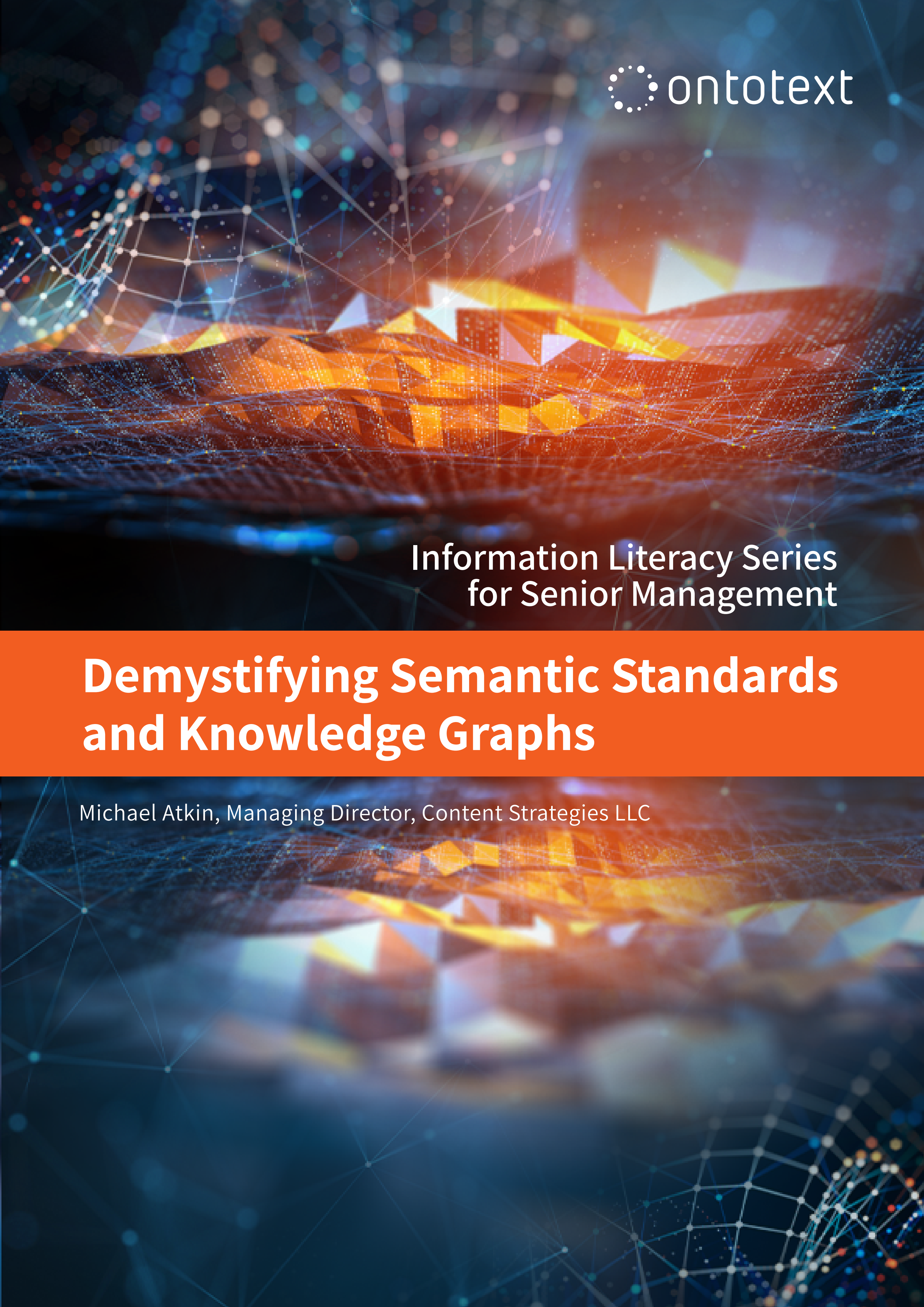


ontotext

Information Literacy Series for Senior Management

Demystifying Semantic Standards and Knowledge Graphs

Michael Atkin, Managing Director, Content Strategies LLC

Knowledge graph is a marketing term popularized by Google in 2012 that describes a set of standards for expressing the identity, location and granular meaning of data. Remember, data is simply a representation of real things. It represents our customers, products, people and processes. It represents the commitments firms make and the obligations they accept. It is an essential factor of input into absolutely every aspect of operations.

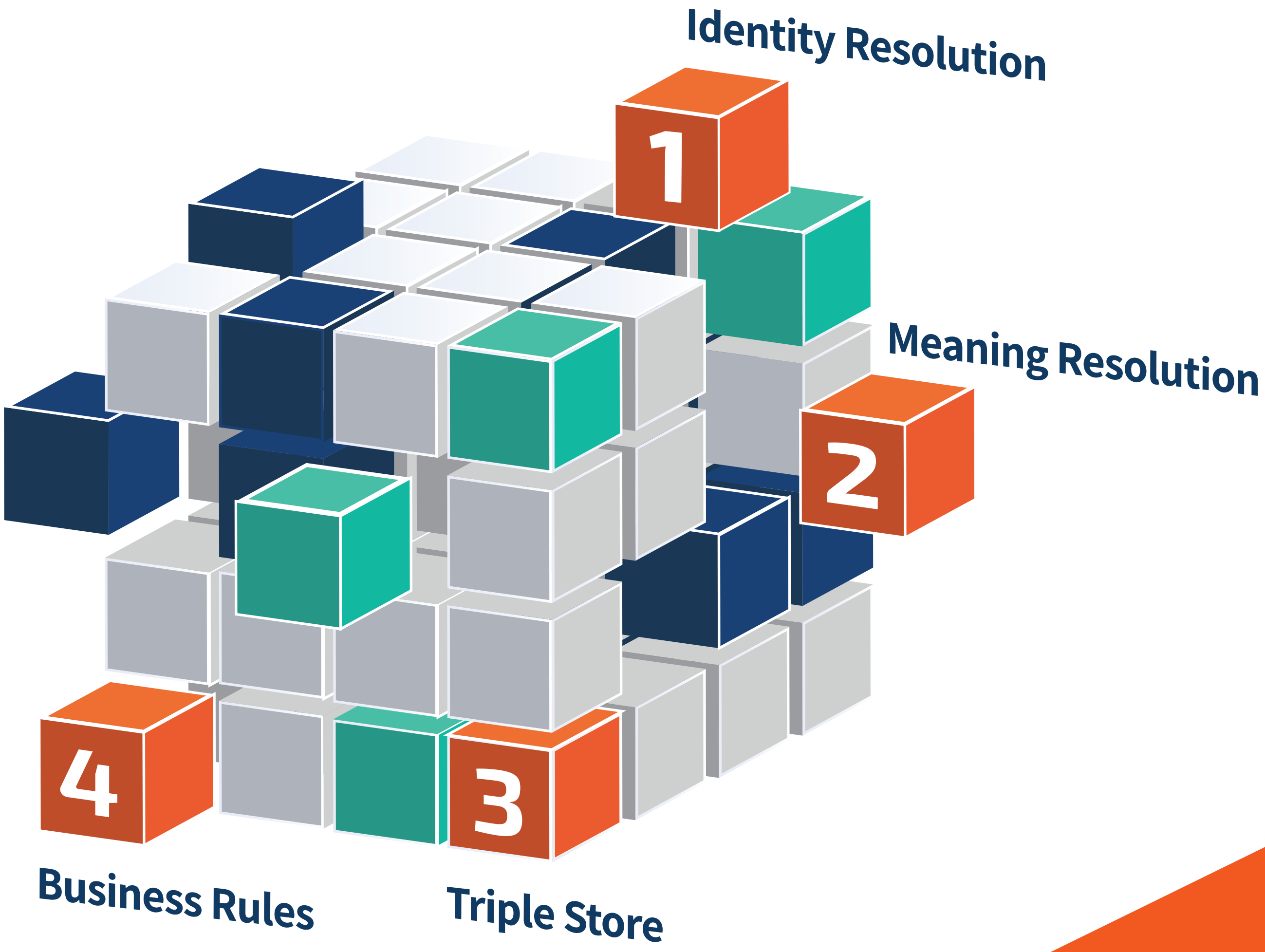
Despite its essential nature many organizations and companies have a data problem. The problem is based on two realities that stem from technology fragmentation. The first reality is that we have allowed the meaning of data to become mismatched across systems, databases and operational boundaries. We have done so because we have transformed and independently renamed data to match the software that drives our applications. We've created this 'data incongruence' because we seek to manage context between front-office activities (transaction-related or diagnostic) and back-office activities to address legal, contractual, procedural and analytical requirements.

Not only do we suffer from data incongruence, we suffer from the limitations of proprietary technology that was state-of-the-art two generations ago. This is the legacy of relational databases where data is organized into columns and stored into tables linked together using internal keys. We know that organizations are supporting many thousands of tables many with conflicting column names and all with relationships that must be explicitly structured as well as definitions managed separately from the content. As a result, we spend significant effort moving data from one place to another – and countless person hours reconciling data and its meaning.

The net of all this is that we've allowed data to become isolated, incongruent and inflexible because of technology fragmentation and rigid technology environments. These problems are now recognized as serious liabilities. This diverts resources from business goals. It extends time-to-value and inhibits analytical flexibility. It leads to business frustration. And it fosters mistrust across organizational boundaries.

To fix these problems, we have to fix the data. And we can state unequivocally that this is a solvable problem. One that does not require a big investment in new technology or the 'rip and replace' of existing infrastructure. The pathway is simple and straightforward – adopt the principles of data hygiene and take advantage of semantic standards for identity, meaning and business rules. This is the prime goal of data management – to ensure that the meaning of data is consistent, precise and nuanced as It flows across processes and between entities. Once you do (using the language of the Web) you will be able to turn data from a "problem to be managed" into data as a "resource to exploit."

Our goal with this paper is to demystify these concepts for executive stakeholders and demonstrate that this new form of 'information literacy' is a capability that is both easy to understand and worthy of being elevated as a 'top-of-the-house' priority.



ESSENTIAL BUILDING BLOCKS

The Four Concepts about Semantic Standards

ESSENTIAL BUILDING BLOCKS

The approach to finding, interpreting and linking data is now available to be used by companies and organizations to harmonize data, unravel risk and capitalize on business opportunity. The application of these standards (termed a knowledge graph) solves the data harmonization problem. It gets us out of the business of data wrangling and into the business of using data for innovation. And it does so in a way that is cost-efficient, non-intrusive, based on open-source standards and governed by trusted processes. Below are the four concepts that executive stakeholders need to know about semantic standards:

1 **Identity Resolution** (IRI): Knowledge management starts with identity. In the knowledge graph, all objects are identified with at least one universally unique, permanent and web-resolvable identifier in the form of an Internationalized Resource Identifier (IRI). The IRI is a meaningless 'identifier' (what something represents) as well as a 'locator' (where it resides). Instead of downloading copies of a database, managing cross-referencing tables, updating APIs and managing a whole suite of testing, you just point to the IRI. This eliminates the task of moving and mapping data. Think of the IRI as the Rosetta stone for data harmonization because all the content in your organization is linked to its own unique (never changing) identifier.

2 **Meaning Resolution** (ontology): We know that one of the drivers of the 'data problem' is that data has been modified, transformed and renamed many times over its lifecycle. We also know that ensuring a unified view of data is challenging because it can have different data structures, definitions and contextual meanings. All of this makes integration difficult and expensive – particularly when there are dozens of systems of record all serving various operational processes and independent lines of business.

This process of reconciling glossaries that reflect the local "business language" of bespoke applications is complex and best accomplished using modeling processes and content standards that describe what the data means as well as how concepts are connected. This is what an ontology is for. Ontology is simply a data modeling and communication process that is used to ensure a shared understanding of requirements between business stakeholders and applications developers. It starts with the ability to capture concepts and relationships as defined by subject matter experts.

The Semantic Web standard uses conceptual data models to precisely describe what the data means as well as how concepts are connected. The meaning of every data point is directly resolvable to a machine-readable definition. Ontologies are linked to business glossaries that can be directly translated into physical data structures. The properties in each data point are linked to their definition so the meaning is never in doubt. Expressing data at a granular level allows ultimate flexibility for it to be sliced, diced, combined and aggregated.

3 **Business Rules** (SBVR, SHACL): Business rules are needed to make sure the data is fit-for-purpose. These 'conditional expressions' are established by criteria specified by subject matter experts and translated into validation rules, calculation rules, classification rules, transformation rules, workflow rules, business definition rules – many types of rules from simple to complex. These rules can be expressed in standard language and stored in the knowledge graph. They are linked to data and process quality as well as to the ontologies to ensure that meaning is shared (not obscured by vague terms or cryptic codes). The logic is captured and expressed as executable models and consistently enforced across all systems and processes.

4 **Triple Store** (RDF, OWL): The big contribution from DARPA was to shift from data that is 'location-based' as a coupled pair stored in tables - to data that is 'meaning-based' in the triple store language of the Web. To grasp the value of triples, understand that data is organized into groups of three that contain subjects and objects that are linked together by predicates and verbs. It is just a sentence structure. These concepts are all precisely defined based on the knowledge of subject matter experts in the form of an ontology. And once you define these concepts at their most atomic level, you can link them together. These ontologies link the meaning of data to business glossaries that can be directly translated into physical data structures that drive our applications. So, instead of location-based, the data is meaning-based.

FOUNDATIONAL CAPABILITIES

01 Quality by Math

02 Concept Reuse

03 Context

04 Access Control

05 Lineage Traceability

06 Governance Simplification

07 Machine-Readable

08 Continuous Testing

FOUNDATIONAL CAPABILITIES

Using the four building blocks described above, the knowledge graph provides eight foundational capabilities that work together to create business value.

1 Quality by Math: In a knowledge graph, data is aligned to precise meaning and embedded into the structure of the content itself so that users always know what the data represents even as it moves across organizational boundaries. This means that errors and definitional conflicts are verified at source before they are introduced into operational systems. Quality is rules-based and unhooked from both schemas and data models that are often tailored to specific applications. The rules are linked to structured vocabularies and resolved to the unique IRI to ensure that meaning is both discoverable and able to be shared. The goal is automated quality assurance. This is done at a granular level so that users have confidence they are getting the information they need to understand context and examine ad hoc business questions. And from a compliance perspective, data in the graph is immutable because lineage can be traced, and nothing can be deleted except by policy.

2 Concept Reuse: One of the challenges associated with conventional database design is the problem of 'hard-coded assumptions' (i.e., doing the same thing in a slightly different way based on some design objective). Engineers and architects often make explicit assumptions about their domain and code them directly into their applications. Hard-coding these design choices in programming language makes them hard to find and hard to change – particularly when documentation or programming expertise is lacking.

Using Web standards and ontologies for modeling eliminates this problem of hard coding because it focuses on concepts, not specific applications. Users always understand what the data represents at its most granular form. This enables an efficient reuse of important concepts across systems and processes. Consider the example of time. Different domains require different ways to model time – including the notions of time intervals, points in time and relative measures of time. With a detailed ontology, all concepts of time are captured, so that the appropriate dimension can be selected as needed (not reinvented) for the specific application.

3 Context: Semantic standards allow architects to separate business logic from code. And the business logic can be expressed just by looking at what the data element represents. This is accomplished by reference to the ontology and by its singular identity. This realization of precision can be moderated by a time stamp to express exactly when it occurred and by source, so you know where the data came from. Time is important for analysis and source is important when you are seeking to determine if the data can be trusted. With semantic standards, we can understand all data in context by examining these four dimensions of identity, meaning, time and source.

4 Access Control: Technology that grants and enforces access rights to data must be managed at the data, platform, applications and role level. The rules for entitlement and access control must be linked to lineage and transformation processes, tracked and audited. This is mandatory for managing security and ensuring privacy and must be kept synchronized as individuals move across departments and perform a variety of roles.

The problem is that many systems come together at the enterprise warehouse, each with their own entitlement expression. Linking access control to this proprietary technology locks organizations into specific approaches. This becomes a huge, complex and messy administrative burden when trying to replicate entitlements across technology environments. The knowledge graph is able to solve this dilemma by modeling business rules (in context) for all circumstances. The entitlement capability in the graph automatically executes these models by assigning access control at the data and applications level. Security is embedded in the design of the data and not constrained by either systems or administrative complexity.

5 Lineage Traceability: In the knowledge graph all data is linked to a single identifier. That means firms can trace the data as it flows through systems. Data professionals and business users know what the data represents as well as how it is used in the data production process. Data can be transformed and renamed many times as it flows across systems without losing the knowledge of where it came from, what it represents and where it is going. Lineage and provenance objectives are automatic and fully auditable – as well as tested on a continuous basis. The knowledge graph becomes the logical point of distribution because it traces data flow and is fully auditable by source, purpose and responsible party.

6

Governance Simplification: The knowledge graph uses the capabilities of resolvable identity, precise meaning, structural validation and lineage traceability to shift the governance focus from people-intensive data reconciliation to more automated data applications. With semantic standards, firms can create a connected inventory of data (i.e., what exists, how it is classified, where it resides, who's responsible, how it is used and how it moves across systems). Data is traceable to all applications enabling users to run flexible queries and perform contextual search. Data quality is structurally enforced so consistency is ensured across repositories. Issues are identified by the ontology and able to be resolved when and where they arise. The knowledge graph changes the governance operating model by simplifying operations, automating issue management and facilitating a collaborative environment for integration testing.

7

Machine-Readable: Semantic standards are written in a language that both humans and machines can understand. The meaning of data is standardized at a granular level. Data is linked to machine-executable rules with audit trails. Policies can be modeled as machine-executable rules. Semantic standards are rules-based and not connected to data models that are tailored to specific applications. The use of machine-readable standards facilitates automatic validation and provides assurance of data quality.

8 Continuous Testing: In the knowledge graph, requirements, use cases and individual user objectives are linked to automated testing procedures and issue management. All data pipelines have a full and structured test coverage for every change. Without automation, the cost of introducing new components and new functionality is high. With semantic standards, every change in the ontology is linked to a testing process for both logic and circular reasoning. There is a defined and automated governance process for change management. If there are changes to authoritative sources, the downstream implications and dependencies are tracked and tested.

These four open standards result in eight foundational capabilities that yield what could be described as the "Data Bill of Rights." You have the right to expect the data to be true to original intent. You have a right for it to be defined at a granular level, self-describing and reusable. You have a right to have the data available and accessible when needed as part of your asset inventory. You have a right for the data to be in a format that is flexible to use and not stuck in rigid schemas. You have a right for the data to be traceable as it flows across processes and testable as for-for-purpose. With semantic standards, all of these rights are achievable without a huge investment in technology or massive disruptions to the way your organization operates.

VALUE DRIVERS AND USE CASES

Put it all together. Four critical standards for identity, meaning, business rules and expression to deliver the data Bill of Rights. And it is just a short jump to make the leap from understanding these building blocks and capabilities to articulating the overall value proposition. The best way to think about it is by referencing the three "C's" of cost, capability and control. These are standard KPIs that resonate with executive stakeholders (who think about growth and velocity) with technology executives (who think about resilience and scalability), with business executives (who think about use cases and time to market) and with compliance executives (who think about transparency and traceability).

From the cost side of the equation, we start with factual certainty. This is the prerequisite for data integration - that we simplify by standardizing meaning, resolving identity and tracing data flow. With factual certainty we know precisely what the data represents, with context. This enables us to construct our connected inventories of assets to better allocate resources. It enables us to automate processes by reducing reconciliation and mitigating process failure. It enables us to consolidate and scale systems. And it supports efforts to simplify data governance by locking down meaning. Conservative calculations suggest this cost savings can amount to at least 30% of total operations.

From the capability orientation, this is about understanding relationships for better customer profiling and predictive marketing. This is about flexible inquiry by giving business analysts the tools they need to follow their intuition. Adopting semantic standards allows users to perform scenario-based (what if) analysis by asking questions of the data rather than restructuring it and reconciling its meaning. Flexibility and the capability to both construct and navigate relationships is the best tool we have for competitive analysis, for managing the supply chain, for targeted selling and for determining both customer and product ROI.

And from the control perspective, adopting semantic standards supports our ability to consistently aggregate data across lines of business. This is the key to managing systemic risk and ensuring compliance with our legal obligations. It is about being able to look at interrelationships from multiple viewpoints whether it be for regulatory compliance, traceability, privacy protection, access control or the management of intellectual property rights. And (of course) it supports the goal of security. We can control access at a data level, not just a systems or process level. We can trace the flow of data. We can unravel our business calculations. We can prevent fraud and secure sensitive data from falling into the wrong hands.

No matter how you examine it, the value proposition is overwhelming. No matter what your initial use case drivers are, you get all these capabilities. Semantic standards are the mechanism for addressing the data problems caused by technology fragmentation. And not only does it solve the data challenge, it adds operational capabilities that were not previously possible. And it does so without a huge investment in new technology and in a way that fully integrates with your existing environment.

CONCLUSION

I described the objective of this paper as enhancing the ‘information literacy’ of executive stakeholders. I hesitated to use that phrase because suggesting that someone is illiterate might be perceived as an insult. But that is not the case, and nothing could be further from the truth. Information literacy is a new capability.

Most of our organizations have come of age in a world dominated by technology. We have seen multiple technology revolutions with new capabilities coming at us faster and faster. We have been racing to catch up and, in the process, created big organizational departments to make it all work. In the midst of all this activity we didn’t realize that the data paradigm has not really changed all that much. We are still managing data as a coupled pair stored in tables with mismatched column names where relationships are explicitly defined, and meaning is managed separately from structure.

The problem is that we have thousands (sometimes tens of thousands) of locations that exist at the intersection of a column and row in a relational environment. We have modified the meaning of data to make the proprietary software that drives the applications work in context. We are victims of our own innovation. We have neglected to be stewards of what the data really represents – particularly as it gets aggregated across lines of business and calculated by complex and nuanced rules.

Information literacy is about understanding this fundamental truth. It is about understanding that the goal of unambiguous shared meaning is an instrument to transform the business. It is about understanding that the causes, implications and liabilities of data that is structured in rigid processing environments which are a terrible legacy. It is about getting our analysts out of the business of being data janitors. And it is about recognizing that we are not going to fix the problems of technology fragmentation by using the same conventional approaches that created the problem in the first place.

There is a business rationale that we must all work to adopt. The inability to automate processes, explore ‘what if’ questions, aggregate data with confidence, secure sensitive data, respond to client needs and turn analytical ideas into action will add up to competitive disadvantage in our complex and interdependent world. The pathway out of the morass is straightforward – implement the principles of data hygiene and adopt semantic standards for identity, meaning and business rules. This is a solvable problem. Think of it as building the data infrastructure for the digital world.

Michael Atkin has been an analyst and advocate for data management since 1985. His experience spans from the foundations of the information industry to the adoption of semantic technology. He has served as an advisor to financial institutions, global regulators, publishers, consulting firms and technology companies.

The views and opinions expressed in this white paper belong solely to the author, and not necessarily identical to those of Ontotext.

ontotext
info@ontotext.com