ENTERPRISE KNOWLEDGE GRAPH

The Enterprise Knowledge Graph (EKG) is a potential universal repository for knowledge management with various use cases. It is the cornerstone of the semantic platform promoted by the TRAIDA platform. It serves as the unique point of contact for all AI systems within the company (digital twin).

1. CONDITIONS OF SUCCESS

To properly train AI systems, it is necessary to gather the maximum amount of knowledge according to three levels:

1. Data available on the Internet, paying attention to usage rights. Large AI models like ChatGPT or Llama are trained on these data. As a user of these LLMs, you benefit from the training already done on large amounts of information. However, keep in mind that with an open-source LLM like Llama (Meta), you will still need to find a solution to run it on sufficiently powerful infrastructure, likely in the cloud.
2. Your company's data that already exists in your databases, office files, and physical documents (paper). This data is essential to enhance the training of LLMs in order to personalize their behaviors to your company. This is a fine-tuning task.
3. Your company's data known by your teams but not yet formalized in databases, files, or even in writing. This wealth of data is a reservoir of tacit knowledge that represents a significant percentage of the total knowledge the company possesses, around 60% to 80%. This includes the know-how of operators, how they adapt work procedures to the realities on the ground, information exchanged between actors and stakeholders to meet objectives, etc. This informal knowledge must be transformed into formal knowledge to enrich AI systems and improve their profitability.

This data is of all kinds, both structured and multimedia. It evolves with the company and requires version management. For example, a set of data used to train an AI system in an initial version may become obsolete later and will then need to be removed from the AI system's training. In other words, for each AI system training, it is necessary to keep the sources of data used and ensure that rights and security are respected.

This management is particularly delicate because the structures of the collected data are very diverse. Indeed, the training scope of AI systems encompasses the entire company. For example, starting from an internet-based LLM like ChatGPT, the company will proceed to a first level of global fine-tuning to its activity before carrying out finer settings for its different activities, such as its marketing, manufacturing, human resources departments, etc. As these activities coordinate through cross-functional processes, other knowledge will enrich AI systems to optimize operations at the boundaries of departments. Thus, it is a bidirectional movement of AI system training that operates from global to local and vice versa.

From this description, a need for data and knowledge governance (multimedia, tacit data) emerges. It must be flexible enough to instantly adapt to new data structures whose prior modeling would be too cumbersome, lengthy, and even impossible in some situations. However, it must also be able to work with metadata to classify, document, organize, manage versions, rights, security, traceability, and more for all the knowledge used to train AI systems.

Finally, a last criterion to consider is the dual mode of knowledge exploitation:

1. First, in asynchronous mode for the massive training of AI systems. Thus, the company trains its AI from the company's data as OpenAI or Meta does from internet data. This training mobilizes the maximum knowledge to personalize large LLMs.
2. Then, in synchronous mode at the time of prompt execution, through real-time enrichment of information injected into AI systems. This principle relies on RAG (Retrieval Augmented Generation) technology. It is no longer about mobilizing the maximum knowledge but the minimum necessary to help the AI better respond in the specific context of a query. For example, it may involve adding updated data from a customer relationship management database to a prompt analyzing a customer file.

Need for a Knowledge Management Repository

To address the needs described earlier, it is pertinent to set up a knowledge management repository that offers the following characteristics:

- Data injection in compliance with a predefined data model (schema-oriented) that relies on ontology modeling.
- Data injection with free loading (schema-free) and the ability to automatically generate the data schema reflecting the source data (automatic ontologies).
- Management of rules for quality and security controls.
- Management of all types of data, both structured and multimedia.
- Version management at the data schema level.
- Version management at the data level.
- Connectors for data transformation during import and export.
- Management and visualization UI for data.

This repository is both schema-oriented and schema-free. An interesting implementation can be found with knowledge graph-oriented database technology, which is the subject of this TRAIDA card. The EKG (Enterprise Knowledge Graph) repository then becomes the pivot of the semantic management platform necessary for the governance and execution of AI systems. It is also identified as the digital twin of the information system, providing a single and secure access point to all AI systems within the company.

How to Integrate the EKG with Other Data Repositories

The EKG repository should be the unique integration point with AI systems, thereby centralizing governance. This enables tracing the versions of data used to train AI systems as well as the data sources requested during real-time prompt enrichment (RAG).

However, upstream of the EKG, it is essential to consider the use of two other repositories addressed in specific TRAIDA cards:

1. Master Data Management (MDM) for reference and master data. It is the natural source for feeding the EKG with this data. It also provides the data models for ontologies. In other words, the MDM is the pivot repository for ontologies. The EKG can have its own ontologies during free data injection (schema-free) but must rely on ontologies from the MDM for schema-controlled data injections (schema-oriented).
2. The Operational Data Store (ODS) also relies on MDM ontologies. It contains operational data that are injected into the EKG as needed, both during massive AI system training processes and for RAG processes.

Data integration must be carefully designed. For example, it starts with an information modification in a production database that triggers an update in the ODS, then propagates it to the EKG to make it available to AI systems. This integration preferably relies on an event-based architecture to avoid tight coupling between subsystems. In other words, the ODS listens to a data injection stream from the application to initiate its update. Similarly, the EKG listens to a data injection stream from the ODS to trigger its own update. If the MDM repository is also involved in propagating reference and master data, the ODS will listen to a data stream from the MDM.

Is it Possible to Merge MDM, ODS, and EKG Repositories?

It is not straightforward to merge the three repositories into one, for the following reasons:

- MDM requires a transaction-oriented database with strong data typing, meaning a technology that relies on a formal data schema. This may involve an internal meta-schema to ensure sufficient flexibility when updating data structures and their relationships. Moreover, MDM is also an application system with specific governance functions for business teams that are not found in ODS and EKG solutions.
- ODS is an operational data repository that unifies structured data from application systems. It requires schema-oriented database technology to guarantee integrity and transactional management for large volumes and high-frequency access (multi-user in parallel).
- EKG is a more flexible repository that benefits from a schema-free approach to absorb multiple types of data beyond structured data. As indicated in this TRAIDA card, it provides a relevant solution for accumulating the knowledge necessary for AI system execution.

Depending on the technological quality of the software products used, the following integration scenarios are possible:

- For a company with data volumes and transactional requirements compatible with the processing power of the MDM repository, an additional ODS is unnecessary. The MDM also serves as the ODS. In this scenario, the company has two repositories: the MDM, which also handles ODS, and the EKG.
- Conversely, a company with high data volume and transactional demands and a simple governance requirement for its reference and master data can use the ODS as an MDM repository. In this scenario, the company has two repositories: the ODS, which also handles MDM, and the EKG.
- If the EKG technology is robust enough to handle the company's structured data volumes and transactional requirements, it is conceivable to use it as the single repository that also addresses governance needs for reference and master data (MDM) and operational data unification from application systems (ODS). As of the writing of this TRAIDA card and to our knowledge, such technology does not exist. Indeed, graph-oriented databases are not well-suited for multi-user transactional management on large volumes (ODS) and lack business governance functions for reference and master data (MDM).

The TRAIDA MDM and ODS cards provide a more detailed description of these two repositories.

For small businesses (start-ups, SMEs), it is also conceivable to implement a NoCode database (such as Knack or Airtable) to handle the three repositories MDM, ODS, and EKG within the same technology. If the company is starting its information system, it is even possible to consider the ODS as the core system data. In other words, operational processes are directly built around the ODS without needing to install application systems like ERP.

How Does a Knowledge Graph Database Work?

The technology behind knowledge graph databases implements a meta-schema for data storage based on the description of triplets: subject, predicate, object. For example, a customer (subject) buys (predicate) a product (object). Therefore, to inject data into the database, it is not necessary to model data structures beforehand. This meta-schema leads to a schema-free mode of operation, which is very agile for data manipulation. Conversely, in the absence of a data model, there are few rules for data quality verification. However, these rules can be added in addition to the triplet meta-schema. Thus, the knowledge graph database reconciles the flexibility of schema-free operation with the power of quality controls configured according to the use cases implemented in the company.

However, this flexibility generally comes with weaknesses in response times with large volumes, particularly during queries that traverse multiple nodes (equivalent to table joins in relational databases), and in cases of massive concurrent and transactional access (difficulty in guaranteeing ACID criteria: Atomicity, Consistency, Isolation, and Durability).

As mentioned earlier in this TRAIDA card, the state-of-the-art in knowledge graph databases offers an advantage for integrating AI systems, but they do not yet have the sufficient capabilities to serve as ODS and MDM.

2. IMPORTANCE OF THIS CARD FOR YOUR TRANSFORMATIVE AI

Establishing a unified knowledge repository exposed to AI systems is an indispensable condition for mastering AI and managing data at the enterprise level. Without this repository, point-to-point calls between AI systems and core-system databases (applications, software) would be necessary, as well as all other information sources like files and other documentation. The quality control of data used by AI and their traceability would be compromised. To avoid this point-to-point mode, it is essential to build an EKG repository. It forms a central element of the semantic platform adopted by the TRAIDA framework, in conjunction with the MDM and ODS repositories.

This unified repository thus enables the implementation of security rules, traceability, version management, and more, across all the knowledge used by AI systems.

PERSONAL AND COLLECTIVE KNOWLEDGE ACCUMULATION

An important aspect of the successful and profitable deployment of AI in the enterprise is the ability of stakeholders to document their tacit knowledge and elevate it to a collective level. According to most studies, this knowledge represents about 60% to 80% of the information used by the enterprise. Therefore, a comprehensive program must be initiated to support stakeholders in formalizing their know-how in writing. In the Engage-Meta community, this work relies on the WASI process for Write, Analyze, Share, and Innovate (see the community website for more information).

Of course, all these writings must be stored, versioned, quality-controlled, shared, and used to train AI systems. Therefore, a knowledge storage repository is necessary, and naturally, the EKG stands out as the solution.

This knowledge is stored according to two classifications: either in an ontology specifically created for the classification of knowledge, for example, following the main functions of the company such as marketing, sales, production, etc., or in the business ontologies already in place to operate the organization's

processes. In this case, it involves attaching the knowledge directly to the business concepts that form the ontologies.

This new and decisive approach to formalizing tacit knowledge is an ongoing, daily activity that the organization must adopt. Strategies for classifying new knowledge in the EKG may vary depending on use cases and cultural practices within the company, such as in terms of information sharing. However, in all cases, the goal is to have the best possible and up-to-date knowledge to train AI systems and thereby increase their intelligence and value-creation potential.

ONTOLOGY MANAGEMENT (ANALYTICS, OLTP)

The basic principle to remember is the unification of ontologies across the three repositories: MDM, ODS, and EKG. In other words, the list of business concepts, their hierarchies and relationships, as well as their life cycles (business states) are shared by the three repositories. As mentioned earlier, it will be necessary to decide on the pivot repository for the ontologies, the one capable of managing their versions. This is generally the MDM. The EKG, which powers these ontologies shared within the company, resembles the OLTP usage mode.

Next, given the flexibility of the EKG, it can be used to generate tactical or temporary data analysis ontologies. Indeed, when dealing with a data set whose underlying data structure is unknown or partially compatible with the shared ontologies, it is useful to load them in schema-free mode into the knowledge graph database. This allows for data discovery by navigating through the information triplets, and even automatically generating an ontology based on the injected data. This process helps in better understanding the data and even calculating the differences between the default ontology obtained and those officially shared within the company. The EKG that powers these automatic and non-shared ontologies within the company resembles the Analytics usage mode.

How to Build Shared Ontologies?

In the TRAIDA core-system data card, we describe the process of creating a draft of shared ontologies through AI analysis of existing databases. Further construction work is then initiated to build and maintain the shared ontologies within the company, taking into account the needs for optimization, automation, and value creation. This work requires expertise in semantic data modeling, which can be provided by AI systems specialized in this field. In the TRAIDA Initial Engagement offer, each step of data and process modeling is accompanied by an AI assistant under ChatGPT, helping teams produce the ontologies. Here is an excerpt from the work process, with the full description available on the Engage-Meta community website:

- Process
- Business Concepts
- Business Concepts Control
- Ontology
- Data Modeling
- Identifiers Design
- Business Concepts States

- Process Modeling Refinement
- Integration
- Database Implementation
- Process Implementation
- Security Policies
- Governance
- Review

REGULATORY MANAGEMENT

The EKG is very powerful for implementing regulatory compliance monitoring based on these basic principles:

1. A regulation describes rules that the company must apply to different business concepts and processes.
2. The company implements these rules and references them in documentation.

When regulations are extensive, they can include dozens of rules whose implementations in the company's applications can result in hundreds of impact points. Beyond the initial system compliance, the company must find a way to monitor changes in both the regulations and the application systems, whose updates may render the impact points obsolete.

To establish this governance, the EKG repository is used as follows:

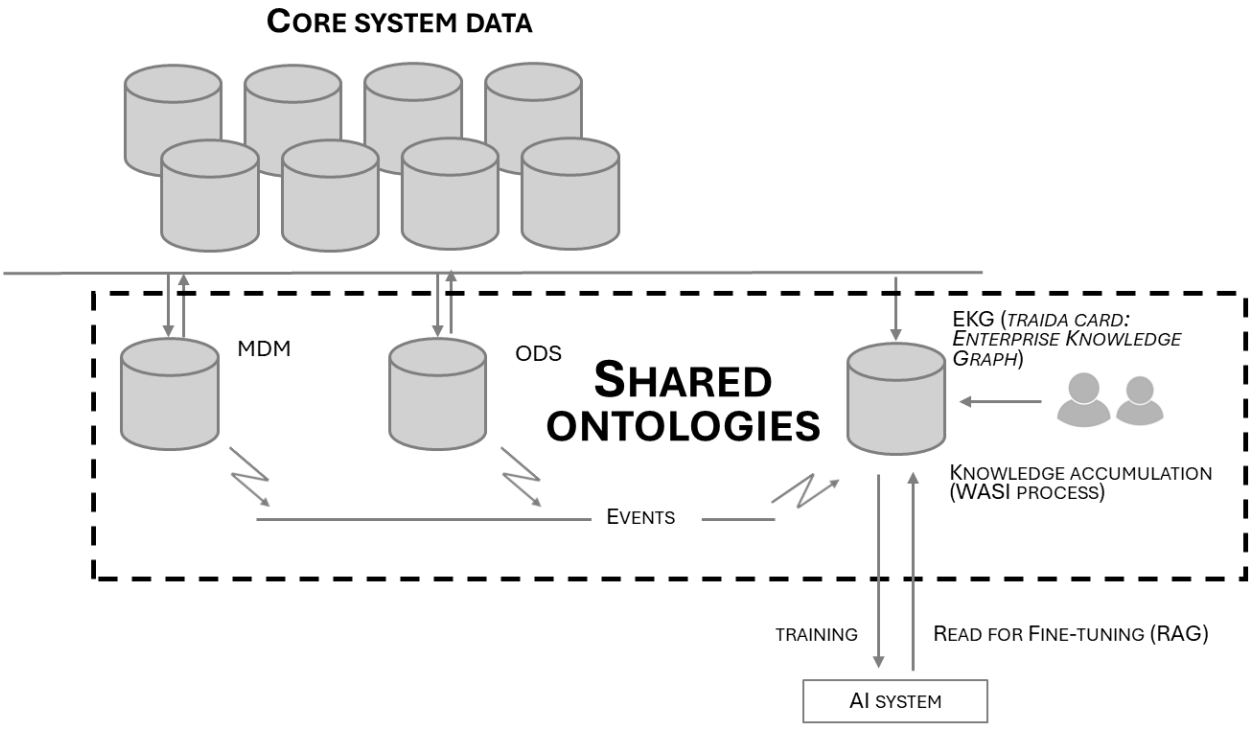
- The regulatory text is analyzed by an AI (LLM) to inject it in the form of triplets, for example (Rule, Objective, Business Concept), into the repository.
- A regulatory manager verifies and enhances the relevance of the triplets and complements them with references to the impact points in the company's application systems and processes.

Once this repository is in place, it is used for regulatory monitoring at two levels:

1. The documentation of the application systems and processes, as well as their IT descriptions (source code, data flows, etc.), are processed by an AI (LLM) and injected into the EKG repository to calculate discrepancies between the intended impact points and the reality of the systems. This is an audit operation to ensure regulatory implementation compliance.
2. When a regulatory update arrives, a new repository is built using the same process as described previously. An analysis of the discrepancies between the two versions of the regulations then allows for automatically identifying the impact points to be considered.

Without the power and flexibility of knowledge graph databases and LLMs, it would not be possible to automate the governance described above. Only manual management would be feasible, with the associated costs and risks of non-compliance errors.

3. BLUEPRINT



4. YOUR SITUATION & OBJECTIVES