ABSTRACT / SUMMARY

Data governance is a lifecycle-centric asset management activity. To understand and realize the value of data assets, it is necessary to capture information about them (their metadata) in a connected way. Capturing the meaning and context of diverse enterprise data in connection to all assets in the enterprise ecosystem is foundational to effective data governance. A data governance environment must represent assets and their role in the enterprise using an open, extensible and “smart” approach. Knowledge graphs are the most viable and powerful way to do this.

In this short paper, we outline how knowledge graphs are flexible, evolvable, semantic and intelligent. It is these characteristics that enable them to capture the description of data as an interconnected set of information that meaningfully bridges enterprise metadata silos. But not every graph technology has these qualities.

TopBraid Enterprise Data Governance is the first data governance solution that uses true knowledge graph technology based on W3C standard RDF graphs that are connectable — exactly like the web. We conclude by illustrating how, using knowledge graphs, TopBraid EDG delivers integrated data governance by addressing all three aspects of data governance — Executive Governance, Representative Governance, and Applied Governance.
The term data asset is used in data governance to describe any data element or data structure that has value to an organization and enables it to perform its functions. A data asset may be a database or a dataset. It may also be a structure (e.g., a table or a view) within a larger data asset, down to a specific column or field.

Some information about data assets can be collected directly from the data sources. Known as technical metadata, it can tell us the number of tables in a database, the number of columns in a table or the datatype of a column, whether its values are unique, the min and max range of values, and so on. While this information is important, it does not tell us the meaning of the data, how a data asset is being used and by whom, what security and regulatory requirements it must satisfy and other key information that identifies in what way the data asset delivers value to an organization. In other words, what actually makes it an asset.

To provide useful context to data and enable its management as an enterprise asset, we need to capture and relate it to information about other assets in the enterprise. So, what do we mean exactly when we use the term “asset” more broadly?

Virtual assets such as data or intellectual capital tend to be less understood than physical assets (buildings, equipment, etc.). Oxford Advanced Learner’s Dictionary describes an asset as a valuable or useful quality, skill or person; or something of value that could be used or sold to pay off debts. This definition helps in establishing that an asset could be described as any entity that has value, creates and maintains that value through its use, and has the ability to add value through its future use.

Enterprise assets include activities, processes, functions, business applications, technical infrastructure and policies. And they also include the user communities that perform activities during the course of which they create and use data.

The value provided by an asset can be both tangible and intangible in nature.

As an economic entity, an asset’s value may be different in different stages of its lifecycle such as design, development, active use and retirement. Management of assets requires: 1. understanding the economic benefits of the assets, and 2. implementing strategies for preserving and realizing their value throughout their respective lifecycle stages.

Just as with other enterprise assets, to understand and realize the value of data assets, we need to connect them to other enterprise assets by capturing information about them (their metadata) in an integrated, connected way.

Today, enterprise information and data are among the most important assets of an enterprise, and their optimal management and use are the motivations behind the rapid growth and urgency of interest in data governance. So, the cataloging of diverse enterprise data by capturing its full (technical and business) context and meaning through connections across all assets in the enterprise ecosystem is foundational to effective data governance.

This means that a data governance environment must represent assets and their role in the enterprise using an open, extensible and “smart” approach. One powerful way to do this is through knowledge graphs.
KNOWLEDGE GRAPHS FOR INTEGRATED DATA GOVERNANCE

You may have heard the term “knowledge graph” from Google or Microsoft; they use knowledge graphs for smart search. In this paper, we describe how knowledge graphs provide a powerful platform for integrated data governance. Search is also an important feature since stakeholders need smart search in order to find assets of interest.

A knowledge graph is an interconnected set of information that is able to meaningfully bridge enterprise metadata silos. Here are some characteristics of knowledge graphs. They are:

- **Flexible** — graphs are the most flexible formal data structures (making it simple to map other data formats to graphs) that capture explicit relationships between items so that you can easily connect new data items as they are added and traverse the links to understand the connections.

- **Evolvable** — able to accommodate diverse data and metadata that adjusts and grows over time, much like living things do.

- **Semantic** — the meaning of the data is stored alongside the data in the graph, in the form of the ontologies or semantic models. This makes knowledge graphs self-descriptive, a single place to find the data and understand what it’s about.

- **Intelligent** — the semantics of data are explicit and include formalisms for supporting inferencing and data validation. As a self-descriptive data model, knowledge graphs enable data validation and can offer recommendations for how data may need to be adjusted to meet data model requirements. They also enable drawing conclusions and new information from the available data.

A knowledge graph allows you to store information in a graph model and use graph queries to easily navigate highly connected datasets. For example, if a user is interested in a specific data source that participates in some specific business process or stores certain type of data (e.g., customer data), using graphs, you can help them discover other data sources that are used by that process, or other data sources that store customer data. You can build and query complex models of regulatory rules and organizational policies, or model any general information, like Wikidata. These qualities make knowledge graphs an ideal and the only viable foundation for bridging and connecting information about all enterprise assets.

Graph technology is gaining adoption and you may hear more and more about it. It is important to understand that not all graphs are created equal. Flexibility of the data structure is common to all graph technologies. However, other knowledge graph qualities described here are not necessarily present in all graph technologies. Thus, not every graph is a knowledge graph. An essential quality of a knowledge graph is its ability to capture the description of data. Two graph data models that are the most popular today (property graphs and RDF graphs) differ strongly in their support for this capability:

- **Property graphs** — an example of a database implementing a property graph data model is Neo4J, but there are a number of others including open source property graph databases. No standard definition of the “property graph” data model or a standard serialization of data they store exists. Different vendors implement their own variants of the property graph data model which limits their interoperability and the ability to connect and query them in a uniform way. Property graphs are flexible and evolvable like all NoSQL databases, but they are not semantic. They lack a language (a representation) for storing the meaning of data. Consequently, they can’t, on their own, be regarded as “intelligent.”

- **RDF (resource description framework) graphs** — a number of databases are based on the RDF data model. Commercial vendors offer RDF database products, with no single marketplace leader at this time. There are also commercial software-as-a-service offerings, such as Amazon Neptune. Open source software plays an important foundational role with projects such as Apache Jena providing publicly tested implementations of W3C...
standards. The generic name for the database technology that implements the RDF data model is **triple store**. A name that reflects the fact that each graph statement consists of three parts: subject, predicate and object, just like a simple sentence.

RDF is a W3C standard (just like XML and HTML). W3C stands for the World Wide Web consortium, a web standards body led by the creator of the internet, Tim Berners-Lee. Thus, the RDF data model is fully interoperable across all vendors. It offers standard data serialization formats (for bringing data in and out) and a standard query language.

W3C also standardized languages for describing data semantics and inferring new data facts.

- The most commonly used is a very simple language called RDFS — RDF Schema language. It offers very minimalistic support for data description.
- A new W3C standard that provides a more complete RDF schema language (and more) is called SHACL (Shapes and Constraints Language).

Further, every resource in an RDF graph has a globally unique, dereferenceable, web identifier — a URI. With its URI, a resource can be reliably referred to and accessed from any application.

Just as all graphs are not created equal, not all GUIDs (globally unique identifiers) are created equal. Web technologies and architecture provide us with an authoritative system for globally unique dereferenceable identifiers — URIs. Dereferenceable URIs are based on the well-established theory and practices of “data access by reference”. Use of URIs to identify resources in a knowledge graph ensures reference-ability and connectivity across an enterprise ecosystem through standardized protocols such as HTTP.

The Web and, more recently, Blockchain technologies have made former, proprietary or system specific approaches to global identity undesirable because they can result in silos that your enterprise will then have to spend significant efforts on connecting — often in brittle, expensive and inefficient ways.

**Using URIs as identifiers, on the other hand, ensures that enterprise assets are truly unique and addressable.** You can have two different systems use the same identifier (URI) when they talk about the same assets. Thus, graphs of data from such systems come together seamlessly without the need to employ data integration efforts. You can also refer to a URI and dereferencing will deliver data about the asset it represents.

From the description above, you can see that RDF graphs are indeed knowledge graphs. Another key feature of RDF graphs is that they are connectable — exactly like the web. You can think about one giant knowledge graph spanning the world. And you can also think about multiple, separate (but connectable) knowledge graphs — like an individual website or a web page.

Property graphs, on the other hand, can’t be used as knowledge graphs unless a higher level knowledge representation is defined for them. In principle, it is possible for an organization to build its own knowledge representation on top of property graphs. However, it would be proprietary, not interoperable and any organization doing this would be duplicating many years of effort expended by the members of the World Wide Web consortium to define and ratify its semantic standards.
DATA GOVERNANCE VERSUS METADATA MANAGEMENT

As the topic of data governance becomes increasingly important, many organizations try to implement it by focusing solely or primarily on technical, system-oriented concerns. In some ways, this is not really surprising. For years, the ETL and metadata management vendors have been saying that their products provide “data governance.” Traditionally, they are used in a context of a specific project — such as building a data warehouse.

More recently, dedicated data scanning and profiling tools have become available. They do not perform data processing, but they offer a way to scan various repositories to re-capture technical metadata. While all these tools provide some (often, sophisticated) ability to capture technical metadata and technical data lineage, they do not provide true data governance or lineage as business understands it. As we outlined previously, data governance is an asset management discipline.

To treat data as an asset, its value needs to be understood from the business context of the data’s use and importance. Its lifecycle needs to be understood and governed to ensure data delivers value in a sustainable way. Thus, project centric approaches and/or approaches that focus on technical metadata can’t deliver data governance.

When IT needs to integrate systems for a specific project, they identify data that needs to be brought together based on project specific requirements and analysis of the sources and targets. These findings are usually in various Excel spreadsheets, Word documents, PowerPoint presentations, Visio flowcharts and similar artifacts. The documents will include details on what data will be moved or aggregated. They may include information on how this is to be accomplished as well as some partial business context including who will be using it and why, the frequency of data movement, mapping rules, quality thresholds, and more. After design is completed, IT will implement code to support these objectives. The implementation process represents and captures some of the knowledge assembled during the design, but it is a technical translation of the knowledge and a substantial part of the business context remains only in documents, and in designers’ and analysts’ heads.

Once the solution is implemented, people who worked on the original project will sooner or later move on. The documentation of the design is rarely maintained and becomes outdated. It may also be misplaced and/or hard to understand by new stakeholders that were not part of the original team. New players faced with the need to understand exactly what was implemented and how changes may impact the system — and more importantly, the business — are likely to look at the data management tools that have been processing the data. Often, there is the hope and belief that using tools that reverse engineer the solution will help them solve the problem of recovering the lost knowledge.

They may use scanning and profiling tools for this purpose and they will get some information. The problem, however, is that much of the relevant knowledge is simply not available in the code so the information is not complete. Moreover, outputs of such reverse engineering are typically too low level. Using reverse engineering alone will not yield information that is sufficiently meaningful to the business.

As a partial, ad hoc solution to this problem, various custom fields may get added to different data management tools in order to supplement technical metadata with business context information (additional metadata). Project-specific knowledge re-capture efforts may be started to populate these fields — or to create new Excel spreadsheets. In either case, the information captured is specific to a given solution, project or a set of tools — resulting in yet another metadata silo.

A medium enterprise today has hundreds of systems. A large enterprise has thousands. The number of dataflows across them is typically an order of magnitude larger than the number of systems. Capturing information about them is a large job, especially when done after the fact.

Automation is important — but a retrospective approach will not deliver capabilities needed by an enterprise in order to manage its critical data as an asset. In the next section we will describe an integrated pro-active approach to data governance.
THREE ESSENTIAL ASPECTS OF INTEGRATED DATA GOVERNANCE

Data governance requires pro-actively implementing an asset management discipline — addressing all three aspects of governance shown in the figure below.

Executive Governance is concerned with putting in place control processes and policies, or, formalizing and instrumenting them if they already exist informally.

Representative Governance is focused on having models of the information you will be capturing such as glossaries, data sources, applications, reference data, and so on, and using them to describe these assets. In knowledge graphs, a key point to remember is that models of data are uniquely identified, stored, and can be queried and used for reasoning in exactly the same way as the data they describe.

Applied Governance is about using the information you have captured to address specific needs. For some, it could be the ability to assess the impact of a change in data sources. For others, it could be about tracing data lineage in order to satisfy some regulatory compliance requirements. Yet another common goal is improving the quality and consistency of data. Often, these initiatives and/or needs are related. Yet, they may be implemented only in a project specific way.

In the figure, the circles within the triangle represent examples of where organization’s focus may be currently in their data governance practices:

Process Centric. Organizations may start with a top-down “Executive Governance” approach and work in this area for considerable time. This focus is represented by the large orange circle, typified as a “Process Centric” approach. The danger of the primary focus on executive governance is that without a direct connection and use of it in representative and applied governance, it doesn’t deliver value to an enterprise. Thus, such top-down only, or primarily, programs may lose traction, get de-emphasized or even cancelled after some time.

Metadata Centric. Organizations that focus strongly on “Representative Governance” can be characterized as “information gatherers.” This is a good stepping stone to applied governance and it also has value of its own in creating a searchable catalog of information. This is a knowledge capture activity. It may include automated aspects of data scanning and profiling to gather technical metadata.
In the diagram, the green circle represents efforts that have their prevalent focus in this area typified as “Metadata Centric.” Representative governance must include interaction and participation from business and IT alike — to gather all other types of metadata. Connection to executive governance is important in providing a framework and support for such interactions. Connection to applied governance is important because it represents a specific use or application that the represented information must support.

Without these connections, representative governance can’t be sustainable and will not deliver the organizational transformation that is required for data governance. The trustworthiness of the collected information will likely be questionable, hence its value will likely not be as certain as needed. In other words, there may be a lot of information across the board, but not enough to support any specific use cases.

**Project Value Centric.** A primary focus on “Applied Governance” is common and invariably it can be narrow in scope. As with the earlier example cited, it may be part of an initiative to support some data integration effort. In the diagram, it is depicted as dark blue circle typified as “Project Value Centric.” But, when it is not connected to executive and representative governance, it becomes a one-off effort that is likely to create yet another metadata silo.

**TOPBRAID EDG DELIVERS INTEGRATED DATA GOVERNANCE**

TopBraid EDG is the first data governance solution that is based on true knowledge graph technologies. Further, it supports and enables all three aspects of data governance. It delivers an integrated approach to data governance with customizable workflows, flexible governance operating framework options, enhanced visualizations, and advanced capabilities to create, manage, and view connections between technical and business metadata.

**TopBraid EDG for Executive Data Governance**

Users of TopBraid EDG can define governance areas — business functions or data domain areas and create and assign governance roles for the areas (see the illustrative screen shot below). Different types of asset collections (e.g., business glossaries, reference data, data dictionaries and more) are associated with the respective areas.

Users can also capture data governance metrics and policies, identify and resolve issues, build workflow templates for data governance processes and work with other governance assets.
TopBraid EDG for Representative Data Governance

TopBraid EDG includes over 100 predefined asset types such as Glossary Term, Requirement, ETL Script and many others that represent technical and business context. Some of these are shown in the diagram below.

Customers can configure each one and they can also add new asset types. A new asset type is defined semantically, in the same way as pre-built types. Semantic models in TopBraid EDG are highly expressive to support the richness of information and use cases required by data governance.

To populate the asset models with the data they describe, TopBraid EDG lets users automate collection of information from the data sources. It infers new information based on the available facts to connect and enrich collected data.

TopBraid EDG for Applied Data Governance

TopBraid EDG supports different use cases at the level that is most appropriate for different stakeholders. For example:

- **Data lineage** — where one looks “upstream” for the sources of information.

  Questions about lineage may be motivated by the regulatory compliance requirements. They may also be driven by the need to understand how to bring together data from different sources in an aligned way.

  Different stakeholders may use the same term “lineage”, but they require different levels of information. Business users want to know at the entity, object or an “information asset” level where information is sourced from. Technical stakeholders will typically need to drill down from a logical level to specific data structures.

- **Impact** — this is another use case, where one looks “downstream”, to where information is used after it first appeared in some system.

  Users are interested in impact analysis in order to understand how some change may affect an enterprise.

  - A change will often start at the business level — for example, an enterprise may change the definition of a business concept such as an employee. As the change is worked through, it will impact a variety of processes, systems and data assets.

  - A change may also be initiated at a technical level. For example, the software used to hold some data may need to be upgraded. This will have impact on other technical as well as business assets.

In the screenshot on the next page, we see a high level view of how data flows between business applications to produce a report selected by a user interested in understanding data lineage.
Each of the connections between applications can have different levels of detail:

- **Logical Level** — describes the data flow at the business level. An example of the data flow information at the logical level is shown in the diagram below. We can see what business entities and information are being passed between two applications without going into technical details of how this information is stored.

- **Data Level** — expands on the logical level to describe the data-level details behind the logical flow: what specific fields in each data source are being used and how they are transformed as data moves between systems and sources.

- **Pipeline Processing Level** — additionally describes any intermediate data processing steps that may happen in ETL and other scripts.

As this example shows, irrespective of the focus (business vs. IT), the level of granularity and the reason for using asset information, TopBraid EDG can deliver the insights and key effectiveness and efficiency improvements organizations expect from data governance.
NEXr STEPS?

Your concerns and priorities may be informed by common questions such as:

1. How would you characterize your current data governance initiatives?
   Are they establishing an enterprise-wide foundation? Or, primarily focused on individual project needs as they arise?
   Are you providing support and appropriate enablement to all stakeholders?
   Are you addressing the need for governance processes, consistent and rich metadata representation (technical and business) and delivering immediate value to stakeholders? Or are you focusing mainly on one or two of these aspects?
   If you are not addressing all three aspects of data governance, your next step should be to consider how you may add to or reshape your data governance initiative to address the full scope.

2. If you are addressing all three aspects of data governance, are you doing this in a connected way? Or are you addressing each using its own silo approach?
   The goal and value of data governance is predicated on its ability to bridge data and metadata silos and enable integrated asset management. A “toolkit” approach to this mission can’t work unless all tools in the kit use the same open and standards-based foundation. Otherwise, each effort becomes a “band aid” that may provide a solution to some immediate need or to resolve a pain point in an expedient way. But longer term, such isolated efforts create more silos of information. Your next steps should be to determine how to implement a more holistic and interoperable approach to data governance.

3. Does the technology you are using support integrated data governance?
   As a business initiative, data governance requires the right combination of people, processes, skills and technology. It is not about the technology alone. However, given the scope of data governance and the complexity of enterprises’ data ecosystems, it is not possible to implement data governance without capable technology. As described in this white paper, knowledge graphs are unique in offering a powerful, fully integrated way to govern all your data assets. If you are not familiar with this new technology, consider learning more about knowledge graphs to better understand how they can help.

About TopQuadrant

TopQuadrant helps organizations succeed in data governance. Its flagship product, TopBraid EDG, delivers easy and meaningful access for all data stakeholders to enterprise metadata, business terms, reference data, data and application catalogs, data lineage, requirements, policies, and processes.

TopQuadrant’s customer list includes organizations in financial services, pharma, healthcare, digital media, government and other sectors.

Governance Packages Available in TopBraid EDG

In ramping up a data governance program, different organizations may have differing starting points. With TopBraid EDG, you can start incrementally and add capabilities as you go. For details on available EDG packages and additional modules, visit [topquadrant.com/products/topbraid-edg-gov-packs/](http://topquadrant.com/products/topbraid-edg-gov-packs/)

For more detail or to schedule a demo, contact us at: edg-info@topquadrant.com