White Paper
The Perfect Semantic Platform to Harmonize Your Data and Provide New Insights
The Smartlogic + MarkLogic solution
When trying to answer critical business questions or build new business applications, organizations are faced with the challenge of making sense of all of their data. They have information distributed across disparate data silos that tend to be rigid and tailored for specific uses. And, even within data silos, information is often disjointed and devoid of meaning.

To make sense of all the information available, organizations are leveraging the power of Smartlogic and MarkLogic as a combined semantic platform to handle heterogeneous data, extract meaning from data, understand data in context, provide more relevant search, and discover new relationships. From investment banks onboarding clients to large healthcare providers tracking risk factors among patients, organizations today are using Smartlogic and MarkLogic as a proven semantic platform to gain more value from their data.

Introduction to Semantics

A semantic platform provides a universal framework to describe and link data so that it can be better understood holistically, allowing organizations to see and discover relationships in the data. A semantic platform combines two elements: a content intelligence platform (such as Smartlogic Semaphore) to create and work with the semantic data, and a database (such as MarkLogic) to natively store data as RDF triples.

Content Intelligence

Unstructured data and text content are difficult to manage for computers, as are concepts and relationships. Content intelligence refers to an array of semantic technologies that are able to perform activities such as text mining and ontology management—which address these challenges and make more sense of unstructured data.

- **Text mining** – The ability to identify what content is, what it is about and what facts and relationships it contains. Examples of text mining in use are:
  - extracting the medication and quantity used to treat a disease and the outcome
  - identifying which company acquired another company, for what price and who was the adviser
  - discovering what equipment failed, the symptoms and the resolution
  - detecting what opinion a customer has about a product or service.

  Once extracted this information is stored as data as a series of RDF triples.

- **Ontology management** – Ontologies are used to describe what is important to a domain, business or industry, the relationships that exist between things and the vocabulary used to describe them. For example NASA's ontology describes missions, systems, research, equipment and facilities while a healthcare provider's ontology describes symptoms, diseases, treatments, hospitals and physicians.

  The ontology is used to power the text mining process, harmonize information of different types and sources and to enhance the way users search, explore and analyze information.

Text mining and ontology management are designed to make sense of information in a consistent way. They reduce confusion and provide order to unstructured data—something that traditional relational data models have long struggled to do. With 80% of the world's data being unstructured, this task is critically important to all organizations.
RDF Triple Stores

The Semantic Web is often referred to as a web of linked data because it focuses on the relationships among facts, which provide context. The core component of linked data is a data format called RDF triples (Resource Description Framework). RDF triples contain a subject, predicate and object. Two examples of triples are:

- Sears Corp. (subject) owns (predicate) Kmart (object)
- Aspirin (subject) treats (predicate) fever (object)

These triples can be linked to form a graph-like representation without hierarchy, are machine readable, and can be used to infer new facts about the world.

The standard language for writing triples is RDF, and the standard query language is SPARQL. Both RDF and SPARQL are World Wide Web Consortium (W3C) standards that are designed to describe and relate data on the Web and inside enterprises. W3C also defines other semantic Web standards for rules languages (Rule Interchange Format, or RIF), a language for marking up data inside Web pages (RDF in Attributes, or RDFa), and schema and ontology languages for describing concepts and relationships (RDF Schema, or RDFS; and the Web Ontology Language, or OWL).

When using these standards to link data, organizations gain a wealth of knowledge that even goes beyond the limits of their own data. Trillions of triples—all facts about the world—are freely available to download or query. And, even within the boundaries of an organization, triples can be combined to infer new facts. Organizations are using linked data to form knowledge graphs built using defined vocabularies with rules for handling data—all using a standard format.

The Benefits of a Semantic Platform

A robust semantic platform can be used to solve very complex problems that traditional technologies cannot. Organizations are using the combination of Smartlogic and MarkLogic as a semantic platform to get better answers from today's data, creating an intelligent data layer that powers smarter applications. Some of the specific benefits you can achieve include the following:

- **Extract more value from your content** – Use text mining to automatically extract concepts, entities, and facts such as people, places, and things. This data is contained in unstructured data but can be extracted and stored as triples. As triples, the data can be indexed and searched in real-time.

- **Make sense of your data in context** – Provide context through the use of ontologies. Even if the data is not natively stored as RDF, it can be semantically enriched so that a user can discover facts and information more intuitively.

- **Aggregate and link disparate data sources** – Store your data natively as RDF triples, and use triples as the glue to link documents. You can also use triples to "decorate" existing data with annotations that describe it.

- **Provide more relevant search** – Enable granular searches such as, "everything an analyst said about a company in relation to an employee that worked there at a specific time" or do a general search for “kind of what you are looking for” and get accurate, relevant results.

- **Discover new relationships** – Create inferences (i.e. new triples) based on existing data using rule-based reasoning across hundreds of billions of triples, and visualize relationships to discover new insights.
"Through 2015 organizations integrating high value, diverse, new information types and sources into a coherent information management infrastructure will outperform their industry peers financially by more than 20%."  

Gartner

Smartlogic Semaphore and MarkLogic

Smartlogic Semaphore and MarkLogic provide an enterprise semantic platform to harmonize your data and provide new insights. Semaphore provides the data and the interface to interact with it, and MarkLogic stores the data.

Semaphore is a content intelligence platform that enables organizations to model what is important to their business with ontologies, describe their content with metadata, harmonize internal and external data, and leverage the ontology and metadata to augment search and analytics.

MarkLogic is a powerful Enterprise NoSQL database that can natively store documents, metadata, and triples—without the need to define schemas. And, with built-in search, MarkLogic also allows for powerful combination queries across multiple data types using a single query.

<table>
<thead>
<tr>
<th>Capability</th>
<th>Semaphore</th>
<th>MarkLogic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology Manager</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Ontology Query API with NLP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Auto-extraction of concepts, entities, facts, relationships, sentiment</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Mapping workflows</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Semantic data store for asset metadata</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Schema-less content store</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Search over metadata, triples and content</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Hadoop connection</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Multilingual</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Enterprise scalability</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The Origin of Linked Data

Where do triples come from? That is where entity extraction and ontology management are important. Semaphore has entity extraction, which provides the ability to extract data and entities, and can auto-classify content. It has entity extraction for more than 30 entity types in more than 20 languages, and can extract names,
places, measures, noun phrases, and much more. It is also highly scalable, includes a rule-based text categorization engine, and has support for open standards.

Once you have the data in a standard format, it also has to be meaningful. You need taxonomies, ontologies, models, or controlled vocabularies, that relate data by defining a standard vocabulary for events, people, concepts and things.

Ontologies map linked data to broader categories of data. For example, when the computer processes hundreds of millions of documents and records it is because of ontologies that "Name" means a "Customer's Name" versus an "Employee's Name."

Ontologies map linked data to broader categories of data. For example, ontologies can map a "name" field in one database to "customer name" or "employee name" in other data stores across hundreds of millions of documents.

Ontologies also help us solve linguistic ambiguities. For example, when the word "money" is written in a U.S. publication, it refers to the value of the dollar, whereas in Great Britain, "money" refers to the British pound. Ontologies can map those differences and apply the appropriate metadata during auto-classification. There are all sorts of ontologies for specific domains — in healthcare, financial services, publishing, and many other industries.

Semaphore includes an ontology manager that helps create the ontological classifications used to model the links and structure between language elements. With Semaphore multiple users can create, enhance and browse all types of semantic models, whether they are lists, controlled vocabularies, taxonomies, thesauri, or ontologies.

Semaphore Ontology Manager

Semaphore's Ontology Manager allows organizations to create and manage their own ontologies, as well as leverage linked open data, which are domain-specific standard vocabularies that can be mapped to your organization's own requirements to quickly establish a custom ontology. Organizations can further accelerate ontology development by applying advanced text mining techniques to sample content to identify concepts, vocabulary, and relationships that are important to the organization.

Once a semantic model is in place, Semaphore auto-classifies content and identifies the concepts, topics, products, resources—and more—contained in it and adds metadata or creates triples that describe the content and information. All of the data is then stored in MarkLogic, including all of the metadata generated by
Semaphore, bypassing the issue of database alignment. This makes it possible to crosswalk between multiple systems, sources and even organizations.

Organizations can combine the data generated by Semaphore to answer questions such as:

- Why are certain products failing and what is the best resolution to the problem?
- What treatments are most successful for patients with a given condition?
- How are customers talking about our brand, products and competitors?
- Which customers are likely to leave us and why?

Organizations can use this ability to govern, organize, and search content—unlocking knowledge and providing new insights.

**MarkLogic Semantics**

MarkLogic is an Enterprise NoSQL database that can store and query across a combination of XML or JSON documents, structured data, and RDF triples. It is also designed with the enterprise in mind—it has massive speed and scale, high availability and backup recovery, and government-grade security.

**Documents + Data + Triples**

MarkLogic is the only Enterprise NoSQL database platform that can store and query a combination of documents, data and triples. This single platform gives you the flexibility to choose the data model that works best to store your data and provides the ability to query across everything holistically.

**Holistic Search**

You can embed triples in documents, triples can refer to documents, or triples can connect documents. In any case, users can search across data with a single query. In the real-world, this gives you critical flexibility. For example, on Data.gov.uk, there are 10 primary data formats, including documents, XML, CSV, and RDF, but only 1 percent of the data is in RDF. Only MarkLogic can store and query the XML, CSV and RDF data together within the same semantic store.

**Speed and Scale**

MarkLogic has a specialized triple index to ensure querying triples is fast. MarkLogic also has a triple cache to better manage the use of memory to ensure optimal performance at scale. Some triple stores insist you store the whole triple store index in-memory, but MarkLogic uses memory-mapped index files to maintain speed without the limitations of physical memory.
Both of these features—the specialized triple store and the triple cache—make MarkLogic a scalable, elastic, high performance triple store. With other triple stores, volume quickly becomes an issue. Some triple stores claim to scale in clustered systems, but for parallel query only—that is, they can have three node clusters but only if each node has the same data on it. MarkLogic’s shared-nothing architecture supports elasticity and scalability.

MarkLogic can store 1 billion triples per node, at about 350 bytes per triple, and can scale to hundreds of billions of triples.

**Enterprise Features**

MarkLogic’s triple store comes with all of the features MarkLogic has built and proven over the past decade—including ACID transactions, scalability and elasticity, high availability and disaster recovery, government-grade security, and performance monitoring tools. With semantics in particular, government-grade security gives users the ability to define exactly which users are able to see which triples by using Role Based Access Control (RBAC).

**Smartlogic and MarkLogic in Practice**

**Investment Banking: Faster, Smarter Customer Onboarding**

A Tier-1 investment bank onboarding clients has to process more than 100 types of documents for thousands of customers per year, extracting and cross referencing millions of facts between documents and other data stores.

The bank uses Semaphore to identify each kind of document, extract its most important facts, and provide these as triples for ingestion into MarkLogic. MarkLogic stores the original unstructured documents as well as all the extracted facts, which are cross-referenced with other data. When the data does not match—or suspicious facts are found—exception workflows are triggered in MarkLogic so that analysts can investigate further. At the end of the process, the Bank is confident that it can meet its compliance obligations and trustworthy reference data is ready for the enterprise.

**Healthcare: Improving Patient Safety**

A healthcare provider needs to reduce the number of falls, slips and trips among clients older than 75. After that age, the consequences of falls tend to be more severe. Injuries are often followed by a loss of independence and the need to move from home and into care accommodation.

The provider uses Semaphore to harmonize the vocabulary from internal and external data sources—both structured and unstructured. Examples include: drug data showing important side effects such as drowsiness, loss of balance or reduced concentration; medical history showing use of medications and injuries affecting mobility; and visit reports showing risks in the home such as steep stairs, loose rugs, and exposed cables or icy steps.

The combined information that includes documents and triples to describe risk factors is stored in MarkLogic, where powerful combined queries are used to identify high risk individuals. The system also links risk factors to suggested actions that could be taken to reduce serious accidents. Changes in medication, provision of mobility aids, and low cost home improvements are recommended to prevent injuries and save lives.
Additional Resources

This paper provided a high level introduction to what is possible using the powerful combination of Smartlogic Semaphore and MarkLogic together as a semantic platform to get more value from your data.

If you are interested in learning more, check out the resources:

- **Smartlogic Knowledge Center** – [http://www.smartlogic.com/home/knowledge-zone](http://www.smartlogic.com/home/knowledge-zone)
- **MarkLogic Website** – [http://www.marklogic.com](http://www.marklogic.com)

If you have additional questions or want to talk to a sales representative, please contact us at info@smartlogic.com.