditional AI systems** are designed to perform specific tasks based on predefined rules or patterns learned from data. Depending on the application, their outputs typically take the form of predictions, classifications, scores, recommendations or decisions rather than newly generated content. Examples include fraud detection, credit risk assessment, predictive maintenance, recommendation and image recognition systems.
- **Generative AI systems**, by contrast, are designed to generate new content, such as text, images, audio, video or software code, based on patterns and representations learned from training data. They can typically perform a broader range of tasks and increasingly support multimodal inputs and outputs. Large Language Model (LLM)-based assistants are a prominent example.

[6] OECD (https://oecd.ai/en/ai-principles), Terminology AI from the Confederation's Competence Network for AI (https://www.bk.admin.ch/en/terminology-ai) & Strategy Use of AI systems in the Federal Administration (https://www.bk.admin.ch/en/sb021-strategy-use-of-ai-systems-in-the-federal-administration).

[7] "Generative AI" is a broad term that refers to AI systems that are trained on large amounts of data from the physical and virtual world in order to generate data themselves (e.g., texts, imagery, sound recordings, videos, simulations, and codes). They are often multimodal, e.g. with input and/or output in one or several modalities (e.g., text, image or video). Source: Terminology AI from the Confederation's Competence Network for AI (https://www.bk.admin.ch/en/terminology-ai).

[8] Large language model-based generative AI systems have gained exceptionally rapid and widespread adoption among end users, particularly since the public release of ChatGPT in November 2022. As a result, generative AI and Large Language Models (LLMs) have become a major focus of public attention and are often closely associated with the broader concept of AI. However, as the examples above illustrate, AI encompasses a much broader range of systems and applications.

This distinction is not absolute. In particular, **foundation models** – models trained on broad data at scale and adaptable to a wide range of downstream tasks – can support both generative and non-generative (i.e. traditional) applications. Generative AI systems can also perform tasks associated with traditional AI, such as classification, prediction, information extraction or recommendation. This distinction therefore reflects the characteristic capabilities and outputs of these systems rather than mutually exclusive types of AI.

Generative AI systems can be extended with additional capabilities. **Retrieval-Augmented Generation** (**RAG**) combines a generative AI model with the retrieval of information from external knowledge sources, such as document repositories, databases or other information systems. The retrieved information is provided as additional context to the model when generating an output, helping to ground the output in relevant and potentially more up-to-date information.

AI systems can also incorporate agentic capabilities. **AI agents** enable an AI system to pursue an objective through a sequence of actions, potentially with a degree of autonomy. An agent may interact with models, tools, data sources, applications or other systems to gather information, make intermediate decisions and perform actions. Agents may also use RAG as one of their capabilities. Multiple AI agents may also interact or collaborate within a **multi-agent system**[9], with individual agents performing different tasks or roles in pursuit of a shared or related objective.

RAG and agents represent architectural patterns and capabilities that can be incorporated into AI systems. As such, they need to be considered throughout the AI system lifecycle, as discussed in the next section.

2.2. AI SYSTEM LIFECYCLE

The development and use of an AI system can be broadly structured into two phases: **AI model development** (a) – the "**build**" phase – and **deployment and operation** (b) – the "**use**" phase:

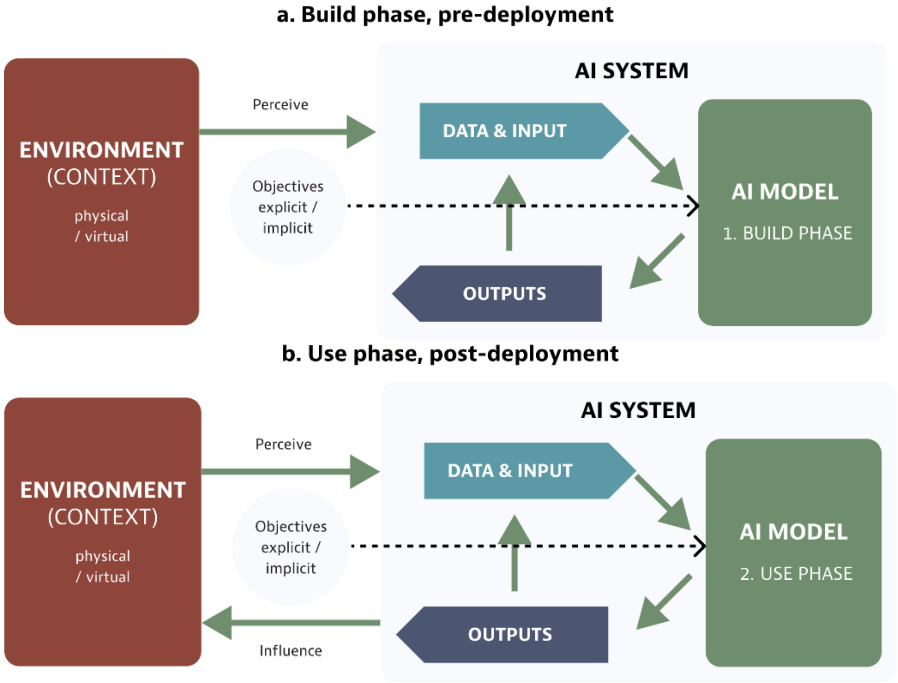


Figure 1: Build and use phases of an AI system[10]

[9] A Multi-Agent System (MAS) is an AI system composed of multiple independent but interacting agents, each capable of perceiving its environment and taking actions. Agents may be software-based or embodied in physical systems, such as robots, and may incorporate AI models and other computational components. Multiple agents can interact or collaborate towards a common or related objective, potentially enabling capabilities, adaptability and robustness beyond those of an individual agent.

[10] OECD (https://oecd.ai/en/ai-principles).

This simplified distinction serves to highlight the different roles and requirements of data during the development and use of AI systems: During the "build" phase, data serve as the basis for training the AI model, whereas during the "use" phase, they may be used to adapt or customise existing AI models for specific applications, and as inputs processed by AI systems. The "build" and "use" phases should, however, be understood within the broader context of an AI system lifecycle. The OECD distinguishes the following AI system lifecycle phases:

> "i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase."[11]

The incorporation of RAG, agentic capabilities or multi-agent architectures introduces additional considerations across these lifecycle phases. During "design, data and models", these include, for example, the selection of knowledge sources, models, tools and permitted actions; during "verification and validation", the resulting system behaviour and interactions need to be tested; during "deployment", appropriate access rights, safeguards and controls need to be established; and during "operation and monitoring", outputs, actions, dependencies and underlying information sources need to be monitored and maintained. For multi-agent systems, these considerations also extend to interactions and dependencies between agents. As an AI system's ability to retrieve external information, interact with other systems and autonomously perform actions increases, appropriate lifecycle management, governance and oversight become increasingly important. At the same time, these capabilities expand the ways in which AI systems depend on, access, process and generate data throughout their lifecycle. This makes the role of data a central consideration in the development and use of AI systems.

2.3. THE ROLE OF DATA

As AI systems are **learning-based systems**, they **cannot operate without data**[12]. As noted above, data play different roles throughout the AI system lifecycle. Data are used to train AI models during the "build" phase, to adapt or customise existing AI models for specific applications, and as inputs processed by AI systems during the "use" phase. Data are also generated as outputs of AI systems or through interactions with them. Such data can, where appropriate, subsequently be used to evaluate, improve, adapt or further train AI models and systems.

The following figure from the Australian Government Productivity Commission[13] illustrates how data can serve as both inputs to and outputs of a (generative) AI system:

[11] OECD (https://oecd.ai/en/ai-principles) & Strategy Use of AI systems in the Federal Administration (https://www.bk.admin.ch/en/sb021-strategy-use-of-ai-systems-in-the-federal-administration).

[12] The Swiss Data Literacy Charter, published in 2024 by the Swiss Academies of Arts and Sciences, defines data as "analogue and digital information units that exist in various formats as numbers, texts, images, videos or audios"; see Swiss Academies of Arts and Sciences (2024). Swiss Data Literacy Charter (https://swiss-academies.ch/publications/data-literacy-charta-schweiz & https://zenodo.org/records/11146023).

[13] Australian Government Productivity Commission (2024). Making the Most of the AI Opportunity – Research Paper 3: AI Raises the Stakes for Data Policy (https://apo.org.au/sites/default/files/resource-files/2024-02/apo-nid325426_1.pdf).

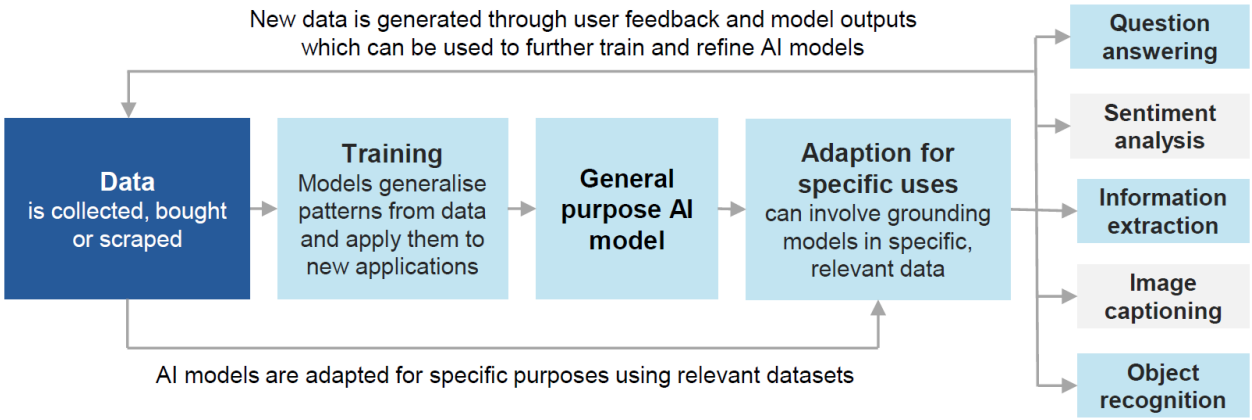


Figure 2: Data as input and output of a (generative) AI system

At the technical level, two fundamental processes help explain the role of data in AI: "training" and "inference". "**Training**", which takes place during the "build" phase, is the process through which an AI model learns patterns and relationships from training data and encodes what it learns in its parameters. "**Inference**", which takes place during the "use" phase, is the process through which a trained AI model processes new input data to generate an output, such as a prediction, classification, recommendation or generated content.

In RAG-enabled and agentic AI systems, data may additionally be retrieved dynamically from external sources during the "use" phase, rather than being incorporated into the model during training. Such external sources may include document repositories, databases or other information systems, as described in the previous section. To broaden the applicability of AI use cases, data that are currently available only in physical form should also be considered as sources of context and inference and transformed (digitalised) into AI-ready datasets.

Data used by AI systems can also differ substantially in their **degree of structure**[14]. **"Structured data**" are organised according to a predefined schema, typically represented in tables with rows and columns, such as in relational databases. Common examples include credit card numbers, financial transaction logs, and inventory records. "**Semi-structured data**" do not follow a rigid tabular structure but contain organisational elements such as tags, keys or metadata, as in JSON or XML files, or semantically structured content formats such as the Swiss open-source standard bitmark. "**Unstructured data**" do not conform to a predefined data model and include, for example, text documents, PDFs, images, audio and video. These distinctions are relevant for AI because different types and forms of data may require different approaches to their preparation, processing, integration, documentation and governance. In practice, the boundaries between these categories are not always clear-cut and are increasingly blurred by modern data architectures and AI systems.

Importantly, the value of data for AI does not depend on data alone, but also on the **context** in which they were generated, organised, interpreted and are intended to be used. Context provides relationships and meaning that may not be apparent from individual data points alone.[15] Consequently, making data available to an AI system is not sufficient: data need to be understandable, appropriately contextualised and fit for the intended purpose. This provides the basis for defining what constitutes "**AI-ready data**" in the next section.

[14] DAMA UK (2026). Structured, Unstructured, and Semi-Structured Data, LinkedIn (https://www.linkedin.com/pulse/structured-unstructured-semi-structured-data-dama-uk-ltd-mpvae/).

[15] Verhulst, S. G. & Singh, R. (2026). The Context Loop: How AI Remembers Us, and Shapes Digital Self-Determination (https://sverhulst.medium.com/the-context-loop-04d473545909). See in particular the discussion of context as the relational structure that gives information meaning.

3. Key Concepts: What are AI-ready Data?

3.1. OVERVIEW

Despite the widespread call for "AI-ready data", the term remains ambiguous and is a topic of intense discussion among practitioners and academics. A large number of definitions for AI-ready data have been proposed by analysts (e.g., Gartner), technology vendors and consultancies (e.g., Microsoft, McKinsey, Accenture), as well as research institutes (e.g., Open Data Institute). However, there is no convergence towards a common definition (see Appendix).

To address this gap, research activities in the Competence Center Corporate Data Quality (CC CDQ) reviewed more than 100 sources, covering academic and practioner publications. It was found that the term "AI-ready data" is often conflated with high-quality or well-managed data or associated with specific enablers such as metadata or data governance. Based on this analysis, it is therefore important to distinguish between the **characteristics of AI-ready data** and the **practices used to make data AI-ready**. This white paper will focus on the former by providing a definition and dimensions of AI-ready data that can guide the assessment of AI-ready datasets.

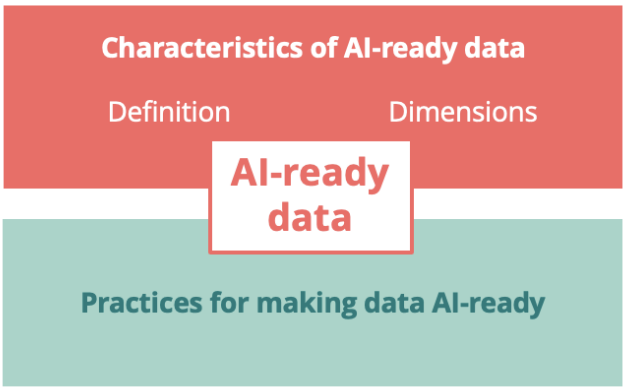


Figure 3: AI-ready data

3.2. DEFINITION

From the comprehensive review of more than 100 academic and practitioner sources, the CC CDQ has derived the following definition of AI-ready data and identified six dimensions that characterise data's AI-readiness.[16] [17]

Definition:[18]

AI-ready data refer to **datasets** that have been **documented, refined and extended** to **train or run AI system(s)**, designed for specific **intended purpose(s).**

AI-readiness is therefore not an inherent property of a dataset, but rather a state that can be assessed in relation to its intended use. It needs to be established through appropriate preparation and assessed in relation to the specific AI system, the relevant AI system lifecycle phase(s) and intended system purpose. For example, data requirements for training a predictive model differ from those for grounding an LLM through RAG or providing an agent with operational context.

[16] Wang, J. & Legner, C. (2025). What makes data AI-ready? European Conference on Information Systems (ECIS 2025) (https://aisel.aisnet.org/treos_ecis2025/39).

[17] Wang, J., Legner, C. & Lehmann, R. J. (2026). "AI-ready Data", Competence Center Corporate Data Quality (CC CDQ) research briefing (https://www.cdq.com/events-insights/publications/cc-cdq-research-briefing-ai-ready-data).

[18] Competence Center Corporate Data Quality (CC CDQ) research (https://cc-wiki.cdq.com/Glossary#AI-ready_data).

AI-ready data provide benefits in different ways. First, the **performance of an AI system** throughout its lifecycle can be improved when data are specifically prepared. As the model learns from the patterns represented in its training data, relevant, accurate, representative, and appropriately prepared data can improve its predictive accuracy. During inference, data can affect how that model performs on a particular input or in an operational context. Second, AI-ready data are an important **prerequisite for trustworthy, responsible and compliant AI use**. They reduce potential biases and limitations and ensure accordance with applicable legal, ethical and organisational requirements. Lastly, AI-data data also impacts the **efficiency of the AI system**.

3.3. EXAMPLE

The example of a pharmaceutical company that implemented an AI-enabled Human Resource (HR) platform offering personalized employee recommendations illustrates what it takes to make data AI-ready. The platform was designed to support two user groups: For employees, it provides personalized career and development opportunities and mentor pairing. For HR offices, it helps identify internal candidates for open positions and facilitating more effective deployment of skills across the organisation.

As the company had been actively driving data and AI initiatives for more than five years, it already had a "data product"[19] for employee data in place. For the new recommendation system, however, additional data from the learning management system were required to capture employees' acquired skills and completed training activities.

The two data products therefore had to be merged into a **combined dataset** that could be processed by the AI system. Making this dataset **AI-ready** required several additional preparation steps. First, the data had to comply with applicable regulatory and organisational requirements, in the case of a company that is active on different European markets, these include local as well as European data protection and AI regulations (i.e. GDPR and the EU AI Act), and internal requirements agreed with the workers' council. Second, data quality checks had to be performed to ensure that the dataset was sufficiently complete, accurate, and consistent for the intended recommendation use cases. Finally, the dataset and its relevant metadata were documented in the data catalogue before the data could be published and made available to the AI-enabled HR platform.

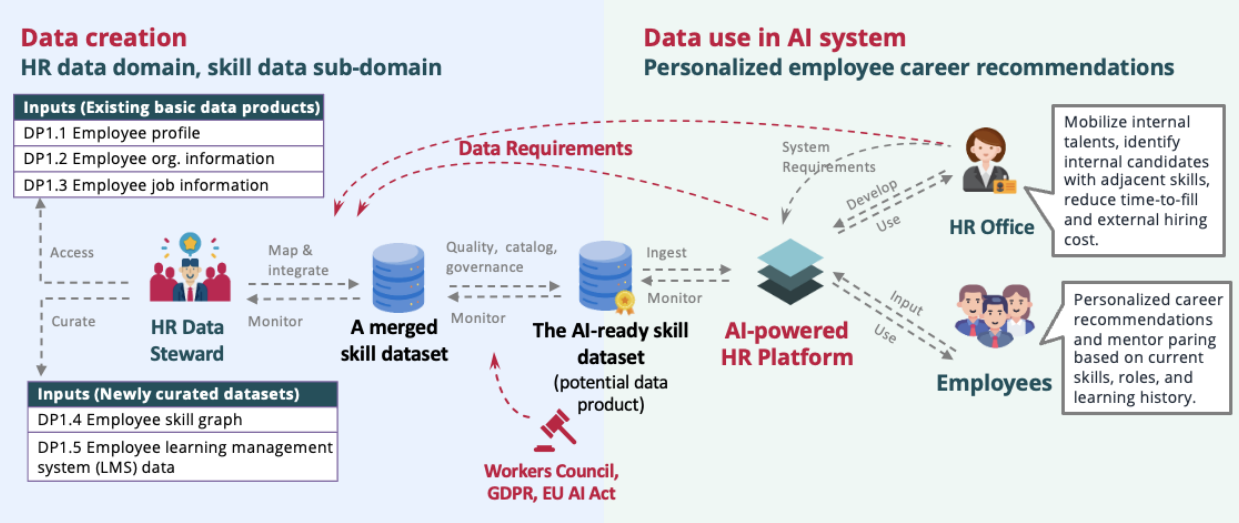


Figure 4: Preparation of AI-ready data for use in an AI-enabled HR platform

[19] A data product is a "discoverable, well-documented, high-quality and reusable dataset that satisfies recurring information needs by consumers"; see Blohm, I., Wortmann, F., Legner, C. et al. (2024; Data products, data mesh, and data fabric, Business Information Systems Engineering, 66, 643–652, https://doi.org/10.1007/s12599-024-00876-5).

3.4. CHARACTERISTICS OF AI-READY DATA

Based on the comprehensive analysis of more than 100 sources, AI-ready data can be characterized by six dimensions[20]:[21]

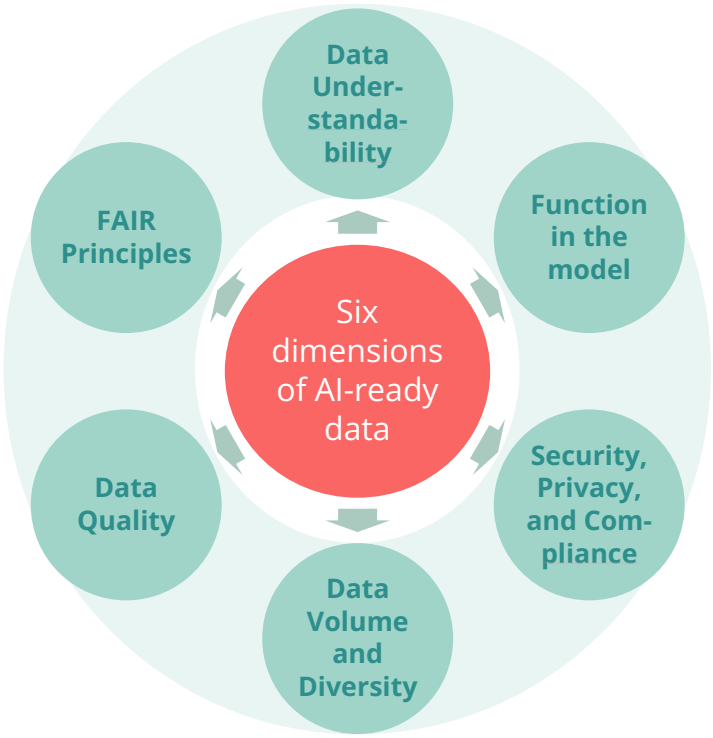


Figure 5: Six dimensions of AI-ready data

Some of these dimensions **build on and extend the established data management dimensions**, such as data understandability, FAIR data, data quality, and security, privacy, and compliance. Many organisations have developed practices to support these dimensions, for example through data catalogues, metadata management and data governance. However, AI systems place more stringent and often more specific requirements on data in these areas. As a result, **existing data management practices** often **need to be adapted and extended** to address the specific requirements of AI systems and their intended purpose(s).

AI systems also introduce **additional dimensions and requirements**, such as sufficient volume and diversity to support representativeness and reduce the risk of bias amplification, and require model-specific of data. These AI-specific data preparation activities may include labelling, enrichment, transformation and feature engineering, and increasingly require continuous monitoring of both training and inference data.

In the following, we will elaborate on each dimension individually, although it is important to underline that the dimensions are highly interdependent.

Data understandability: Metadata[22] should provide the information needed by both human and machines to understand and process the dataset within an AI system.

- *What is required?* Datasets must be **documented through metadata,** describing the structure and content, terminology, **governance** (e.g., sensitivity, access permission, licensing agreements) and context (e.g., provenance or use). This documentation should provide the information needed to understand and process the dataset within an AI system. For agentic AI, machine-readable metadata

[20] Competence Center Corporate Data Quality (CC CDQ) research (https://cc-wiki.cdq.com/Glossary#AI-ready_data).

[21] Tschofen, J. (2026). What Is AI-Ready Data? A Practical Definition (and Checklist) for Enterprise Teams, CDQ (https://www.cdq.com/blog/what-ai-ready-data).

[22] "Metadata" are "data about data" that facilitate access, management and sharing of large sets of structured and/or unstructured data. They help users and systems understand their meaning, structure, origin, format, quality, relationships and intended use. They support the discovery, interpretation, management and appropriate use of data.

are essential for enabling AI agents to interpret and use data effectively. Beyond describing the dataset as a whole, **the data themselves** should make their **meaning explicit:** relationships and roles between pieces of data should be stated explicitly, so this is not implied based on free text or column names. Finally, semantic assets describing the content of the data (e.g., code list, taxonomy, data model) should be versioned, so that changes such as a category being split into two are recorded.

- *Why does it matter?* Metadata communicate basic information about datasets to AI developers and users who have not been involved in data collection and have limited knowledge about the dataset. With agentic AI, metadata are essential for automated processing of the dataset in AI systems and contribute to performance improvements.

FAIR principles[23]: To facilitate reuse and repurposing in AI systems, the dataset should be **Findable**, **Accessible**, **Interoperable**, and **Reusable**.

- *What is required?* Datasets should be clearly identified and discoverable, accessible under defined conditions, described using shared standards and semantics to support reuse across AI applications.
- *Why does it matter?* Identifying and accessing suitable data for AI systems can be time-consuming and complex. Even when relevant data are findable, access may be delayed by lengthy approval processes. Applying FAIR principles facilitates the reuse of existing datasets across AI use cases and reduces the effort required to integrate and repurpose them. Across organisations, findability and interoperability also depend on shared reference data that are maintained outside the individual dataset. Persistent identifiers allow datasets from different sources to be linked reliably and consistently over time. Where such identifiers and reference data are not available or not systematically maintained, interoperability cannot be established through dataset-level documentation alone.

Data quality[24]: The dataset should be systematically **cleansed**, **validated**, and **contextually enriched** to support the AI system(s).

- *What is required?* Datasets should be sufficiently accurate, complete, consistent and timely for the intended AI system and purpose. Errors, missing values, duplicates, noise and anomalous observations should be identified and appropriately handled. Validation and verification become critical to ensure high data-quality.
- *Why does it matter?* Errors or noise in the data compromise AI model performance and the reliability of AI outputs. Even small data issues can propagate through AI models and result in significant, large-scale mistakes. The most damaging issues are those that produce no error: a loading process that skips faulty records reports success, and the data appear complete.

Data volume and diversity: The dataset must have an appropriate level of data volume – in the context of structured data with respect to the number of observations ("size" of the data), but also with respect to the number of variables ("dimensionality" of the data) – and diversity to ensure sufficient representativeness ("bias-free sampling") and avoid other types of biases in AI systems[25].

- *What is required?* The dataset should be extended with new attributes or complementary data for sufficient volume, coverage, or types to support data exploration and prevent biased sampling. Specific attention should be given to "sampling bias", "historical bias", "label bias", and "temporal bias".
- *Why does it matter?* Insufficient or unrepresentative data can lead to overfitting, poor generalisation and "biased AI model" behaviour. Adequate volume and diversity help AI systems learn relevant patterns, capture less frequent cases and perform more reliably across different situations and contexts.

[23] "The FAIR Guiding Principles" (https://www.gofair.foundation/fair-principles).

[24] "High-quality data" means that the data are fit for their intended use by people, organisations or systems, including AI systems, in operations, decision-making, planning or other applications ("right data"), AND that the data are "free from defects" with respect to the relevant data quality dimensions, such as accuracy, completeness, consistency and timeliness ("data are right", i.e. sufficiently correct for that intended use).

[25] "Bias-free sampling" refers specifically to bias in the sampling process. AI systems may be affected by many other forms and sources of bias throughout their lifecycle; see Srinivasan R. & Chander, A. (2021; Biases in AI systems, Communications of the ACM, 64, 8, 44-49, https://doi.org/10.1145/3464903 & https://cacm.acm.org/practice/biases-in-ai-systems/).

Security, privacy and compliance: The dataset should support secure, lawful and responsible AI use in accordance with applicable regulatory and ethical requirements.

- *What is required?* The collection, processing, access and use of the dataset in the AI system should comply with relevant regulations (e.g., GDPR, EU AI Act), internal policies, and applicable security and ethical standards.
- *Why does it matter?* Data used in AI systems may contain sensitive, personal or otherwise protected information, and AI models, specifically generative AI models, risk leaking these data. New AI regulations impose transparency and data governance (e.g., monitoring model outputs). Appropriate safeguards are therefore required to protect privacy and security, ensure lawful use, and support responsible and compliant AI.

Function in the model: The dataset should meet the specific requirements of both the AI model and its lifecycle stage, aligning with the defined intended purpose(s) of the AI system(s) it serves.

- *What is required?* Data preparation should reflect how the dataset will be used in the AI system, for example for training, validation, fine-tuning, retrieval, grounding or inference. Depending on the AI system, this may require labelling, feature engineering, transformation, enrichment or adaptation to specific input formats including semantically structured formats.
- *Why does it matter?* AI models require specific input data. Feature-engineered datasets can misalign with downstream model requirements. A dataset is often safe for retrieval but not for training where the runtime access control boundaries are removed.

4. Use Cases: What are Practical Challenges and how are Data Prepared for AI Systems?

4.1. OVERVIEW

Based on an open call for AI-ready data use cases, exemplary cases to illustrate the relevancy of AI-ready data for different types of AI systems and contexts. Table 1 below provides an overview of the use cases, which will be presented in more details in the following sections. The majority of the cases aim at making existing (public or private) data ready for consumption by a variety of AI systems, mostly LLM-based like chatbots and RAG systems. In contrast, Apertus provides insights into the training of foundation models.

	Use Case	Motivation for AI-ready data	AI systems	Datasets and types	Understandability	FAIR principles	Data quality	Volume and diversity	Security, privacy and compliance	Function in the model
1	**Swiss AI-Ready Tourism Data** *HSLU, Switzerland Tourism et al.*	Findability and accuracy of touristic information	LLM-based; compound AI system	Websites and other touristic information	**X**	X	X	-	X	X
2	**Meta-driven AI for search & analysis** *Stadt Zürich*	Error reduction of AI outputs during search and analysis	Generative AI	Structured and unstructured	**X**	X	X	-	**X**	-
3	**AI-ready Knowledge for Reliable Content Generation in Vocational Education** *Altus Forge; LSCP*	Reliable content generation, reduction of expert rework and improvement of traceability and consistency	Generative AI (LLM), supported by agent-assisted domain modelling and graph-based contextualisation	Training material (PowerPoint, PDF) and legal sources (Fedlex)	**X**	X	X	-	*	X
4	**Trusted data for Apertus** *Opendata.ch, Swiss Data Alliance, ETH AI Center, EPFL AI Center, et al.*	Reduction of responsible data & compliance overhead, output quality	Foundation model	Large-scale web text, articles and books in all languages	**X**	**X**	X	X	X	X
5	**Swiss Life Digital Heritage** *Swiss Life, Accenture, bitmark Assoc., et al.*	Availability and accessibility of physically retained and archived data for AI systems	LLM-based; generative AI; RAG-based chatbots	>10 000 books dating back to 1850	**X**	**X**	**X**	X	X	**X**
6	**MODI Sandbox** *Beyond Civic, et al.*	Availability and accessibility of parking information for users	Conversational AI powered by agentic AI	Parking space descriptions and availability	**X**	**X**	X	-	X	**X**
7	**Automated invoice matching in procurement** *Samedo, Rinascai*	Verifiable invoice release without interfaces or data migration	Agentic AI with browser automation	Purchase orders (inventory management), supplier invoices (PDF, many layouts)	**X**	**X**	X	X	X	X
8	**Ari – your City Companion**	Aggregation of relevant city	RAG-enabled and agentic AI	City information from different sources	X	X	X	X	X	X

	LiveMap AG (Axon Group)	information, combined with map		(geodata, Web, transportation, etc.)						
9	**4LM legal chatbot on Legilux** *Luxembourg Government, Mistral AI, Cognizone*	Chatbot answers grounded exclusively in the law applicable on the day of the question	Generative AI + Knowledge Graph; RAG-based chatbot	Official Journal acts and consolidated versions (XML text, RDF metadata). Note: The Swiss FedLex data follows the same format and standards.	**X**	**X**	**X**	X	**X**	**X**
Legend: X or **X** (dimension is relevant or highly relevant), - (dimension not observed)										

Table 1: Overview of use cases

4.2. USE CASE: SWISS AI-READY TOURISM DATA

Use Case Name	**Swiss AI-Ready Tourism Data**
Overview	
Context and motivation	Tourists increasingly rely on LLM-based AI systems like ChatGPT, Claude, or Gemini before and during their trips. Swiss tourism organisations already observe declining search engine traffic while LLM-based referrals grow — but Swiss tourism data are fragmented across various IT systems with heterogeneous standards, making them poorly suited for AI model and/or system consumption.
Description	SwissAIRTD develops a prototype to assess and improve the AI-readiness of these data, including a measurement system to track how accurately LLMs reproduce Swiss tourism information, and evaluates concrete measures across selected use cases and geographic areas.
Actors: Who is involved?	HSLU, discover.swiss, Made in Bern, Zürich Tourismus, Switzerland Tourism, FH Ost Rapperswil
AI system: For which type of AI was the data prepared?	- LLM-based systems like ChatGPT, Claude, or Gemini used by tourists before and during their trips; - Compound AI systems (LLM + RAG)
Phase: Which phase(s) is covered?	Training (LLMs) and Inference (LLM + RAG)
Status	Prototype
AI-ready data	
Data: What datasets / type were required?	Websites and other information, mostly related to Points of Interest (PoI), that is relevant for visitors
Key data challenges:	- Highly fragmented data landscape (many systems, heterogeneous models and standards) with unclear licensing for AI use. - Little established knowledge on which concrete measures effectively improve AI-readiness in tourism, as the field is evolving rapidly.
Data preparation: How is data prepared for the AI-system?	Adapting data models and metadata for AI consumption (standardisation, licensing, enrichment with AI-relevant attributes)
AI-ready data dimensions	
Data understandability	**X** (metadata, schema.org, with tourism relevant objects organised in a hierarchy)
FAIR principles	X
Data quality	X (tourism organisations check accuracy and work on completeness according to their mandate)
Volume and diversity	-
Security, privacy and compliance	X (Creative Commons licenses are widely used for pictures; text is normally not licensed)
Function in the model	X (adaption of data model for LLM+RAG)
Learnings and recommendations	
Learnings and recommendations	- Preparation leads to significant improvements of output quality. For instance, the time and day of the update as well as object attributes such as accessibility make objects appear more often in LLMs. - Adaptation of data model is required to improve consumption for LLM + RAG.
Further information	Project website: https://www.hslu.ch/en/lucerne-university-of-applied-sciences-and-arts/research/projects/detail/?pid=7137

4.3. USE CASE: METADATA-DRIVEN AI FOR DATA SEARCH & ANALYSIS

Use Case Name	**Metadata-driven AI for data search & analysis**
Overview	
Context and motivation	General (Wikipedia-like) knowledge isn't enough to answer specific questions. Specific problems require specific data. In many cases, AI results must be explainable and traceable. Hallucinating AI undermines trust in the results.
Description	The metadata-based AI approach combines the strengths of LLMs (e.g., natural language processing, general knowledge) with specific data relevant to answering organization-specific questions.
Actors: Who is involved?	Stadt Zürich, Organisation & Informatik (OIZ)
AI system: For which type of AI was the data prepared?	Generative AI
Phase: Which phase(s) is covered?	Inference
Status	Proof-of-Concept
AI-ready data	
Data: What datasets / type were required?	Unstructured data (e.g. documents, images)
Key data challenges:	LLMs establish connections based on textual matches ("text snippets"). Synonyms, homonyms, and context-dependent meanings of words can lead to erroneous connections. Correlations should be based on semantics.
Data preparation: How is data prepared for the AI-system?	The PoC includes an AI search assistant. This assistant uses internal documents (including metadata) and an ontology to answer user questions accurately and in a way that can be explained. **Data Preparation** The data source is a records management system in which the documents (including metadata such as access rights) are stored according to the filing plan. From the documents (and their metadata), a knowledge graph (instance data as nodes and edges) and an ontology (classes, properties, and relationships) for knowledge representation are automatically generated (with AI support) and stored in a graph database. The ontology is manually verified and, if necessary, corrected and refined. The refined ontology is stored in the citywide data catalog ("Stadtweiter Datenkatalog") as a domain-specific data model and can thus be reused for other tasks. **Using the AI Search Assistant** Based on a user's question, a query is first sent to the graph database. This identifies documents and contextual knowledge relevant to answering the question. A corresponding excerpt from the graph database and the selected documents are provided to the prompt (GraphRAG). The user's access rights are taken into account in this process. The contextual knowledge and the documents used make it easier for the user to contextualize the AI's response and evaluate it.
AI-ready data dimensions	
Data understandability	**X** (metadata, including – business metadata: domain data models with business objects and relationships, value lists, information on business processes and legal bases; – governance metadata: Responsibilities & roles (e.g., contact persons, responsible persons), access levels

	– technical metadata: systems, including golden source (data is created), single source of truth (data is being retrieved) – lineage: domain data models with business objects and relationships, value lists, information on business processes and legal bases;
FAIR principles	X
Data quality	X
Volume and diversity	-
Security, privacy and compliance	**X**
Function in the model	-
Learnings and recommendations	
Learnings and recommendations	Added value of a metadata-driven approach: – metadata (including ontology) enables the generation of more precise answers, as these are based on quality-assured information and contextual knowledge (reduces hallucinations); – access permissions for the (internal) data or documents can be taken into account; – the result is explainable.
Further information	-

4.4. USE CASE: AI-READY KNOWLEDGE FOR RELIABLE CONTENT GENERATION IN VOCATIONAL EDUCATION

Use Case Name	AI-ready Knowledge for Reliable Content Generation in Vocational Education
Overview	
Context and motivation	Vocational education providers invest significant expert time in creating and maintaining high-quality learning and assessment content. Initial experiments with generating microlearning content directly from source documents showed the limitations of document-based AI approaches: outputs required substantial expert review because the AI lacked an explicit understanding of the relationships between competencies, learning objectives, taxonomy levels, domain knowledge and applicable legislation.
Description	Instead of providing the LLM with raw documents, this use case explores a different approach to AI-ready data , transforming domain knowledge into a structured, machine-readable semantic knowledge graph. The graph connects curriculum requirements, competencies, learning objectives, taxonomy levels, domain concepts and relevant legal sources. SHACL validation is used to verify structural consistency, and graph traversal is used to assemble the relevant context for a specific learning objective before content is generated. In the prototype, this approach was applied to two different content-generation tasks: a) **Microlearning content and assessment questions** The system generated microlearning and assessment questions based on the structured domain knowledge. Domain experts assessed the resulting question catalogue as directly usable, including more complex multi-scenario questions. b) **Learning material for a regulatory transition** The same domain model was used to create learning material for learners who had previously studied the TARMED tariff system and needed to prepare for an examination based on the newly introduced TARDOC methodology. Based on this structured knowledge, the AI generated a PowerPoint training document that outlined the relevant differences between TARMED and TARDOC and introduced the new methodology. More than 90% of the final presentation used by the trainer was based on the automatically generated output, with only limited expert adaptation required. This demonstrates that the semantic knowledge foundation can support different types of learning content without having to reconstruct the relevant domain context for each individual use case. The use cases demonstrate how AI-ready semantic data can shift quality assurance upstream: from repeatedly correcting individual AI outputs towards modelling and validating the knowledge on which those outputs are based.
Actors: Who is involved?	Altus Forge GmbH; LSCP Life Sciences Consulting Partners
AI system: For which type of AI was the data prepared?	Generative AI (LLM) for question generation, combined with a SHACL-validated knowledge graph as semantic data foundation. The LLM performs graph-traversal-based prompting – it only generates

	content from structured, interlinked graph data, not from unstructured data (documents).
Phase: Which phase(s) is covered?	Inference
Status	Prototype/PoC completed
AI-ready data	
Data: What datasets / type were required?	- Training materials (PowerPoint, PDF) - FedLex legal texts (AKN / ELI)
Key data challenges:	AI systems have no structured view of the relationships between legislation, learning objectives, taxonomy levels, and domain knowledge. Without semantic context, the AI could not reliably combine concepts into valid, educationally sound questions.
Data preparation: How is data prepared for the AI-system?	Creation of a structured semantic knowledge graph (JSON-LD, SHACL-validated) containing: - **Swiss vocational curriculum** (BiVo 2023) with competency areas A–E, learning objectives, taxonomy levels K1–K6, curriculum blocks and training days (16 üK days), question types (MC 4/5-choice, true/false, flashcards), domain-specific terminology and rules. - Additionally: complete **Legal Domain** with all relevant Swiss laws (KVG, UVG, AHV etc.) from Fedlex (AKN, ELI) modelled and linked to learning content – giving the AI full context from legislation through to individual learning objectives) Build semantic model as in LLM: 1. **Domain Modeling (DMA):** All entities and relationships modelled according to their real-world origin logic using the Domain Modeling Agent. The Model validation is supported by a 3D Visualization. 2. **Legal Domain Integration:** Relevant Swiss legislation modelled and linked to curriculum objectives, providing full traceability from law to learning content. 3. **SHACL Validation:** Model consistency verified before use. 4. **Graph-Traversal Prompting:** AI generates questions exclusively from linked graph data of the selected curriculum block.
AI-ready data dimensions	
Data understandability	X (metadata)
FAIR principles	X
Data quality	X
Volume and diversity	**-**
Security, privacy and compliance	**-**
Function in the model	**X (semantic knowledge graph, domain modeling)**
Learnings and recommendations	
Learnings and recommendations	- **Structure before generation**: Reliable AI output depends strongly on the quality, structure and explicit semantics of the underlying data. - **Quality assurance can move upstream**: Validating the domain model reduces the need to identify the same types of errors repeatedly in generated content. The 3D visualisation improved significant the communication between subject matter experts and the data architect. - **Traceability matters**: Generated content can be linked back to learning objectives, competency levels and relevant source knowledge.

	- **AI-ready knowledge is reusable**: The semantic model is independent of an individual prompt or LLM and can serve as a persistent knowledge foundation for further AI applications.
Further information	-

4.5. USE CASE: TRUSTED DATA FOR APERTUS

Use Case Name	Trusted Data for Apertus
Overview	
Context and motivation	LLMs like Apertus require massive amounts of training data – in the order of trillions of tokens. This training data must be trusted and compliant and reflect the pluralism of our society.
Description	Large-scale public data and pipelines for pre-training Apertus as fully open, transparent and compliant LLM: "Our primary goal is to establish an open, reproducible, and high-quality foundation for model training, focusing on general language modelling, multilingual breadth, mathematical and coding capabilities, and limiting ourselves to permissively licensed data." (1) To support the preparation of AI-ready, trusted data for Apertus, the Swiss Data Alliance has conceived an inbound/outbound model under a governance that increases data sharing while reducing the compliance overhead of open models (3)
Actors: Who is involved?	- Apertus: Swiss AI Initiative (ETH AI Center, EPFL AI Center, Swiss National Supercomputing Center - CSCS), - Trusted Data for Apertus: Opendata.ch (prototype fund), Swiss Data Alliance, ETH AI Center, EPFL AI Center
AI system: For which type of AI was the data prepared?	Foundation large language model (Apertus)
Phase: Which phase(s) is covered?	Training
Status	in production (Latest release 1.5, July 2026)
AI-ready data	
Data: What datasets / type were required?	- Large-scale open data with global representation: over 1'800 languages, 40% non-English content. - Unstructured data (web texts, articles and books) as well as code, mathematical, and structured data, with explicitly no specific limitation.
Key data challenges:	Preparation of large-scale data, specifically: - Compliance overhead for data protection and AI regulations (sensitive data, opt-out preferences, etc.) - Clearing of legal restrictions (copyrights, etc.)
Data preparation: How is data prepared for the AI-system?	Apertus has developed **Responsible Data Practices** that better satisfy the provisions of the EU AI Act, Swiss data protection laws and copyright requirements: 1. **Start with Public & versioned data only**, primarily sourced from open repositories (such as CommonCrawl) 2. Respect **AI opt-out preferences** of authors and publishers (removes 8% English, 4% non-English data) 3. **Prevent model memorisation** of data, to address the reproduction of copyrighted content 4. **Quality:** Remove personally identifiable, toxic & harmful content, various forms of low-quality data (e.g., spam, non-human text, etc.). All pre-processing steps are documented to ensure full transparency and facilitate replication of results. The pipeline code is released to recreate all of the data that was used for training the models.
AI-ready data dimensions	
Data understandability	X (metadata)
FAIR principles	X

Data quality	**X** (removing toxic & harmful content, and various forms of low-quality data, such as spam, non-human text, etc.)
Volume and diversity	**X** (multilingualism, i.e. global representation of languages)
Security, privacy and compliance	**X** (license filtering, removing PII and respecting AI opt-out preferences)
Function in the model	X
Learnings and further information	
Key learnings and recommendations	**Responsible data practices** ensure that the large language model is trained only on permissive content, ensures privacy and filters toxic content. (2) **AI-ready data require management and governance**, e.g., through Data Commons, and Data Spaces (3)
Further information	(1) Project website: https://apertus-ai.org (2) Technical Report: Apertus: Democratizing Open and Compliant LLMs for Global Language Environments (2025, https://aclanthology.org/2026.acl-long.2172/) (3) https://www.swissdataalliance.ch/news/apertus-challenge-prototype-fund

4.6. USE CASE: SWISS LIFE DIGITAL HERITAGE (BITMARK – RAG SYSTEM)

Use Case Name	**Swiss Life Digital Heritage**
Overview	
Context and motivation	Swiss Life's enormous knowledge treasure gathered over the past 150 years which has been locked away in the archives is to be made accessible and useful. More generally, we want to make knowledge that is currently available only in physical form digitally usable and accessible
Description	Transformation of physical library into AI-ready content and creation of a digital library including structure and accessibility of the knowledge through RAG-system.
Actors: Who is involved?	Swiss Life, Accenture, bitmark Association
AI system: For which type of AI was the data prepared?	- Conversational RAG enhanced AI chat with the library: A LLM allows to chat with the library, books and content; the AI Chat is integrated with the Swiss Get More Brain-platform for a seamless experience, access via existing Swiss Life SSO - Further use cases are planned
Phase: Which phase(s) is covered?	Inference
Status	PoC (early 2026), go-live (testing) in June 2026, full go-live end of 2026
AI-ready data	
Data: What datasets / type were required?	Book archive: 120m^2, 64 shelves, each comprising 200-300 books (total of approximately 700 linear meters).
Key data challenges:	- Size of the archive, requiring efficient digitalization - Diversity of input data (fonts, languages, quality, types, layouts, images, formulas etc.) - Interoperability across different platforms → use of open-source bitmark standard
Data preparation: How is data prepared for the AI-system?	**Data preparation:** 1. **Books are scanned:** 3 million pages (double sided) converted to 6 million scanned images 2. **Content is digitized and structured:** Text, tables, formulae, and images parsed, metadata extracted, with both digital & manual QA 3. **Content is represented for retrieval:** Vectorized representation of the bitmark-format is created to enable low latency RAG-system 4. **Content is transformed into the Digital Library:** Digitalized content is converted to Swiss open-source standard bitmark (https://docs.bitmark.cloud/) and stored in the Get More Brain-platform as digital library. Further steps include the creation of Linked data, thus permitting further uses cases
AI-ready data dimensions	
Data understandability	**X** (LLM-based pipeline for content transformation, metadata for retrieval, bitmark for relationships)
FAIR principles	X (bitmark format supporting interoperability across platforms)
Data quality	X (three step validation: AI-based, statistical, manual; the AI-based digitization process includes both automated and manual quality assurance to ensure the accuracy, completeness and consistency of the extracted content. Identified errors or quality issues can be addressed through targeted reprocessing of the

	affected content. Statistical methods ensure, that quality across content is measurable, validated and actionable.)
Volume and diversity	X (large volume and diversity of languages, fonts, layouts, content types, images, formulas and source quality)
Security, privacy and compliance	X (Copyright, access rights and lawful use of the digitised content were explicitly considered. Access to the digital library and RAG system is controlled through Swiss Life authentication and authorisation mechanisms)
Function in the model	**X** (historical library content is transformed into semantically structured and metadata-enriched knowledge objects (bitmark chunks) that can be retrieved and referenced by a RAG-enabled LLM at inference time. The structured content, metadata and contextual relationships captured in bitmark provide a valuable foundation for effective retrieval, grounding and context-aware AI-generated answers).
Learnings and further information	
Learnings and recommendations	- The pre-existing structure enforced by the bitmark standard improves the preparation of content for AI use cases and systems. - It is a chunking format for PDFs and XML-based documents and includes references to the original content, enabling RAG systems to reference and display the content used, including its context.
Further information	Project website: N/A Standards: bitmark standard (https://docs.bitmark.cloud/)

4.7. USE CASE: MODI-SANDBOX – CONVERSATIONAL ANALYTICS FOR PARKING DATA

Use Case Name	MODI-Sandbox: Conversational analytics for parking data
Overview	
Context and motivation	Mobility data are becoming increasingly indispensable for a functioning and efficient mobility system. Simplifying the exchange and use of mobility data across all modes and means of transport makes an increasingly important contribution to ensuring that infrastructure and services can be optimally planned, operated, and utilised. The Federal Council therefore proposes the establishment of a national MObility Data Infrastructure (MODI), through which the various stakeholders can network more efficiently and provide, link, and access mobility data in a standardized manner. The goal of the sandbox is to support users requesting parking related information, i.e. what is the capacity of a parking or how many parking spaces are currently available.
Description	Conversational analytics powered by agentic AI solves a critical bottleneck: traditional Business Intelligence (BI) is too slow, static, and disconnected from daily operations. While legacy BI tools require specialised teams to build rigid dashboards, agentic AI introduces autonomous data assistants. These agents don't just answer questions; they plan, reason, fetch, and analyse data across an entire organization and beyond (data spaces). MODI supports users requesting parking related information, i.e. what is the capacity of a parking or how many parking spaces are currently available. The AI agent translates plain question such as "how many parkings are available in city X?" into the right tool call (query with SQL against the data product) without you needing to know the schema, the table name or that the data lives in a nested JSON file.
Actors: Who is involved?	- Beyond Civic (Provider of governance and tech stack provider of sandbox - Fondation des Parkings: Data (Product) Provider - Parkker: Data User
AI system: For which type of AI was the data prepared?	**Conversational AI / Agentic AI.** For simplicity reason, Claude Desktop (Anthropic) is being used as agent. The MODI-Sandbox offers MCP endpoints so that the agent can act based on an open protocol. The MODI-Sandbox is agnostic to whether the data product is accessed via a REST API or file-based cloud storage (i.e. CSV, JSON or Parquet file on an S3 bucket).
Phase: Which phase(s) is covered?	Inference
Status	Prototype
AI-ready data	
Data: What datasets / type were required?	Data on available parking spaces in parking garages in Geneva. This includes availability (occupancy), descriptions of the parking spaces (e.g., with EV charging, etc.) and the pricing structure.
Key data challenges:	One of the main challenges is to properly define the underlying data schema (metadata) of the data product so that agents are able to semantically understand the data product and build the right queries to answer questions in natural language.
Data preparation: How is data prepared for the AI-system?	Essentially, there are two types of data products: one consists of static data—such as how many parking spaces the parking garage has, how many of them have an EV charging station, whether there are women-only parking spaces, and so on. The other data product

	is dynamic and indicates "occupancy" or the number of available parking spaces. This data product is updated every 60 minutes via an automated job. The data stewards describe the data products using a comprehensive and open metadata standard called Open Data Product Specification (ODPS – https://opendataproducts.org/). The data (product) owner or data steward needs to enrich the data product with context (i.e. in which use case the data product can be used, machine-readable data schema description, quality metrics, SLA's, etc.) so that the agent can understand the semantics and come up with trustworthy reasoning.
AI-ready data dimensions	
Data understandability	**X** Comprehensive metadata description using Open Data Product Specification (ODPS)
FAIR principles	**X** Data catalogue and data product accessibility via MCP endpoints
Data quality	**X** Data quality is being described via ODPS: • Processing frequency • Consistency • Timeliness • Accuracy • Completeness • Validity • Uniqueness
Volume and diversity	**X** Description of all parking facilities in Geneva (managed by Fondation des Parkings) with comprehensive data attributes provided by the APDS standard (https://allianceforparkingdatastandards.org/).
Security, privacy and compliance	- The data products do not contain any PII data, nor do they contain any business sensitive data.
Function in the model	**X** Semantic modelling to support conversational AI. Parking occupancy data is updated every 60 seconds to make sure it supports the use case related to timeliness.
Learnings and further information	
Learnings and recommendations	- Use open and widely adopted standards such as ODPS, MLCommons Croissant and MCP. - Invest time in understanding to what degree data products must be described so that the agent can act with high confidence on the reasoning.
Further information	Project website: https://www.bav.admin.ch/en/data-for-an-efficient-mobility-system Documents: n/a

4.8. USE CASE: AUTOMATED INVOICE MATCHING IN PROCUREMENT (SAMEDO)

Use Case Name	**Automated invoice matching in procurement**
Overview	
Context and motivation	Supplier invoices are checked by hand in most small commerce organisations: open the purchase order in one system, find the invoice in another, compare article by article. Under time pressure the routine is shortened first, and price increases, duplicate orders and faulty documents pass unnoticed. Preparing the data upstream is out of reach at this size, so the case shows what AI-readiness means when the data cannot be moved.
Description	Every morning an agent checks each purchase order that has arrived and is still unpaid. It reads the open orders from the inventory management system, finds the matching invoice in the invoice inbox and compares position by position: unit price against the maintained purchase price, quantity against the ordered quantity. Accounting receives one traffic-light report: green for release, amber for a price deviation with the exact difference per article, red for a missing or unattributable invoice. Every entry links back to the order and carries invoice number and payment terms. The agent operates both systems through the browser, so no interface was built and nothing in the existing IT changed.
Actors: Who is involved?	- Samedo (online retailer); - Rinascai (AI Office as a Service)
AI system: For which type of AI was the data prepared?	Agentic AI (Claude, Anthropic) with browser automation and document understanding. The comparison rules are held in a documented, reusable skill and the agent operates the source systems as a user rather than through integrations.
Phase: Which phase(s) is covered?	Inference
Status	Pilot in production, built in a few days
AI-ready data	
Data: What datasets / type were required?	- Purchase orders from the inventory management system connected to the main webshop: article, ordered quantity, maintained purchase price, delivery and payment status - Supplier invoices as PDF in the dedicated invoice inbox on the internal mail server, one layout per supplier - Comparison rules, tolerances and exception cases, held in the skill rather than in either system
Key data challenges:	- High format diversity at low volume: every supplier has its own invoice layout, so documents have to be understood rather than parsed. - The reference value depends on maintenance, and one order to one invoice does not always hold (duplicate orders, part deliveries, collective invoices). - The data cannot be moved: no API access, no development project, no change to the existing IT.
Data preparation: How is data prepared for the AI-system?	- Comparison rules, tolerances and exception paths written out in a documented, reusable skill. - Ordered quantity and maintained purchase price declared the authoritative reference; the invoice is the claim tested against them.

	- Order and invoice positions brought into one structure per run (article, quantity, unit price), so the comparison happens per article. - Access instead of extraction: the agent reads both systems through the browser inside existing user permissions, so no data copy and no additional store are created.
AI-ready data dimensions	
Data understandability	**X** (comparison rules documented in a reusable skill; authoritative reference fields designated; article-level mapping between order and invoice)
FAIR principles	X (every result carries its deeplink, invoice number and terms, and can be verified without repeating the check)
Data quality	X (validation against maintained master data; deviations, missing invoices and duplicate orders surface daily)
Volume and diversity	X (diversity rather than volume: one invoice layout per supplier, read by document understanding instead of per-supplier parsing rules)
Security, privacy and compliance	X (existing user permissions, no data copy and no shadow store; every release decision documented with its source link)
Function in the model	**X** (the agent acts on structured, matched positions and an explicit rule set, not on invoice text)
Learnings and further information	
Learnings and recommendations	- Where data cannot be prepared upstream, readiness can be produced at the point of use: an explicit rule set, designated reference values and document understanding at run time, in days rather than quarters. - Master data maintenance sets the ceiling, not the model. Declaring the purchase prices the reference makes their neglect visible, so the use case improves the data it depends on. - An AI-ready output carries its evidence: source link, reference value and exact difference. A verdict without them has to be checked again.
Further information	Company website: rinascai.com

4.9. USE CASE: ARI BY LIVEMAP AG – YOUR CITY COMPANION

Use Case Name	Ari – your City Companion
Overview	
Context and motivation	People are stuck in their routines. 51% of Zurich residents say they always frequent the same places, and 69% say they travel the same route, even though they know there are better ones (LiveMap Research in ZH, May 2026, 206 participants). They want to experience their city to the fullest, but they don't know where to start. Information is scattered across different platforms and touchpoints and is often not well structured or easily findable. This is where we come in. We aggregate all relevant city information from different sources, make it AI-ready, combine it with a custom map and a personalisable routing engine, and present it all in one intelligent platform: Ari. All of this is supported by a sovereign infrastructure (on-premises inferences with our own GPUs on open-weight models, locally hosted, etc.).
Description	All relevant data are automatically integrated into a data pipeline, where they are structured, metadata is created and the data is processed for use in widgets and on our map within the app. This data is then findable and usable by Ari's AI-Layer.
Actors: Who is involved?	LiveMap AG (Axon Group)
AI system: For which type of AI was the data prepared?	RAG-enabled and agentic AI-System
Phase: Which phase(s) is covered?	- Inference on open-weight models (run) - Training of data and city geo-model
Status	Production (available in app stores)
AI-ready data	
Data: What datasets / type were required?	- public and permissively licensed GIS data from the federal government, cantons and municipalities or other sources. - social media - website scraping (events, sports clubs, etc.) - exclusive partner data - background tracking data - own processed geodata
Key data challenges:	- Geodata that varies in format, geographic coverage and update frequency. Websites often contain useful details that lack a consistent structure. - Ari requires entries containing clear details of locations, dates and sources. Furthermore, a distinction between outdated or missing information and confirmed facts must be clearly indicated (quality).
Data preparation: How is data prepared for the AI-system?	- LiveMap collects data through feeds, APIs, file imports and website extraction and own geoprocessing. Language models help extract structured information from website content. - Automatically mapping source fields to standardised definitions, harmonising dates and geographical coordinates, and checking datasets for inconsistencies as part of the data pipeline.
AI-ready data dimensions	
Data understandability	X Ari needs to know what each dataset contains and how to interpret it. Catalogue metadata describes field meanings, sources, geographic coverage and available queries.

FAIR principles	X Findable: Data are identified and sourced either through web-accessible datasets or directly provided by the organisation. Accessible: Data are accessed through geospatial services, organisational access, web sources, or data provided directly by our users. Interoperable: The data pipeline automatically standardises, harmonises, and integrates heterogeneous sources into a common structure. Reusable: Data sources, metadata, provenance, and processing steps are systematically documented and preserved within the pipeline.
Data quality	X In the data pipeline, various tests are conducted and data from different sources is consolidated; for example, valid geometry is checked, duplicate events are filtered out, and unstructured data is tested in a structured manner. Verification processes are additionally integrated (reference points and standards).
Volume and diversity	X LiveMap requires a vast number and wide variety of datasets, which come from the federal government, cantons, municipalities and private and partner organisations. These include all kinds of websites and APIs, as well as social media. The data types range from 3D to background tracking data. For example, we have integrated around 130 datasets for the city of Zurich. The aim is to integrate all data from every municipality. This amounts to approximately 73 000 datasets.
Security, privacy and compliance	(X) We not only manage data intended for the public, but also for private organisations. This data contains specific metadata and passes through a separate pipeline and a different database. Our local GPUs ensure that data never leaves our system and remains secure.
Function in the model	X Data are structured to cover all use cases: training, validation, fine-tuning, retrieval, grounding or inference,
Learnings and further information	
Learnings and recommendations	- Reliability: Monitoring data pipelines, managing API limits and outages, versioning raw data. - Quality: Documenting data sources and collection methods; identifying gaps and distortions; levelling the gap. - Providers: Building relationships, creating incentives, proposing optimization reg. formats. - Verification/Merging: Defining reconciliation rules, resolving conflicts and carrying out manual checks in borderline cases. Scaling at 100% quality is still hindered by this. - Closing gaps: Using verified data as a reference value.
Further information	Project website: https://www.myari.app/en

4.10.USE CASE: 4LM – LEGAL CHATBOT ON LEGILUX (LUXEMBOURG LEGISLATION)

Use Case Name	**4LM – Legal chatbot on Luxembourg legislation as AI-ready knowledge base**
Overview	
Context and motivation	4LM ("Luxembourg's Legal Large Language Model") is a legal chatbot on Legilux, Luxemburg's official legal portal, that gives citizens, companies and administrations first guidance in their legal research. This example is relevant to Switzerland, since the collection of Swiss federal legal documents "FedLex" is published in the same format and follows the same standards as LegiLux.
Description	4LM is developed with Mistral AI, under the strategic partnership signed between Luxembourg and Mistral AI in June 2025. The publication back-office of the Service central de législation (SCL), which produces the Official Journal and Legilux, acts as the single source of truth and decides which texts the chatbot's knowledge base contains, and when. Through Mistral's ingestion API, it sends the XML of each relevant act or consolidated version to Mistral – in real time at publication and through a daily batch – and withdraws texts that are no longer applicable. Mistral vectorises the texts and runs the RAG-based chatbot, which is embedded in Legilux. The Mistral services currently run on Mistral's SaaS platform; a move to the state's infrastructure (CTIE) is planned for early 2027.
Actors: Who is involved?	- Luxembourg's Service central de législation (SCL) – publisher and data owner; - Luxembourg's Centre des technologies de l'information de l'État (CTIE) – API gateway and access control; planned host of the Mistral services; - Mistral AI – ingestion API and everything behind it (OCR, embeddings, knowledge base), LLM, RAG pipeline and chatbot; - Cognizone – development and maintenance of Legilux and its publication back-office, integration of the publication back-office (Casemates) with Mistral's ingestion API: real-time synchronisation at each publication, a nightly batch, and an administration page in the back-office to check, enrich and delete documents in Mistral's knowledge base
AI system: For which type of AI was the data prepared?	Generative AI: RAG-based conversational assistant built by Mistral AI (the "Le Lion" widget, embedded in Legilux through an iframe), using Mistral's LLM and embeddings.
Phase: Which phase(s) is covered?	Inference
Status	Test phase (SCL is testing and Mistral is refining the model based on SCL's feedback). The ingestion integration (real-time and nightly batch) is implemented; extension to codes and thematic compilations ("recueils") planned.
AI-ready data	
Data: What datasets / type were required?	**Acts** published in the Official Journal and their applicable consolidated versions, as XML. **Metadata in RDF** (JOLux ontology) with ELI identifiers: dates of entry into force and applicability, in-force status, and relations between acts (amends, repeals, consolidates). The **unit of processing** is the "ComplexWork", i.e. an act together with all its consolidated versions. Texts mostly in French, some in German. Not yet included: codes and recueils.

Key data challenges:	The main data challenge is **temporal validity** – the applicable law changes on specific dates, and not all amended laws have been consolidated yet (969 laws in force still to be consolidated in November 2025). More generally, the challenges are: - **Temporal validity:** a text published today may only enter into force later; repeals take effect on future dates. The knowledge base must reflect the law applicable on the day of the question. - **Versioning:** one act can exist as a base text, several successive consolidations and separate amending acts; only one authoritative version should be retrieved. - **Consolidation backlog:** where no consolidation exists yet, amendments only exist in separate amending acts. - **Metadata edge cases surfaced at ingestion:** e.g. acts carrying several entry-into-force dates, or placeholder consolidations without a published text. - **Synchronisation and deletion** of obsolete versions across two systems. - **Sovereignty and access control** for a state service built with an external AI provider.

Data preparation:
How is data prepared for the AI-system?
	- **Explicit selection rules, owned by the publisher** and documented in a bilingual specification:
	  - only texts in force and applicable are sent;
	  - the most recent applicable consolidation is preferred and replaces the base text or an older consolidation;
	  - an amending act is kept as long as no consolidation integrates its amendments;
	  - texts with a future entry into force are sent on that date; repealed texts are withdrawn on the repeal date.
	- **Source of truth stays with the publisher:** the back-office decides what to send and what to delete; Mistral only vectorises new texts and deletes obsolete ones.
	- **Two ingestion modes:** real-time synchronisation at publication, plus a daily batch as safety net.
	- **Administration and reconciliation:** administrators can check, enrich and delete content in the chatbot's knowledge base and, in a next step, list all URIs it contains.
	- **Structure-aware chunking** (specified in the initial design): never split an article, never cross title or chapter boundaries, target chunks of about 800 tokens with overlap, and attach metadata (act, version date, title/chapter path, article identifiers, language) so answers can cite the precise article.
	- **Roadmap:** vectors stored in Legilux's own search infrastructure, hybrid (keyword and semantic) search and role-based access to the indices.

AI-ready data dimensions	
Data understandability	**X** (ELI identifiers; JOLux RDF metadata describing temporal validity and relations between acts, consolidations and amendments; ComplexWork model)
FAIR principles	**X** (ELI persistent identifiers; open data on data.legilux.public.lu; standard formats; groundwork laid since the 2017 restructuring of the Official Journal)
Data quality	**X** (temporal correctness as the key quality dimension: in-force status and applicability dates; ingestion surfaces metadata

	inconsistencies; multi-step proofreading by the publisher before publication)
Volume and diversity	X (completeness of the corpus of applicable law matters more than volume; the consolidation backlog and the content types not yet included are coverage gaps; French and German texts)
Security, privacy and compliance	**X** (sovereign deployment at CTIE planned; machine-to-machine authentication through the state API gateway; access restrictions; limits on session length and number of sages; moderation of out-of-scope and malicious requests)
Function in the model	**X** (the dataset is the RAG knowledge base: exactly one applicable version per act; deletion as important as insertion; article-level structure for precise citation)
Learnings and further information	
Learnings and recommendations	- **For legal RAG, AI-readiness is above all temporal:** the right dataset changes on a given date. Validity metadata must be machine-readable and drive the selection automatically. - **Selection rules belong to the data owner, not the AI provider:** documenting what is sent, when and why makes the chatbot's knowledge base verifiable. - **Deletion is a first-class operation:** a RAG knowledge base needs withdrawal rules and reconciliation (e.g. a list of all URIs it contains), not only ingestion. - **AI-readiness builds up over years and is consumed in months:** the restructuring and standardisation of the Official Journal since 2017 (structured texts, ELI, RDF metadata) is what made the chatbot integration feasible; the government describes this groundwork as essential for 4LM. - **AI projects surface latent data issues:** metadata inconsistencies and the consolidation backlog become visible – and urgent – once an AI system depends on them. Interim rules (keeping amending acts) mitigate gaps, but answer quality is capped by consolidation coverage.
Further information	4LM project: ai4lux.public.lu/en/projets/legal.html Legilux: legilux.public.lu – open data: data.legilux.public.lu Government press release on the AI4LUX campaign and the Mistral AI partnership (March 2026): gouvernement.lu Answer to parliamentary question n° 3082 on consolidation and 4LM (27 November 2025): csv.lu (PDF)

5. Learnings and Recommendations

5.1. FROM SHARED CHARACTERISTICS TO OPERATIONAL PRACTICES

This white paper aims at finding a common language for what the notion "AI-ready data" implies. It does not intend to be instructional, e.g., to guide project teams what to do exactly, or infrastructure teams how to manage IT infrastructures, specifically. It is only one step leading from an encompassing claim (such as the theme of the Geneva AI Summit 2027[26]: "*Bridging Innovation and Trust*") towards more specific applications. The characteristics of AI-ready data set out in this white paper translate this high-level ambition into more specific data requirements. They describe the properties that data should possess if it is to support the development and use of AI in a reliable, responsible and value-creating manner. In this sense, the characteristics provide a common frame of reference. They help organisations, policymakers, technology providers and data holders discuss AI-readiness using shared concepts rather than isolated technical, legal or organisation-specific vocabularies.

Obviously, a common language is only the starting point, not the endpoint. The next step is therefore to move from describing what AI-ready data should be to determining how AI-readiness is established and maintained in practice. In the next sections, we therefore describe the learnings from the use cases and first insights into practices that help making data AI-ready. We also highlighted selected areas of action to be further investigated in the AI Action Plan for Switzerland.

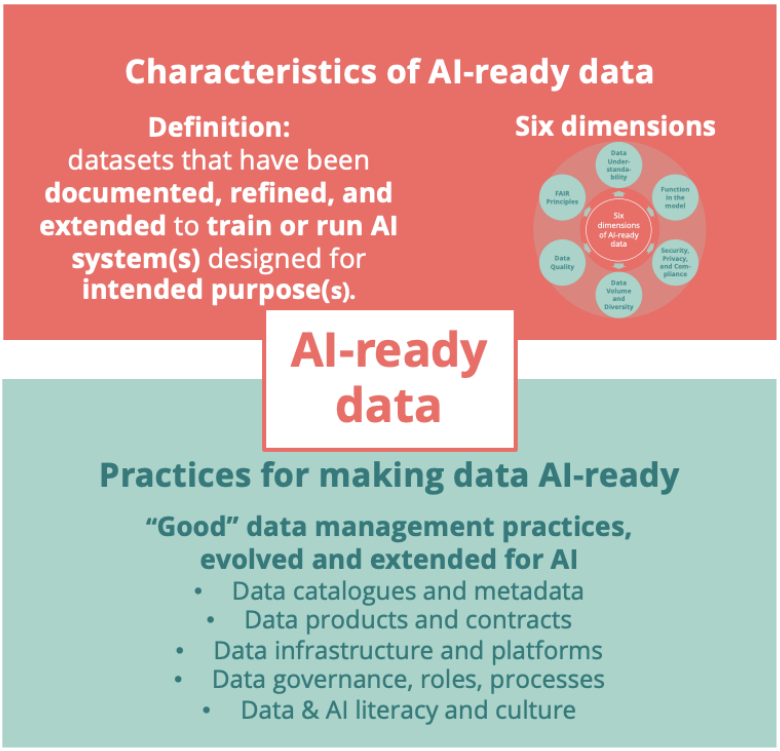


Figure 6: AI-ready data – characteristics and practices

[26] "Geneva AI Summit 2027" (https://genevaaisummit.swiss/). The theme of the Geneva AI Summit 2027 expresses the overarching ambition: enabling AI to contribute to prosperity and progress while ensuring that it is used in a trustworthy, responsible and safe manner. At this level, innovation and trust are not competing objectives. They define the two sides of the same bridge. Innovation without trust will not be sustainable; trust without the capacity to innovate will remain without practical effect.

5.2. LEARNINGS RELATED TO AI-READY DATA DIMENSIONS

During the work evidenced in this white paper, the project team has been able to find extraordinary use cases, showcasing not the key characteristics of AI-ready data as well as the underlying data work. An important learning from the use cases is that AI-ready data leverage the wealth of data that organisations have collected over the last decades. With generative and agentic AI, these data become more accessible and are the core of innovative. Yet, even well-maintained datasets need specific preparation for their use in AI systems.

- **Data understandability**: The use cases demonstrate that metadata is key for any AI system, and that agentic AI and communication between the agents (e.g., "A2A") push for machine-readable metadata. Unsurprisingly, a key theme in the use cases is metadata, with increasing emphasis on semantic definitions, but also for facilitating traceability and data governance which are requirements for high-risk AI systems under the EU AI Act.

  Recommendations:
  - Provide comprehensive metadata documentation and leverage existing metadata standards, such as ML Croissant or the Open Data Product Specification.
  - Introduce semantic definitions (semantic models, knowledge graphs) so that identity, relationships and changes to vocabularies are explicit in the data, not only described in documentation.
  - Document accountabilities and permissions/restrictions as governance metadata.
  - Shift from static data catalogues to dynamic, machine-actionable metadata.
  - Record changes to vocabularies and classification schemes (versioning), so that AI systems interpret data consistently across time.

- **FAIR principles**: Findability and interoperability are key for reuse of data in AI systems. Here, we find that AI-ready data also heavily depend on shared reference data that are maintained outside the individual dataset. Where such identifiers and reference data are not available or not systematically maintained, interoperability cannot be established through dataset-level documentation alone.

  Recommendations:
  - Use persistent identifiers to allow datasets from different sources to be linked reliably and consistently over time.
  - Publish data through "MCP" to make it accessible for "A2A" consumption.
  - Use open formats such as the Swiss open-source standard "bitmark" to support the transformation of unstructured or physical only content into AI-ready data for retrieval, grounding and reuse across AI applications.

- **Data quality**: AI amplifies previous data quality issues on the input side. As illustrated in the use cases, it also challenges the existing data quality practices, because agentic systems operate on data in real-time. Data quality therefore must be embedded in pipelines and platforms and become a continuous and automated effort. As AI systems do not report failures, but rather provide hallucinated answers, output verification and validation become a key concern.

  Recommendations:
  - Ensure continuous and automated evaluation, directly built in data pipelines and platforms.
  - Validate data against explicit, machine-checkable rules on every update, and report rather than silently drop records that fail.
  - Build validation and verification into the AI systems, for instance using validation and verification agents.
  - Test your AI system for verifiability (e.g., a practical test for any AI output: can one show which data it rests on, where they came from, and whether anything was received but not used?).
  - Apply the same validation to data produced by AI systems (e.g., automated classification) as to other data and keep such data identifiable as machine-generated.

- Increase lineage and traceability: record source, date and method for each fact, including whether it was produced by an AI system.

- **Volume and diversity**: Many of the observed use cases built on existing curated databases that were created and curated over many years but still must ensure relevancy, completeness or representativeness of the datasets. Others – like Swiss Life or Apertus – had to invest substantial resources to create a data foundation from the outset, going as far as to digitise a full physical archive and transform massive volumes of unstructured data into AI ready data. This highlights that for broader applicability of AI, there is a need to increase the number of AI-ready datasets that can be used for training and at inference by transforming yet unstructured or even only physically available documents.

  Recommendations:
  - Manage unstructured data systematically, with specific focus on completeness, consistency, versioning and validity.
  - For the broader applicability of AI use cases also yet physical only available data should be considered for context and inference and be transformed (digitalised) to AI ready data sets.

- **Security, privacy and compliance**: The use cases, specifically Apertus, illustrate the significant burden to achieve compliance with relevant data privacy and AI regulations. Additional challenges are licenses, copy-right and other author rights, and the required management of access rights and permissions.

  Recommendations:
  - Implement authentication and authorization, access rights and restrictions for AI-ready datasets.
  - Keep provenance and validation records as evidence of data governance and traceability, e.g., for high-risk AI systems under the EU AI Act.

- **Function in the model**: Several use cases, specifically in Vocational Education and the 4LM legal chatbot on Legilux, demonstrate the importance of semantically structured content models and metadata-rich representations for generative and agentic AI systems.

  Recommendations:
  - Maintain metadata (e.g., by automatically extracting and enriching content-descriptive metadata) for efficient retrieval in RAG systems and for knowledge creation.
  - Create domain models in the form of a structured semantic knowledge graph.
  - Adapt data models for unstructured data in generative and agentic AI systems and organise content for AI consumption (e.g., PDF docs by chapters, paragraphs inserted structured data, e.g., a table, an image).

5.3. LEARNINGS RELATED TO PRACTICES FOR MAKING DATA AI-READY

As mentioned beforehand, during the work evidenced in this white paper, the project team has been able to find extraordinary use cases, showcasing not only the key characteristics of AI-ready data but also the underlying data work. Among the most often cited practices for preparing AI-ready data are:

- Existing **data catalogues** and **metadata documentation** need to be extended for AI-ready data.
- A "**data product**" approach[27] facilitates the preparation and consumption of AI-ready data.
- The **preparation of AI-ready data** is not a one-time effort but must be performed in stages from use-case agnostic to use-case specific data.

[27] See, for example, Big Data Value Association (BDVA)'s publication "AI-ready Data Products" (2025; https://bdva.eu/news/ai-ready-data-products-publication/), and Hasan, R. & Legner, C. (2026). "Data Products", Competence Center Corporate Data Quality (CC CDQ) research briefing (https://www.cdq.com/events-insights/publications/cc-cdq-research-briefing-data-products).

- **Data platforms and infrastructures** are essential for supporting the preparation and consumption of AI-ready data. Architecture patterns like the "Medallion architecture" support.
- **Governance policies and processes** must be implemented in data pipelines ("governance as code"), rather than published as policy documents, while orchestration is led by humans.
- Without **data & AI literacy** and culture it is difficult to achieve AI-readiness.[28]

It is also noteworthy that three interdependent streams are particularly relevant for preparing AI-ready data:

- **Technical curation**: Technical teams prepare, structure, maintain and make data available so that they can be used reliably by AI systems. This comprises work on extending data platforms and infrastructures to support AI-readiness.
- **Content curation**: Teams monitoring content – both when data are input to an AI system, and when data are an output of an AI system – preserve the meaning, context and relevance of information, including the explicit and implicit knowledge required in a particular process or use case. This comprises work on data governance and stewardship.
- **Legal curation**: Legal and compliance teams establish the conditions under which data may be accessed, combined, used, retained, versioned or deleted, with particular attention to provenance, purpose and accountability.

None of these streams is sufficient in isolation. AI-ready data emerges from their interaction.

This perspective makes it possible to formulate more concrete questions. How should IT and data infrastructures be organised so that relevant information remains accessible and controllable? And how can it be achieved that oversight leads to tangible results, permitting clear findings about where procedures must be improved, within an organisation? It becomes increasingly apparent that processes within organisations must adapt to modern needs of AI, to implement the characteristics identified in this white paper. How should responsibilities and decision rights be allocated so that governance requirements are connected to evidence from actual system operation?

How can content standards and metadata mechanisms bind context closely to an individual information object or snippet, effectively enshrining the context with the content? The work undertaken during this white paper has found use cases that contribute largely to these objectives.

How must surrounding systems be designed so that they record not only final outputs, but also the sources, tools, intermediate steps and actions involved in producing them? Organisations increasingly rely on automated tooling to organise observability of their data flows (see dimension "function in the model").

These questions show that AI-readiness cannot be achieved through a static dataset assessment alone. It requires an operational environment in which data, context and system behaviour can be observed and governed over time. Relevant measures may include central inventories of AI systems and agents, clearly assigned owners, technical controls outside the model, task-specific and time-limited access rights, traceability of individual actions and monitoring of the full execution path rather than only the result.

5.4. IMPLICATIONS FOR SHARED DATA INFRASTRUCTURES AND SUSTAINABLE FUNDING

One of the learnings is that the preparation of AI-ready data is not a one-time activity, but ongoing data work that happens in stages from use-case agnostic to use-case specific data and depends on data infrastructures.

[28] Data literacy is a strict prerequisite for AI literacy (see Swiss Academies of Arts and Sciences, 2024, Swiss Data Literacy Charter, https://swiss-academies.ch/publications/data-literacy-charta-schweiz & https://zenodo.org/records/11146023), summarised by the principle: "*No data literacy, no AI literacy*"; see Kuonen, D. (2024; "*Ohne Datenkompetenz keine KI-Kompetenz!*", https://www.netzwoche.ch/news/2024-08-28/ohne-datenkompetenz-keine-ki-kompetenz; "*Sans littératie des données, pas de littératie en IA !*", https://www.ictjournal.ch/articles/2024-09-06/sans-litteratie-des-donnees-pas-de-litteratie-en-ia) and Kuonen, D. (2026; "*Warum Daten allein nicht reichen – Data Literacy als Schlüssel zu einer lebendigen Datenkultur und erfolgreichen digitalen Transformation*", Swiss Data Alliance 2026 General Assembly, https://www.swissdataalliance.ch/news/data-literacy-nfp-77 & https://www.swissdataalliance.ch/s/SDA_MV_DL_May18_2026.pdf).

At the use-case agnostic end of this staging, data are often not prepared by the consuming organisation itself. Registers, master data, metadata services, semantic standards and interoperability services are provided and maintained across organisational boundaries, and their provision is a continuous service rather than a one-off technical exercise. Preparing data for AI at national scale therefore depends on shared data infrastructures whose operation, maintenance, quality assurance and further development are governed and funded on a durable basis. Their reliable long-term availability is a precondition not only for public-sector AI applications, but equally for the reuse of public data in research and in the private sector.

For shared data infrastructures, an assessment of AI-readiness consequently extends beyond the individual organisation. Four questions recur. (1) Which infrastructure services are used across organisational boundaries and therefore constitute a shared basic provision? (2) How are such basic services distinguished from specific services requested in addition? (3) How are operation, maintenance, quality assurance and further development secured over time, and not only initial development? (4) And how are responsibility for an infrastructure, its operation, its use and its costs allocated consistently? In a federal state, these questions also concern the roles of cantons and communes, which use national data infrastructures alongside the Confederation. In Switzerland, these questions are currently being taken up in the context of the federal data strategy and ongoing work on the national data ecosystem.

6. Next Steps on "AI-ready Data" in the AI Action Plan for Switzerland

AI will only be as trustworthy as the data it learns from and acts upon. For Switzerland, this is not a constraint – it is an opportunity. Our vision: by making AI-readiness a shared and assessable quality of data, Switzerland becomes a place where trustworthy AI is built on trusted data – strengthening the competitiveness of Swiss companies and the quality of services in research, education and the public sector.

Today, the ambition to use AI is widespread, but most organisations do not yet have data that are fit for it. Few start from scratch: their data sit in established systems, archives and processes. The use cases in this white paper show that AI-readiness can nevertheless be achieved – by preparing data upstream, as for Apertus; by unlocking historical knowledge with the help of AI, as in the Swiss Life case; or by establishing readiness at the point of use, as Samedo demonstrates in a small company without replacing its systems.

The next step within the AI Action Plan for Switzerland (see Table 2) is the development of an assessment framework for AI-ready data, building on the definition and six dimensions presented here. The framework should reflect maturity levels rather than a pass/fail logic and meet organisations where they stand. It should enable data providers to demonstrate the AI-readiness of their datasets, help data consumers to evaluate them, and show decisionmakers in business and government where investments in data create the greatest value.

We invite companies, AI practitioners, public authorities and researchers to contribute further use cases, to test the framework in practice and to help shape it. The authors of this report and leaders of the Working Group "AI-ready Data" will coordinate these activities and serve as main points of contact.

AI Action Plan for Switzerland – AI-ready Data (https://ai-actionplan.ch)		
Actions	**Objectives**	**Selected Deliverables**
1) **Define and provide an assessment framework for AI-ready data**	Support the availability of high-quality datasets and strengthen the production of AI-ready data, by providing common foundations and an assessment framework.	- **Foundations of AI-ready data** (this white paper) - **Assessment of AI-ready data** – Checklist - Target group-specific recommendations (**user guides**)
2) **Promote secondary use of data to train or run AI system(s)**	Enable and promote the responsible, secure, and efficient secondary use of data for training, testing, operating, and improving AI systems across business, research, education, and the public sector.	- **Legal framework** for enabling secondary use of data, incl. AI systems ("*Rahmengesetz für die Sekundärnutzung von Daten*") - **Specific requirements** for secondary use of data for AI: Standardised metadata and interoperability approaches, mechanisms for secure data access and sharing, etc.
3) **Standards and accessible platform for public sector data across federal, cantonal, and municipal levels**	Promote the documentation and showcasing of relevant use cases, datasets, and practices related to AI-ready data across business, research, education, and the public sector in order to promote reuse, collaboration, learning, and responsible AI adoption.	- Define **metadata standards** that improve use of up to date, high quality, and linkable data in AI systems, starting with a core "**minimal viable metadata standard**" (e.g., per sector) - Specify **semantic standards** to harmonize data - **Extend (data) platforms** to support AI systems

4)	**Promote hosting of international Trusted Data Observatory in Geneva**	Position the Trusted Data Observatory (TDO – https://www.trusteddataobservatory.org/), a global metadata platform for data discovery, as part of the AI Summit 2027 in Geneva to strengthen the Swiss moderation role globally.	- Agreement on a **Minimum Viable Metadata (MVM)** set, associated standards, technical specifications, and Proof of Concept (PoC) participants - Development of a **scalable TDO prototype**, implementation of the PoC, training of participants - Review
5)	**Promote use cases for AI-ready data**	Document and showcase relevant AI-ready data use cases, datasets, and practices across business, research, education, and the public sector to promote reuse, collaboration, learning, and responsible AI adoption.	- **Documentation model** for AI-ready data use cases: template (with datasets and application contexts) - **Collection of AI-ready data use cases** (e.g., classification by sector, domain and maturity level) – in collaboration with data initiatives in domains

Table 2: Five Actions for "AI-ready Data" in the AI Action Plan for Switzerland

Appendix: AI-ready Data – Literature Review

The Competence Center Corporate Data Quality (CC CDQ) has conducted a systematic literature review on AI-ready data, analysing more than 100 sources, including academic publications, reports from analysts, consultancies, and solution vendors and analysts as well as professional associations. Some of the most relevant sources are displayed in the following table:

Source	Definition	Characteristics and assessment
Big Data Value Association (2025): AI-ready data products	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**.	Technical aspects of an AI-ready data product: (1) **Data quality for AI:** continuous and automated evaluation of data quality, model-aware quality checks (tailored to the specific machine learning models), adaptive and intelligent data quality solutions (addressing semantic and contextual nuances) (2) **Preparing AI-ready data:** Feature optimization, data enrichment, data augmentation, synthetic data generation
Gartner (2025): What Is AI-Ready Data? And How to Get Yours There	**AI-ready data** means that your data must be **representative of the use case**, of every pattern, errors, outliers and unexpected emergence that is needed to train or run an AI model for a specific use. It's a **process and a practice** based on **availability of metadata** to **align, qualify and govern** the data.	Three main questions to prove data meets AI requirements: (1) **AI use case** requirements (2) **Data goverance** requirements (3) **Data qualification** to meet AI-expected confidence requirements
Open Data Institute Research (2025): A framework for AI-ready data, Version 2.0 (December 2025)	The **AI-readiness of a dataset** over four components: its **technical optimization for machine learning**; its **overall quality and adherence to standards**; its **legal compliance**; and its **responsible collection**.	Assessment framework comprising four categories, covering data-related properties as well as governance practices and infrastructure, with criteria and sub-criteria: (1) **Dataset properties** (e.g., adherence to standards, semantic and source consistency, class and source imbalance, de-identification, format) (2) **Metadata** (e.g. machine-readable metadata, technical specification, supply chain information, legal and socio-technical information) (3) **Surrounding infrastructure** (e.g., accessibility, version control and monitoring), (4) **Governance** (e.g. policy-as-code, roles and responsibilities, identifiable points of contact, data access processes)
Hinduma et al. (2024): AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI, in: Proceedings of the 36th International Conference on Scientific and Statistical	AI readiness of data = readiness of data for AI processes (machine-learning pipeline), characterised by seven categories of data readiness for AI	Seven categories of AI-readiness for AI: (1) **Data Quality** (Completeness, outliers, deduplication) (2) **Understanding and usability of data** (Metadata availability and quality, provenance) (3) **Data structures and organization** (Data types used, data schema quality and consistency, file format and storage system)

Database Management (SSDBM '24), https://doi.org/10.1145/3676288.3676296		(4) **Data value** (Feature importance, label quality, data point impact, uncertainty quantification) (5) **FAIR and Unbiased** (Class imbalance, Bias mitigation) (6) **Data governance** (Security, privacy, collection) (7) **AI application-specific metrics** (model-specific metrics)
Wang, J. & Legner, C. (2025): What makes data AI-ready? In: in: Proceedings of the European Conference on Information Systems (ECIS 2025), TREO Talks	AI-ready data is data that has been **documented, refined, and extended to train or run the AI-based system**, which are designed for specific intended purpose(s). The underlying data work – incl. annotation, curation, and preprocessing - is routinized and standardized to ensure compliance and continuous performance improvement of the AI-based system.	Six dimensions (derived from the literature review): (1) **Data Understandability**. (2) **Findability, Accessibility, Interoperability, Reusability (FAIR)** (3) **Data Quality** (4) **Data Volume and Diversity** (5) **Prepared to Intended Purpose(s)** (6) **Security, Privacy, and Compliance**

Table 3: AI-ready data in existing literature

