Enterprise knowledge management

How to drive your DataOps framework with a single layer of business knowledge that supports self-service, data governance, AI, natural language processing, and accelerated data lake generation.

Mike Nicpan (ING)  Ron van der Starre (IBM)
Stefhan Van Helvoirt (ING)  Pat O’Sullivan (IBM)
The context for knowledge management

The past 10 to 15 years have challenged companies to improve their data management. After the 2008 financial crises it became clear that the information available for decision makers and auditors was not adequate and often it was not even clear what the data represented.

Moreover, regulatory bodies have continued to tighten rules and propose new legislation that affects many industries. Financial institutions have had their capital requirements changed via Basel IV and Solvency II, and regulators also increasingly demand a clear view of how data was used to produce a specific outcome. GDPR then extended regulatory considerations to virtually all organizations doing business in or with the EU. In parallel, the advent of big data has required management and governance of large amounts of unstructured—or badly structured—data. Finally, the general competitive climate means organizations need to be more agile and adaptable when leveraging their data to achieve business needs.

All these factors are driving businesses to manage their data with modernized information architectures (IA) and software products. This is a good start, but it may still not be enough. That’s why IBM worked together with ING Bank on a new information architecture for a wide variety of regulatory and analytical purposes: “The Governed Data Lake Reference Architecture”. This brought together existing technologies and added a big change: “Governance by Design”.

Companies that want to become data driven and implement successful governance must go beyond data management and take a comprehensive approach to managing their enterprise knowledge. Enterprise knowledge management is not just about the ongoing management of various important assets across the organization; it also offers new opportunities. This ebook elaborates why you should achieve knowledge management, how it is constituted, and how DataOps plays a crucial role.
Enterprise knowledge management is about new opportunities, not just the status quo.
In addition, even when an organization breaks down silos of knowledge management within one area, such as analytics, there is no attempt to consider how such knowledge could be used elsewhere in other functions. For example, the knowledge gathered as part of data governance processes is typically not reused to support the enterprise’s digital transformation.

In contrast, comprehensive knowledge management establishes a semantic and technical foundation to drive business and technology choices. This way of working also helps restore the dynamics between business and technical personnel. Both parties need to collaborate seamlessly, meaning control and ownership needs to be vested in the party that is best equipped to take responsibility. When business users can influence the technological landscape and means by which information is managed and governed, the technology part of the organization can focus on strengthening the capabilities that provide this freedom and control.

Despite technological advances, including new machine learning capabilities for tasks such as data discovery or data profiling, many enterprises still rely on labor-intensive and complicated processes for setting up and maintaining their data governance and data management ecosystem. A single change often needs to be replicated across many different models and systems with different vendors and technologies. This lack of automation and integration means that the organization can’t react quickly to changes in their business and technical environments.

Another difficulty arises in promoting communication and consistency between technical and business users, given that in many cases these different users are presented with separate “versions of the truth” and certainly separate levels of detail and abstraction.
A single semantic layer of business knowledge

The essential feature that separates comprehensive knowledge management from simple data management is a semantic layer of business and technical knowledge. This layer establishes the basis for data management, information governance and data operations. This single layer of knowledge is also the basis for self-service activities by a variety of personas, and it provides support for new range of use cases leveraging artificial intelligence (AI), machine learning (ML), and natural language processing (NLP). New use cases can include auto profiling, auto mapping, auto classification, automated model change propagation, and automated pattern discovery and interactions.

ING and a number of other clients have worked with IBM to define the optimal set of industry-specific predefined content that not only makes the knowledge management landscape more easily accessible but would also make it better managed, less complicated, more adaptable and more representative of best support for this range of use cases.

Organizations that take a unified and holistic approach to knowledge management can benefit from:

- Labor cost savings associated with the day-to-day processing and management of the data lake
- More timely reactions to business and technical changes
- Reduction of risk and improvement in lineage and provenance
- Tighter integration due to the evolution of the enterprise’s common vocabulary—this provides the semantic underpinning not just of systems of insight but of the broader systems of record and systems of engagement as well
- A richer set of capabilities such as NLP queries for both business and engineering users

Validating the enterprise vocabulary

A data steward can help ensure the integrity of a continually evolving central vocabulary using a set of predefined validation capabilities to enforce alignment with a specific metamodel.

Generating the data fabric semantic layer

A data engineer or ontologist helps ensure that the business knowledge in their catalog is deployable to a data fabric using ontology generation to convert the catalog vocabulary into a format suitable for use in knowledge graphs.

Generating logical data models

A data engineer or data modeller helps ensure all data modeling activities are directly aligned with the business vocabulary by exploiting data model generation capabilities to support central data warehouse and data mart creation.

Generating API specifications

A data engineer can drive business consistency in their digital transformation by using OpenAPI JSON generation to provide standardized business constructs to underpin API development.
Auto-generation of DataOps artifacts

One major longer-term advantage is the opportunity to generate artifacts based on the semantic knowledge that is captured in the data catalog. For example, this would mean that the typical data modeling process starts with the business users leveraging the curated semantic knowledge to define the broad scope to be addressed. Then, during a subsequent generation process, that business-driven scope can be further refined with technical details and choices being made by data engineers or data modelers.

This reuse of semantic knowledge helps ensure that information only needs to be captured once, fostering maintainability and automating the propagation of changes. In addition, because of the focus on generation or regeneration of these artifacts, it also becomes more attractive and feasible to establish an ecosystem that is able to adjust itself and cope with ever-changing needs and circumstances. The pain of having to start over is reduced with every generation and automation option that can be applied.

By abstracting from the underlying technologies and platform and capturing knowledge conceptually, a more adaptive technological landscape arises. The more centrally this knowledge base is set up, the easier it is to connect more components and drive further automation and integration.
Data management, governance and operations are built on a single semantic layer of knowledge.
The components of the semantic layer

Several models or other semantic structures form the key components of the central knowledge base, as shown in Figure 1.
In addition, this central knowledge base can be split into three areas of knowledge that relate to the areas of usage and ownership of knowledge by different parts of the organization:

**Enterprise knowledge**
The set of information elements providing a common framework of knowledge across the enterprise, such as the enterprise-wide specification of a product hierarchy.

**Departmental knowledge**
Sets of information elements intended for use by specific departments within the enterprise, such as a local glossary of business terms.

**External / supportive knowledge:**
Sets of information elements representing the regulations and other external sources of business metadata pertinent to the enterprise, such as a GDPR Regulation Taxonomy.

The main types of knowledge would typically include:

**Business knowledge**
The range of business-oriented metadata needed by the enterprise. A subset of this—for example, the hierarchy of the different types of products relevant to the business—is explicitly exposed to run-time users. However, business knowledge may also extend to external references held by the enterprise but not yet exposed to the majority of business users, such as third-party models, open source models, standard models and any local customer models or glossaries.

**User knowledge**
All the user-specific, user-related knowledge; may include access and security information. This also includes the "tribal knowledge" of the previous choices, searches and other behavior of groups of users for matrix factorization. User knowledge is updated on a constant basis as users—or the apps and APIs they are using—interact with the knowledge base.

**Data design knowledge**
The store of all associated information that needs to be stored to support the knowledge base from a technical perspective, or knowledge that is needed to support the generation of physical artifacts. For example, this category could include the specific additional structures and rules needed to enable generation of DB specifications or OpenAPI JSON specifications.

**Implementation knowledge**
The representation of all the physical metadata of the current data lake environment. This is refreshed on a regular basis to reflect changes or updates to that data lake. This is the set of more technically oriented metadata that accurately reflects the content of the underlying data lake or data warehouse.

In addition, this central knowledge base can be split into three areas of knowledge that relate to the areas of usage and ownership of knowledge by different parts of the organization:

**Enterprise knowledge**
The set of information elements providing a common framework of knowledge across the enterprise, such as the enterprise-wide specification of a product hierarchy.

**Departmental knowledge**
Sets of information elements intended for use by specific departments within the enterprise, such as a local glossary of business terms.

**External / supportive knowledge:**
Sets of information elements representing the regulations and other external sources of business metadata pertinent to the enterprise, such as a GDPR Regulation Taxonomy.
Knowledge management helps you set up and maintain DataOps.
The role of DataOps

A centralized, curated semantic knowledge layer can also facilitate efforts to set up and maintain a DataOps framework.

According to Gartner, “DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization.” Or to put it another way, DataOps addresses this important question: “how can we get the right business-ready data to the right users as consistently and as efficiently as possible to enable genuine and tangible improvements in business outcomes?”

The benefit of a DataOps framework in any organization is heavily dependent on their overall level of maturity in related areas such as:

- Effectiveness and processes used
- Challenges
- Level of support from internal stakeholders
- Tangible benefits and metrics

Even a minimal DataOps practice is beneficial, but more mature DataOps implementations offer additional benefits, as shown in Figure 2.
The role of DataOps

DataOps maturity model

Increased business value speed in delivering business-ready data

No DataOps
- **Know:** Spreadsheets
- **Trust:** Emails
- **Use:** Hand coding

Foundational DataOps
- **Know:** Departmental/LOB catalog
- **Trust:** Data quality program
- **Use:** Data virtualization, data integration and data replication

Developed DataOps
- **Know:** Enterprise catalog
- **Trust:** Data governance program with data stewardship and business glossary
- **Use:** Self-service data prep and test data management

Advanced DataOps
- **Know:** Enforced and enriched catalog
- **Trust:** Compliance, business ontology and automated classification
- **Use:** DataOps for all data pipelines
The role of DataOps

The tools and methods that the organization uses to gather and use knowledge are a crucial factor influencing their level of DataOps. Specifically, an adaptable and actionable layer of semantic knowledge that crosses lines of business is a key prerequisite to achieving the Developed and Advanced levels of DataOps maturity. The persistent actionable automation and levels of trust required for mature DataOps is only possible through a semantic layer of an enforced and enriched business ontology.¹

There are a number of key areas where such a semantic layer of business knowledge can be exploited to support capabilities needed in a DataOps landscape (see Figure 3).

DataOps capabilities

The tools and methods that the organization uses to gather and use knowledge are a crucial factor influencing their level of DataOps. Specifically, an adaptable and actionable layer of semantic knowledge that crosses lines of business is a key prerequisite to achieving the Developed and Advanced levels of DataOps maturity. The persistent actionable automation and levels of trust required for mature DataOps is only possible through a semantic layer of an enforced and enriched business ontology.¹

There are a number of key areas where such a semantic layer of business knowledge can be exploited to support capabilities needed in a DataOps landscape (see Figure 3).
The role of DataOps

Some of the areas where this semantic layer assists the DataOps landscape include:

- Providing a rich, fully integrated layer of business knowledge in the knowledge catalog
- Forming the basis for any user or departmental views needed to underpin self-service activities
- Providing a more domain-specific level of integration with other relevant tools, such as IBM InfoSphere® Master Data Management, or Cognos® BI components
- Aligning vocabulary content and tooling with open standards such as Egeria, and permitting this knowledge to be exploited across a federation of metadata repositories in the organization
- Establishing an ontology-like structure that supports the different automation functions in data integration, data curation and data governance
- Generating specific DataOps artifacts, such as the necessary API specifications for governed data access, or data mart structures to enable the specific views of data by certain users
- Setting a baseline from which to extend the specific industry knowledge being used by the organization, helping ensure a business-centric language for users
A semantic layer and DataOps gather knowledge in a way that can be used.
The role of the semantic layer in a broader ecosystem

After considering the semantic layer and the role of DataOps, the main remaining question is how to set up a landscape that enables the ongoing and effective gathering of this broad range of knowledge in a coherent and extensible form, and in a way that can be exploited by the organization to support business objectives.

Figure 4 outlines an example of the main sources and consumers of such knowledge.
Figure 4. The role of a central knowledge base within a broader ecosystem of knowledge sources and consumers.
The role of the semantic layer in a broader ecosystem

1. The central knowledge base
   The set of curated business, technical, and operational metadata—the base of knowledge that underpins the whole enterprise knowledge landscape.

2. Central metadata management
   The day-to-day curation and management of the network of business knowledge as well as the broader knowledge base. Includes a range of capabilities such as editing, versioning, searching, and scoping the knowledge base contents. Also includes a growing set of AI/ML capabilities to support such management, such as auto term-to-term mapping across different areas of business knowledge.

3. Technical metadata onboarding
   Initial and ongoing importing of the technical metadata from the data lake and other systems. Represents the bulk of the “technical knowledge” in the central knowledge base.

4. Business metadata onboarding
   Initial and ongoing importing of business metadata from IBM Knowledge Accelerators and other standards. Includes any emerging AI/ML-based capabilities to automate and accelerate the transformation of business metadata from a source—such as PDF, DPM files, or XML—to a target. Represents the “business knowledge” in the metadata central knowledge base.

5. User support
   The use of the central knowledge base to support the day-to-day tasks of various users— including business users, data scientists, data engineers, and data stewards. Usage ranges from simple catalog-supported queries to the semantic interpretation of natural language queries and potentially references to broader linked open data beyond the enterprise. In addition to usage by humans, the central knowledge base should also be accessible to applications via APIs.

6. Generation of downstream artifacts
   Deployment of data lake artifacts from the central knowledge base. This could include the generation of specific sub glossaries, ETL specifications, JSON specifications for APIs, data virtualization specifications, data store/DGLL specs, and more. In most cases the deployment would constitute of two steps: scoping the necessary set of central knowledge base elements needed and transforming that scope into the required platform-specific or schema-specific specifications needed for artifact generation.

7. Support for systems of record and systems of engagement
   Expansion of the central knowledge base to support other areas of the enterprise beyond analytics or systems of insight. The same curated set of business and technical metadata could also underpin a valuable digital transformation of the business operations. To take just two examples: it could reveal the intents and entities used to underpin support chatbots, or it could provide input to the dictionaries and type hierarchies used in the processing of unstructured documents such as loan applications.
Further detail about the central knowledge base

Many organizations that are beginning to define a central business vocabulary want to ensure that this asset can grow beyond the typical areas around data governance and data lakes. Although it can make sense initially to focus such an initiative on these data and analytics areas of the business, the initiative ultimately needs to span all aspects of the enterprise in order to fully exploit the available store of knowledge. For instance, the knowledge base should be extensible to also include the knowledge pertaining to the operational processes and rules underpinning the running of the business.

That is because the central knowledge base is the lynchpin for the entire information environment. It needs to support two different but interlinked sets of use cases:

- The ongoing runtime activities of the various end users of the enterprise, such as business users, citizen analysts or data scientists
- Ongoing generation and evolution of further extensions of the ecosystem for other personas such as data modelers, engineers, and database administrators

The second set of use cases opens up a whole separate dimension to how the central knowledge base can be exploited by the organization. The same business layer that is being used to guide ongoing data governance, data classification and self-service activities could, in the future, also be leveraged as the starting point for the creation of a range of important artifacts needed in the ongoing evolution of the data lake or data management environment.

In other words, while each of the components of the knowledge base are defined separately and can exist in their own right, there are key areas of integration (Figure 5). For instance, the business knowledge area might only be concerned with representing all of the necessary business language and associated rules and constraints, while the data design knowledge area is only concerned with defining the details of any transformations needed to enable the creation of downstream artifacts. However, both of these areas are linked so that any deployment of an artifact such as a data mart or API can be based on the correct business scope as defined in the business knowledge area.
One key area of focus for ongoing development is evolving the ability for organizations to leverage the curated set of business knowledge as the basis for the generation of various more technical artifacts. For example, it would make a lot of sense to ensure that any logical data models being created to drive a data warehouse deployment, for example, should be firmly anchored on the needs of the business. One very effective way to achieve such lineage from the logical data model to the overarching business requirements is to enable the derivation or, even better, the generation of the initial instance of that logical data model from the relevant areas of the central knowledge base.

There are potentially many other such technical artifacts that could, in the future, be generated from the central knowledge base. For example, if an organization is looking to drive the digital transformation of their front-end systems, then using the central knowledge base greatly helps with the standardization of the various API specifications. A more recent and evolving use case is the potential role of ontologies, possibly defined via RDF, to provide the semantic layer for the definition of applications in areas such as natural language processing or data fabric. In such use cases, there are significant advantages to having the generated semantic layer derived from the same curated business vocabulary that is used to underpin the data governance or DataOps activities.

When the central knowledge base supports more use cases, it is more useful to the overall organization. In fact, the whole point of building a centralized cross-enterprise layer of knowledge pertaining to the business, technical and use domains is so that a degree of standardization and reuse of this knowledge can be exploited by the whole organization.
A central knowledge base is the lynchpin for your information environment.
Existing data catalog solutions can already enable meaningful exploitation of the business metadata as expressed in business glossaries and vocabularies. They can even support functions such as automated classification of data sets and enablement of better self-service. However, there is significant potential to expand these solutions and drive more automation, ultimately supporting additional use cases and personas across more enterprise operations.

When considering some of the typical personas and the challenges they face, it is possible to imagine some questions that a central knowledge base can help address (Figure 6).

After implementing a central knowledge base:

- The data steward has a base of knowledge that can grow in a coherent and organic way as the business itself grows and adapts—including addressing key regulatory obligations in areas such as data privacy.
- The data scientist has a more user-friendly and extensive business language to use in self-service analysis.
- Executives and other less technical users (citizen analysts) have the potential for natural language access to this knowledge via a chat bot or other interface, underpinned by ontologies.
- Data engineers and other technical users have the ability to quickly generate the artifacts needed to grow the DataOps runtime environment while keeping it tightly linked to the ongoing needs of the business.
Addressing the full range of current and future enterprise needs

Figure 6. Example questions that a central knowledge base can help answer for selected roles

- How many new customers did we have in Germany last week?
- What are the most suitable data assets to support my predictive risk analysis?
- How many new customers did we have in Germany last week?
- What is the correct business scope to use for my database generation?
The central knowledge base also supports future use cases across the enterprise.
Enterprise knowledge management leverages well-known technologies for DataOps, data management and governance, linked together with a common semantic knowledge layer to accelerate and extend its benefits to more areas of the organization. Whatever personas and use cases you want to support, knowledge management approaches that leverage a central knowledge base can bring significant benefits across the enterprise.

IBM continues to evolve solutions in this area, often in partnership with customers such as ING. To learn more about IBM solutions for both integrated data management and comprehensive enterprise knowledge management, visit IBM Knowledge Accelerators and IBM Watson Knowledge Catalog.

To learn how you can organize your data to be trusted and business-ready for your journey to AI, visit IBM DataOps.
The Journey Continues: From Data Lake to Data-Driven Organization. An IBM Redguide publication.

The term "enterprise knowledge" encompasses all of the potential knowledge that an organization needs to collect and manage—including a combination of business knowledge, technical knowledge and operational knowledge.

A semantic layer is a business representation of corporate data that helps end users access data autonomously using common business terms.

Data operations (DataOps) is the orchestration of people, process and technology to deliver trusted, high-quality data to data citizens fast. The practice is focused on enabling collaboration across an organization to drive agility, speed and new data initiatives at scale. Using the power of automation, DataOps is designed to solve challenges associated with inefficiencies in accessing, preparing, integrating and making data available.

A data catalog is a collection of metadata displayed in a management tool.

An application programming interface (API) is a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service.

An ontology is a set of concepts and categories in a subject area or domain that shows their properties and the relations between them. An ontology can be machine-interpretable.

A resource description framework (RDF) is a model for encoding semantic relationships between items of data so that these relationships can be interpreted computationally.