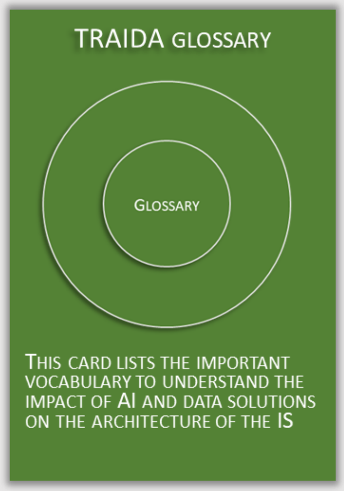
TRAIDA GLOSSARY

To increase your speed of spreading a culture of AI and data management that is understandable by all of your technical and business teams, it is essential to establish and share a glossary of AI and data solutions terms. Although popular, some of these terms do not always have a definition commonly recognized by the market. You will therefore need to decide on your vocabulary choices. This card gives you the starting point for this semantic work, which is fundamental to building and managing your transformation with AI and data management.



The definitions are customized for the TRAIDA framework. They are not intended to conform to the marketing presentations of software vendors or IT analysis firms. Based on these definitions, you can create your own company glossary and update the various cards of the TRAIDA framework according to your context. **However, the more you maintain definitions that are neutral in relation to marketing trends, the more comprehensible your AI and data solutions strategy will be to your stakeholders, and the more robust it will remain over time**. The worst scenario would be to introduce terms and definitions that change too frequently and are challenged by the marketing and sales rhetoric of solution providers, whether they are technology companies or consultants. By relying on the most neutral definitions possible, TRAIDA helps you build a stable communication strategy for AI and data solutions in your context.

D

Data fabric, data hub and data mesh (overview)

Data fabric and **data hub** are complex to define precisely, as different software vendors encompass various concepts within these terms. At TRAIDA, we prioritize identifying the needs of the three fundamental repositories regardless of the chosen data fabric and data hub solutions: Master Data Management (MDM), Operational Data Store (ODS), and Enterprise Knowledge Graph (EKG). No single technology can universally manage these three repositories on the same platform. **To choose the best data fabric and data hub tools for your context, it is important first to have a clear understanding of your needs in MDM, ODS, and EKG** (refer to the respective TRAIDA cards). It is based on these needs that scaling AI and data solutions will be properly managed. Otherwise, you risk selecting technological solutions that are suitable for an initial deployment but not appropriate for scaling AI and data management solutions.

The term **data mesh** is relatively straightforward to define, as it refers to a data architecture that organizes data by business concepts to reduce silos (micro databases).

Data fabric

A data fabric is a comprehensive set of technologies designed to streamline data integration processes, including referencing data sources (data sets), data cleaning procedures, and unifying data structures through semantic modeling. It relies on robust metadata management systems and often uses graph knowledge database technology.

Modern data fabric supports the configuration and testing of AI decision-making algorithms (such as machine learning, AI training, and rule-based systems), as well as the deployment and monitoring of AI processes and data in production environments.

While a data fabric can assume some roles of a data hub (data integration), its primary focus is to enhance data and AI governance at scale. Rather than replacing MDM (Master Data Management), ODS (Operational Data Store), and EKG (Enterprise Knowledge Graph) repositories, it should coordinate them. **However, the overlap between a data fabric and core repositories like MDM, ODS, and EKG must be carefully evaluated before deciding on large-scale deployment**.

In a data mesh context, a data fabric can also offer additional features for controlling micro databases, such as data caching, inter-database transactions, workflow management, and support for long transactions.

Data hub

A data hub primarily functions as a data flow integration bus, incorporating technologies like EAI (Enterprise Application Integration), ETL (Extract - Transform - Load), and ESB (Enterprise Service Bus).

Depending on the solution, a data hub can manage metadata (mainly at the flow level), map IDs across silos, visualize unified data, and store certain operational data akin to an ODS (Operational Data Store).

Coupled with a data mesh approach, it can also handle data caching and long transaction management.

While some vendors market data hubs as universal data management platforms, they often fall short of fully implementing MDM, ODS, and EKG systems. It's typically more effective to use data hubs for integrating data flows and supplement them with dedicated solutions for MDM, ODS, and EKG.

More generally, the concept of a data hub is gradually being absorbed by the broader concept of a data fabric. We can therefore say that a data fabric solution either natively includes or integrates with a data hub solution. Open-source offerings facilitate this kind of integration, which should be carefully considered when selecting tools.

Data mesh

Data Mesh is a data architecture approach that organizes data by business domains or concepts, rather than by functional or organizational silos. It uses semantic modeling and a technical infrastructure to manage transactions between business concepts spread across different micro databases.

Data Mesh enhances data governance and reduces data duplication. It is a set of architectural principles rather than a specific technology. Implementing a Data Mesh requires leveraging data fabric and data hub technologies, tailored to the specific context of each company.

E

Enterprise Knowledge Graph (EKG)

The Enterprise Knowledge Graph (EKG) is a repository specialized in knowledge accumulation. It manages both structured and unstructured data, with the capability to receive data sources without requiring prior modeling. It is based on the technology of knowledge graph-oriented databases.

Unlike MDM, it does not have as advanced governance processes; and unlike the ODS, it does not offer as powerful transactional management (OLTP).

Depending on the context of each company, it is necessary to find the best combination between the needs of the MDM, ODS, and EKG. However, it is important that all these repositories share the same ontologies to avoid the negative effects of siloing. For a small or medium-sized enterprise, it is feasible to manage everything within a knowledge graph-oriented database, that is, within the EKG.

On the other hand, for a larger information system supported by a rationalization policy, it may be necessary to opt for three different technologies for the MDM, ODS, and EKG. The worst approach would be to implement as many EKG repositories as there are functional domains without considering the cross-functional needs of the MDM and ODS. This would lead to siloed EKGs with associated quality issues. In such a case, large-scale AI integration within the enterprise would be compromised.

Each of these repositories—MDM, ODS, and EKG—is covered by a dedicated TRAIDA card.

M

Master Data Management (MDM)

Master Data Management (MDM) is a data repository specialized in managing reference and master data. These are the most shared data between applications. Their lifecycle is less rapid than that of transactional data.

The strength of MDM lies in its agility to accommodate changes in the structures of reference and master data, and in the richness of its data governance processes: quality, security, traceability, data entry UI, reporting, version and variant management, workflow, etc. MDM is also the preferred repository for creating a metadata catalog, which benefits from the full power of its governance. In this context, the MDM includes descriptions of the ontologies managed within the company, which form the core of the semantic platform recommended by TRAIDA.

MDM works in collaboration with the Operational Data Store (ODS) and the Enterprise Knowledge Graph (EKG). Each of these repositories—MDM, ODS, and EKG—is covered by a dedicated TRAIDA card.

O

Operational Data Store (ODS)

The Operational Data Store (ODS) is a data repository specialized in the unified management of operational data. It provides a unified access point to data from multiple sources, meaning data located in heterogeneous databases (silos). Unlike MDM, the ODS deals with transactional data, which has a rapid lifecycle. It is therefore specialized in transaction management and does not offer governance processes like those for reference and master data.

A vertical implementation of the ODS for a specific functional domain leads to the concept of a data hub, such as with Customer Data Integration (CDI) or Product Information Management (PIM). This siloing results in unnecessary data duplication, increasing the risk of poor data quality.

In TRAIDA, the embodiment of ontologies in the semantic platform does not rely on such verticalization. Instead, it is advisable to build a solution that establishes a single ODS. This ODS then works in close collaboration with the MDM, which provides a central access point to reference data, master data, and metadata. Each of these repositories—MDM, ODS, and EKG—is covered by a dedicated TRAIDA card.

Ontology

An ontology is a structured representation of a domain of knowledge. It is based on these four fundamental properties:

1. Exhaustive: The entire semantics of the domain is expressed in the form of concepts and relationships.
2. Unified: There is no redundancy.
3. Explicit: There is no ambiguity.
4. Universal: It is independent of information processing technologies.

Maintaining these four properties throughout the lifecycle of a domain is challenging. Thus, an ontology is a living representation that evolves to improve and accommodate changes. It is necessary, therefore, to plan for version management, variants, and the impact of ontology deployment, which means adopting appropriate governance.

To achieve such a powerful representation, an ontology requires these components:

- Glossary: Unambiguous definitions of concepts.
- Thesaurus: An extension of the glossary with synonymous terms and expression equivalences according to the contexts in which the concepts are used, including multilingual contexts.
- Taxonomy: Hierarchies among the concepts.
- State Machine: The lifecycle of each concept and synchronization between concepts.
- Identifiers: Format and semantics of the identifiers for each concept.

The combination of these components, along with the procedure used for their construction, forms semantic modeling. Semantic modeling is thus the discipline that allows the construction of ontologies.

From the perspective of the tooled representation of ontology, standards such as RDF and OWL are used. It is also possible to opt for a representation with UML, which is also used for semantic modeling. To illustrate these choices of technical representations, here are two possible use cases:

- Using RDF for the ontology, then transcribing it into UML to obtain a semantic model. This model is then derived into a logical model in a database.
- Using UML for the ontology and deriving it to the logical level for implementation.

With TRAIDA, ontologies and semantic modeling form the foundation of the semantic platform that enables the construction of a digital twin of the information system. It is from this digital twin that the integration of AI systems is ensured with associated data management solutions.

S

Semantic modeling

Semantic modeling brings together the design processes for the following components: glossary, thesaurus, taxonomy, ontology, state machine, and identifiers. All of these are necessary to formalize the knowledge of a domain, such as an organization, a business, an activity, or an area of expertise. This formalization is carried out independently of any specific technological implementations.

In TRAIDA, semantic modeling is used to build the semantic platform that powers the ontologies, from which AI systems and associated data management solutions are integrated (digital twin). This approach avoids integrating AI at a lower level of abstraction, that of the physical flows of data and application systems. In most cases, these physical layers do not have the required quality to ensure reliable AI execution and the agility needed to adapt quickly enough to business requirements.