IDC TECHNOLOGY SPOTLIGHT

The Next Generation of Data Integration Enabled by Semantic Technologies

August 2016

by Julia Neuschmid and Thomas Vavra, IDC #CEMA41696616

Sponsored by PoolParty Semantic Suite

In a world characterized by an ever-increasing volume and variety of data — especially unstructured data — managing data in efficient ways and gaining value from it is essential. This paper examines the need for and business value of semantic technologies, and looks at the role of PoolParty Semantic Suite in the market.

Introduction

The exponential growth of data has a multitude of organizational implications. Smart data analytics solutions lead to improved allocation of resources, increased revenue based on better customer insights, and more efficient product and service life cycles. Data fuels the evolution of new business models, and has high strategic relevance across industries.

However, enterprise architectures often lag behind in terms of providing agile data solutions to internal and external stakeholders. Organizational data silos prevent flexible data processing and an immediate transformation to a highly dynamic business environment. Even though unstructured data accounts for 90% of all information (according to IDC), unstructured data, especially in the form of text-based documents, is almost completely left out of organizations' data management frameworks.

Data integration initiatives are challenging due to a variety of factors:

- Historically developed enterprise architectures hardly fit with modern business requirements, and resistance to replacing or modifying existing IT systems remains strong. However, the huge volume and variety of data cannot be used to its fullest potential if it is locked up in data silos.
- Most organizations lack an overview of their digital assets as metadata management is not consistently conducted.
- The lack of standards-based data models complicates data interoperability and makes it an expensive challenge.

Organizations at the forefront of their respective industries have started to extend their enterprise architectures with semantic technologies. Leading companies from sectors including healthcare and pharmaceutical, finance, media and publishing, and ecommerce, are establishing knowledge modeling and graph-based approaches at the core of effective data management. Semantic software solutions give data scientists and subject-matter experts broad leeway for structured and unstructured data operations across platforms. Together with enterprise architects, these specialists are shaping highly customized digital environments for their organizations. The internet serves as a model, as semantic technologies are based on methods and tools from the World Wide Web Consortium (W3C) that have been adopted for enterprise needs.

Semantic Technologies: No Black Box At All

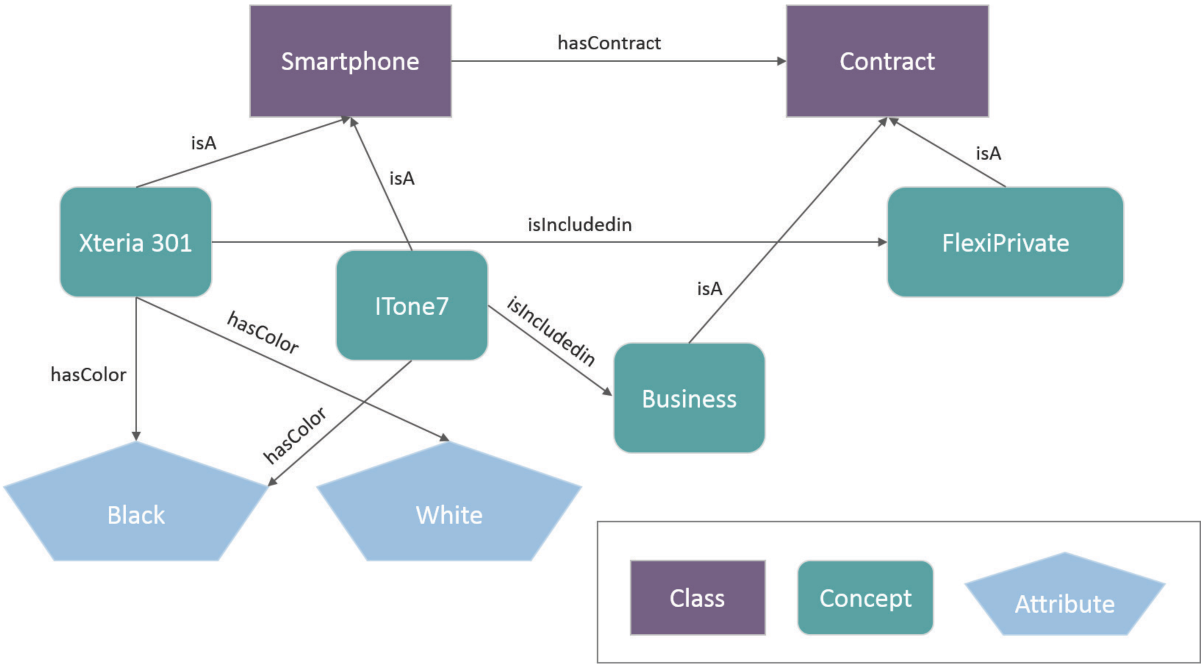
Semantic technologies provide a complete toolset for enriching structured and unstructured data with metadata and establishing a contextual framework for organizational data repositories. Documents are analyzed on an entity-level, and this entity-centric view is the foundation for providing dynamic applications such as semantic search, personalized content services, and knowledge discovery portals. The standards-based technology stack ensures interoperability. Below we outline how the main components of semantic technologies work:

The Enterprise Knowledge Graph

A knowledge graph is a highly efficient way of aligning back-end data management with smart end-user applications. Each graph represents a knowledge domain. It defines the relevant entities and provides a consistent vocabulary for enterprise applications. Leading companies in the field of web search and social media build their solutions based on knowledge graphs. Examples include matching the capabilities of business professionals with like-minded individuals, job positions, and skills requirements, or managing a network of people, interests, and lifestyle offerings. Enterprises can also adopt this information management approach by establishing knowledge models for their specific business domains.

FIGURE 1

Knowledge Graph Example



Source: IDC, 2016

As shown in Figure 1, a telecom operator could model its product, service, and contract portfolio in a knowledge graph. All relevant business objects and possible attributes are included only once. Precisely defined terms such as 'FlexiPrivate' for a type of contract are then related to other relevant business entities. The graph-based approach provides different kinds of querying possibilities. By traversing a graph, you can run queries as follows: "Show me all contracts of FlexiPrivate that include an XTeria 301 smartphone."

Relational databases represent a knowledge domain laid out in tables and rows, with relations between business objects stored in tables. The technical overview of a knowledge domain does not reveal the semantics behind the data. Queries have to be defined upfront, and cannot be decoupled from the schema behind. When names or relations change, a knowledge graph provides the opportunity for making the change once and then applying it to the whole interconnected knowledge model. In an extensive relational database, errors may occur due to the distribution of numerous entities in a variety of constellations. Interconnected data managed in a graph resembles patterns of human thinking, and lays the foundation for cognitive computing applications. Additionally, and most importantly, a knowledge graph can be developed and maintained by subject-matter experts. The IT bottleneck can be easily resolved by enabling experts to reorganize their data in the way that they require. When, in our example below, an 'ITone 7' is only available in white, the product manager can simply remove the attribute relation and the information is updated wherever it is used.

Knowledge graphs provide a wide variety of opportunities. In an enterprise context, it is essential to focus on the relevant entities in terms of the application scenario. The telecom operator in this example could extend the knowledge model by adding entities around customers and internal processes, which would enable queries such as: "Display all customers who had a 'Flexiprivate' contract with an 'XTeria 301' and switched to 'ITone 7'," or "Show me all key account managers for contract type 'Business' who have sold more than 50% 'ITone 7' as the device type."

Knowledge graphs generate linked data that is maintained in graph databases. It is important to stress that graph-based methods and tools are supposed to be used in addition to relational data frameworks. Static data sets are perfectly suitable to remain in established data schemes. Graph-based data management solutions can be beneficial, especially for dynamic data with high personalization requirements on an application level. The bigger the knowledge graph, the more linked data is created, and superior data processing operations along the graph can be executed on the fly.

Entity Extraction

Semantic data management solutions use a knowledge graph in combination with an entity extraction component. Digital assets in different formats across various platforms are matched against the knowledge graph, and classified according to the entities they include. The variety of data formats can include PDF, DOC, CSV, XML, and HTML files. In a data pipeline, the resources get transformed into Resource Description Framework (RDF), which is a standard model for data interchange on the web. This process harmonizes the different metadata schemes and various vocabularies. Finally, the digital assets are enriched with metadata that is derived from the enterprise knowledge graph. This way, content is consistently enriched with semantic data that is machine-readable. Available information can then be automatically reused in different contexts based on an evolving semantic layer around data repositories.

Semantic Layer

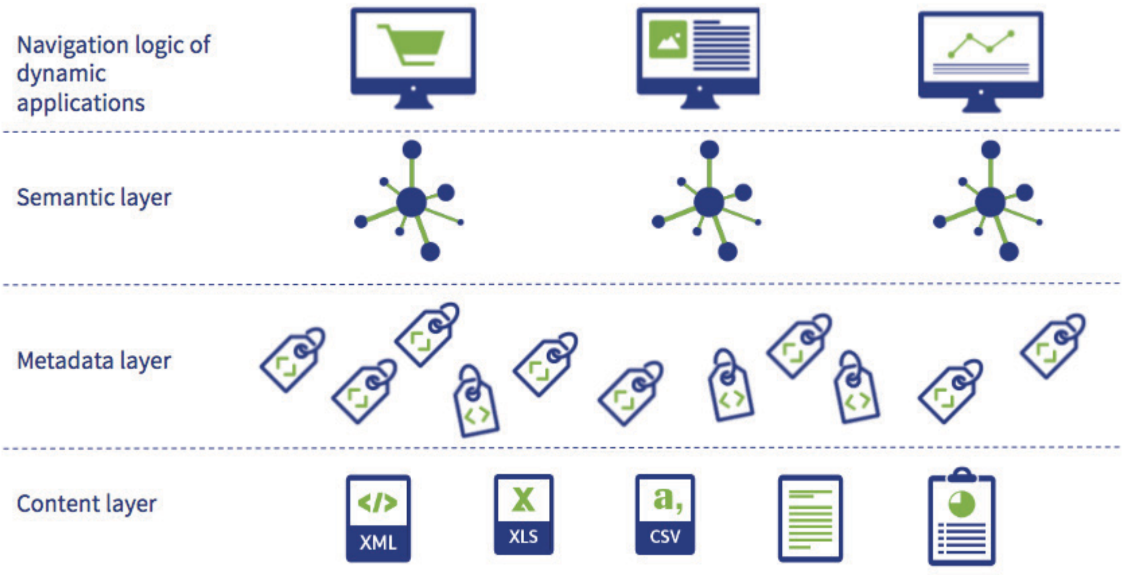
The semantic technology components that are needed to create a semantic layer are available in the form of semantic middleware. This means that organizational challenges caused by data silos can be removed without actual data migration.

The evolving semantic layer is embedded in a four-layered content architecture, and is managed separately from the actual content and document management system. All content types can be transformed into RDF and matched against the knowledge graph to be enriched by consistent metadata. By changing entities in the knowledge graph, content metadata gets automatically changed, which also changes the content display in different dynamic applications.

The semantic layer is a single access point for metadata management, and provides a 360-degree view of business objects and their usage.

FIGURE 2

Four-Layered Content Architecture



Source: Semantic Web Company, 2016

Semantic Technologies Standards for Information Management

Most organizations face the challenge of having multiple proprietary systems in place that cannot communicate with one another. This is why a growing number of companies are embracing standards-based technologies that make data integration easier and also prevent vendor lock-in. Today, the maturity of semantic methods and tools allows corporations to benefit from internet technologies and standards that fulfill demanding enterprise security requirements when bundled with commercial software solutions. The most important semantic technologies, concepts, and methods are:

- **RDF:** The Resource Description Framework is a standard model for data interchange on the web that can also be used in a corporate environment. A variety of schemes can be mapped with RDF, which enables data processing irrespective of the underlying data model.
- **URI:** The Uniform Resource Identifier is used to identify an entity. Every concept is included only once in a knowledge model, but can be linked multiple times depending on the context.
- **SPARQL**: SPARQL is the RDF query language, and enables application developers and data scientists to query and traverse graphs.
- **SKOS**: The Simple Knowledge Organization System is a standard for building knowledge models. It includes a set of modeling rules and is based on RDF. As its name suggests, SKOS can be used by subject-matter experts without specialized skills in knowledge modeling.
- **OWL**: The Web Ontology Language is an advanced knowledge modeling standard that is also based on RDF. OWL enables knowledge engineers to model complex domains in the most specific way.

The Fusion of Cognitive Systems and Semantic Technologies

Cognitive systems are the next generation of smart applications. The underlying technologies are based on statistical computing and highly advanced machine learning algorithms. The ultimate aim is to reproduce human thinking in the digital environment and provide knowledge workers and consumers with smart assistants, including for answering questions and providing recommendations (see *IDC PlanScape: Implementation of Cognitive Systems*, IDC # US41477516, June 2016). Obviously, cognitive systems and semantic technologies share the same goal. By joining forces or adapting tool sets from each other, the common vision can be realized.

Cognitive systems focus on developing knowledge domains and making them actionable in a highly automated manner. Semantic technologies have a qualitative approach at their core, with highly customized concept-based knowledge models created by subject-matter experts. Personalized content services derived using fully statistical methods still often prove to be inaccurate, as human thinking does not apply algorithms. Graphs that develop through concepts and relations mirror more precisely how the human brain functions.

Mature taxonomy management software that enables companies to develop their own knowledge graphs provides semi-automatic features to extend evolving knowledge models. Cognitive systems can accelerate the process of adapting algorithms for entity extraction, provided that they rely on information management that makes the meaning of data explicit.

Barriers between technology disciplines are blurry. It all comes down to practitioners and how they are using available methods and tools to deliver data-driven solutions. The semantic technology stack is well-established in a growing community that has proven successful in implementation projects across various industries. These experts are in high demand and are interested in enhancing their practical skill sets in new implementation scenarios.

Semantic Technologies in Practice

There are numerous use cases for semantic technologies that are applicable to all digitally enabled industries. Healthcare and pharmaceutical, media and publishing, ecommerce, finance, and government organizations are well known for using the semantic technology stack to address their specific challenges related to disruptive industry changes. Below, we present four generic application scenarios to highlight the flexibility of semantic technologies:

- **Dynamic semantic content publishing:** The entity-centric view of text-based information facilitates the dynamic publication of content assets across multiple platforms. This opens up enormous innovation potential for the whole publication life cycle. Configurable landing pages can display varying digital assets depending on subject-matter parameters. With graph queries such as “Display a picture, a video, a blog entry, and the most up-to-date news about XY,” organizations can reuse existing content on the fly and create new digital products. Knowledge workers can search for information and the context surrounding it more precisely, which also makes content creation more efficient.

- **Personalized content recommendations:** Semantic recommendation engines are highly precise. A knowledge graph can model customer interests in advance, and relate them to products, services, and informational materials. This can be enriched with algorithms that take into account an individual's browsing behavior and buying history. The personalization of the customer journey depends on high-quality data that is modeled with regard to relevant processes.

- **Knowledge discovery portals**: Knowledge-intensive organizations benefit from semantic knowledge discovery portals. Resources that are consistently tagged can be easily reused. Knowledge graphs include entity synonyms and relational distances between entities, so that implicit information is also detected. This has a substantial impact on research activities, as the ability to work with information is automatically enhanced.

- **Deep analytics**: Semantic technologies bring structured and unstructured data together, which revolutionizes data analytics. Semantic applications discover and display the meaning of unstructured data. Analyzing unstructured data in combination with quantitative data and methods introduces new business intelligence capabilities. Deep analytics can provide answers to questions as: “How frequently are certain types of complaints issued by clients with a business contract that includes smartphone XY?”

Success Story: Semantic Technologies at Boehringer Ingelheim

The pharmaceutical company Boehringer Ingelheim is a research-intensive organization that employs nearly 48,000 staff worldwide. The company's research and development (R&D) departments are geographically dispersed, which makes both knowledge sharing and performance management equally challenging. As a result, a scientific information tracking service based on semantic technologies was implemented.

One of the main outputs of the R&D departments are scientific publications that are publicly available. Important key performance indicators (KPIs) for research work are the quantity of published papers and the impact factor of each paper. The impact factor is derived from peer reviews and the reputation of the scientific journal that accepted the paper.

The semantic application fetches the relevant publicly available resources and processes them using semantic middleware. Today, the head of research has an analytical dashboard at his disposal that provides extremely precise answers to questions such as:

- Who has published papers about disease X?
- How many papers have been published about disease X over the years?
- What impact did the papers that were published by subsidiary Y in a certain therapeutic area have?

Depending on the structure of the semantic layer, the analytical functionalities can be flexibly extended.

Essential Guidance for Semantic Technology Adoption

When organizations embrace new technologies, they must navigate a learning curve. Technology experts are interested in enhancing their enterprise architecture with semantics, but struggle to convince top management to invest in innovation that has a fundamental impact on how data is managed and used. Below, we outline the benefits of a semantic enterprise architecture, and indicate how to avoid common project pitfalls:

How Organizations Benefit From Semantic Technologies

- The graph-based technology approach enables agile data processing of structured and unstructured content.
- Semantic technologies are standards-based. Organizations avoid vendor lock-in and can start building a sustainable organizational data management infrastructure.
- Semantic technologies are mature. Fortune 500 companies successfully merge a variety of data schemes with RDF, significantly reducing the cost of data integration.
- Companies can hire and obtain support from experienced semantic technology professionals sourced from a large and continuously expanding community of experts.

Recommendations When Starting With Semantic Technologies

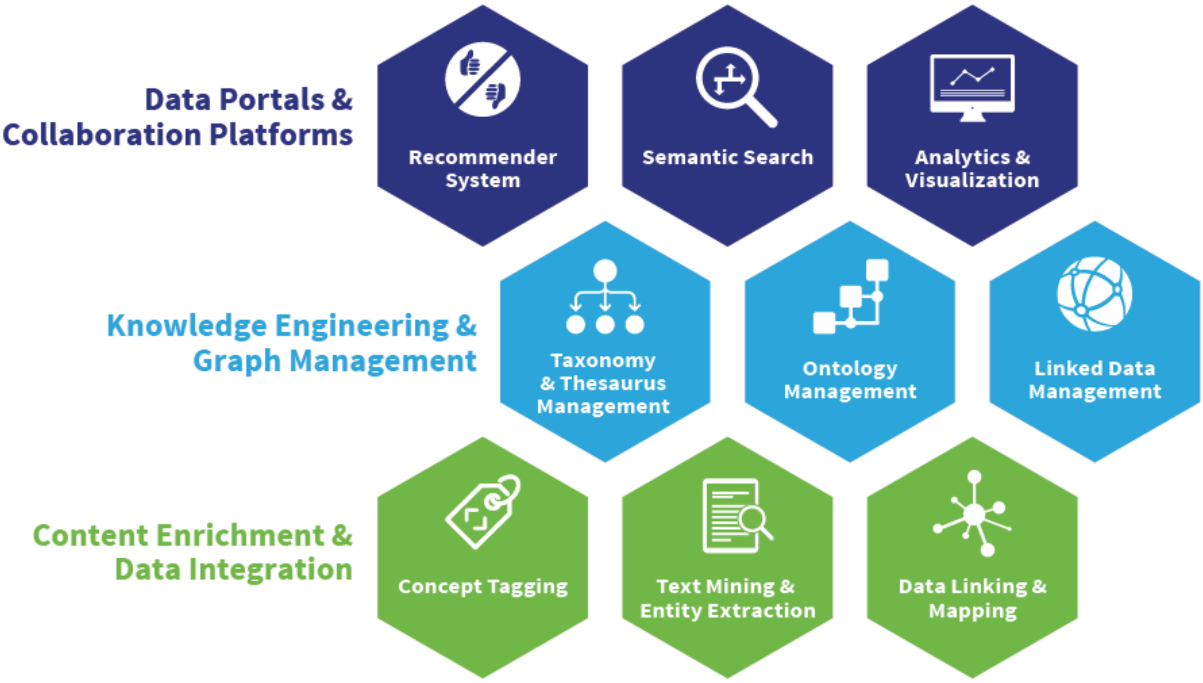
- Linking data is an ongoing effort that increases system intelligence over time. Start with a concrete application as the immediate project goal and a clearly defined data set. Extend the enterprise's linked data initiative over time.
- You do not need to build knowledge graphs from scratch. There are libraries of industry-specific knowledge models that can be licensed and customized to fit your needs.
- Semantic technologies are brought to life by cross-functional project teams. Make sure that you have semantic solutions in place that can also be easily used by subject-matter experts.

Considering PoolParty Semantic Suite

PoolParty is a full-blown semantic technology platform provided by the Semantic Web Company, a pioneer in the semantic web since 2004. PoolParty supports enterprise needs in information management, metadata management, data analytics, and content excellence. Customers from various industries such as finance, government, pharmaceutical, and media benefit from linking structured and unstructured data.

PoolParty Semantic Suite consists of nine highly configurable modules that can be flexibly combined. The platform covers the whole data integration process up to turnkey semantic applications. The solution is completely standards-based. PoolParty's application programming interfaces (APIs) enable organizations to integrate semantic technologies based on widely used web technologies.

FIGURE 3


Data Portals & Collaboration Platforms
Recommender System
Semantic Search
Analytics & Visualization
Knowledge Engineering & Graph Management
Taxonomy & Thesaurus Management
Ontology Management
Linked Data Management
Content Enrichment & Data Integration
Concept Tagging
Text Mining & Entity Extraction
Data Linking & Mapping

PoolParty Semantic Suite

Source: Semantic Web Company, 2016

With PoolParty, subject-matter experts can create knowledge graphs with different levels of complexity. Only minimal training is required. The text mining components are constantly enriched by machine-learning algorithms, and provide highly precise entity extraction. We expect growing demand for semantic technologies and graph-based data solutions in the future. Integration with leading graph database vendors positions PoolParty as semantic middleware that is at the center of the semantic technology ecosystem, and among the leading players in the market.

Related Research

- *IDC PeerScape: Practices for Cognitive Systems Initiatives* (IDC #AP41156315, May 2016)
- *IDC PlanScape: Implementation of Cognitive Systems* (IDC # US41477516, June 2016)
- *IDC TechScape: Cognitive Systems Technologies* (IDC #US41005816, February 2016)
- *Worldwide Data Integration and Access Software 2016 Top 10 Predictions* (IDC #US40332615, January 2016)
- *Worldwide Data Integration and Integrity Software Forecast, 2016–2020* (IDC #US40696116, June 2016)

ABOUT THIS PUBLICATION

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by
various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

For more information on IDC, visit www.idc.com. For more information on IDC Custom Solutions, visit http://www.idc.com/prodserv/custom_solutions/index.jsp.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com