Enterprise Information Integration (EII) is gaining significant credibility in the market as the most efficient approach to linking back-office data with line-of-business content. This business problem is not a new one and has seen multiple approaches being adopted that have sought to achieve this aim but have typically failed to efficiently address a number of significant challenges. These challenges are seen as: the diverse formats of content, the disparate nature of content and the need to derive ‘intelligence’ within the business process from this content. What is required is a more holistic view as to how information can be used to create smarter business processes or applications, irrespective of the location, source or format of the information. The most effective, non-invasive approach to achieving this utilises semantic metadata and is known as Semantic Enterprise Information Integration (SEII).

Outlined in this white paper is a concept known as Semantic Metadata that is central to SEII and paves the way to finally realize liquidity in information and generate tangible business value through the creation of task-specific applications or services.

Metadata and semantics are both proven approaches; metadata has long been used to “tag” documents to make them readable to Enterprise content management and search tools, Tim Berners-Lee’s vision for the next stage of the Web’s evolution is termed the Semantic Web, where semantics plays a pivotal role. Applied together in the form of semantic metadata, enables confident operational decision making across the enterprise. By not just tagging information with syntactic metadata, but by annotating or ‘enhancing’ information with semantic metadata, software programs can automatically understand the full context of what that information means and can make correct decisions about who uses the information and how. This paper explores how semantic metadata meets the current and future informational requirements of an enterprise.

**APPROACHES TO CURRENT INFORMATION CHALLENGES**

There are three main challenges facing enterprises that are trying to gain value and competitive advantage through their information. First is the need to manage the full spectrum of information sources, from structured data through to unstructured content. The growth in the

**SYNTACTIC vs. SEMANTIC**

**Semantic metadata**: Metadata that describe contextually relevant or domain-specific information about content (in the right context) based on a domain specific metadata model (e.g., industry-specific or enterprise specific) or ontology is known as semantic metadata. For example, if the content is from the business domain, the relevant semantic metadata could be company name, ticker symbol, industry, sector, executives, etc., whereas if the content is from the Intelligence domain, the relevant semantic metadata could be terrorist name, event, location, organization, etc. Metadata that offer greater depth and more insight 'about the document’ fall under the semantic metadata category.

**Syntactic metadata** describes non-contextual information about content, focussing on elements such as size, location or date of document creation providing little or no contextual understanding of what the document says or implies. This level of metadata is often the extent of many content management technologies.
variety of types and formats of information is a reflection of the penetration of content centric applications in today’s enterprise and the increased use of richer media types. The second challenge relates to the fact that enterprises have typically deployed multiple point solutions to solve discrete business problems at a business unit level. This typically results in content and data being decentralized and residing anywhere across the enterprise. Finally, there is the business need to integrate this information more directly into the business process from where ‘business intelligence’ and value can be realized.

The areas of IT that has typically sought to address these challenges have been polarised between structured data management and unstructured content management, and as such have been seen as two distinct markets. The structured data market is reaching maturity with large well-known vendors dominating the market, whereas unstructured content management is still in its infancy. The vendors in this latter market have concentrated on creating structure around unstructured content so that it can be ‘filed’ in the correct place for subsequent retrieval. There are many approaches adopted but most are based on traditional offline information management techniques, such as the construction and use of taxonomies.

By combining these two approaches (structured data and unstructured content management), the enterprise is creating the liquidity in information that enables smarter business processes or applications to be created. It is this that generates tangible value within an enterprise and is addressed by the use of semantic metadata in support of Semantic Enterprise Information Integration.

REQUIREMENTS FOR SEMANTIC ENTERPRISE INFORMATION INTEGRATION

Semantic metadata plays a critical role in satisfying a number of requirements that customers are asking from information integration and analysis software:

- Integration of structured data from back end systems with unstructured content from inside or outside of an organization into a business process
- Unified view of relevant information in support of a business process
- Identification of key pieces of information and relevant knowledge (entities such as people’s names, places, organizations), and the relationships held between these entities
- Analyse and co-relate extracted information so as to provide tangible business insight from a large volume of content – showing not only the ‘what’ but also the ‘why’
- Enable high levels of automation in the processes of extraction, normalization and maintenance of knowledge and content within an enterprise
- Make efficient use of the extracted knowledge and content by providing tools that enable fast and high quality (contextual) querying, browsing, and analysis of relevant information

“Semantic metadata is a key enabler of text analytics to derive business value from information”

Dr Amit Sheth, CTO - Semagix

It is only through the use of semantic metadata that an enterprise can deliver on these requirements and create liquidity in its information assets in order to create business value.
CREATING SEMANTIC METADATA

In order to extract optimal value out of a document and make it usable it needs to be effectively tagged by analysing and extracting relevant information of interest. Some of the techniques used by Unstructured Data Management (UDM) vendors go some way to achieving this and revolve around the following:

(a) Dictionary and thesauri: match words, phrases or parts of speech with a static or periodically maintained dictionary and thesaurus. Dictionaries such as WordNet can be used to identify and match terms in different directions, finding words that mean the same, are more general, or more specific.

(b) Document analysis: look for patterns through statistical space vector techniques, co-occurrences and application of pre-defined rules to find interesting patterns, within and across documents.

These approaches however fall short of delivering actionable value back into a business process. The approach of using ontologies is seen as the most effective way of delivering this value and is described as follows:

“Ontologies capture domain (application, industry) specific knowledge, including entities and relationships at both a definitional level (a Company has a CEO), and at an instance or assertional level (Meg Whitman is the CEO of EBay). An ontology is not a hierarchical representation of concepts but expresses domain knowledge though the rich relationships held between concepts and entities.”

Ontology-driven metadata extraction, is the most flexible and comprehensive approach as it allows modelling of fact-based, domain-specific relationships between entities that are at the heart of semantic representations. The evolution of these approaches is highlighted in Fig 1.

![Figure 1: Types of Metadata enabling Business Analytics](image)

An example of semantic metadata extraction
One scenario of the use of semantic metadata is within a financial services company, where large volumes of information are subscribed to and generated throughout the business process, and is used in a variety of functions from analysts through to equity sales. Semagix’ semantic metadata approach helps create new or better services that support these functions. Analysts can be better served with a unified view of all the information that is specific to them, which combines structured information such as real-time stock quotes with unstructured news items, and in addition, the distribution of the reports created by the analysts is enhanced through the annotation of these documents with relevant information. Equity sales people can be supported through the integration of structured data, from CRM systems, with relevant internal and external information (including their own analysts’ reports). This would assist them in understanding who they should be speaking to following an event such as a market announcement, what information would service their customer best, and an understanding of what content their clients have already received.

This is achieved through the creation and management of semantic metadata. The semantic metadata created for the document in Fig 2 below would include both direct relationships and indirect relationships. In the above example the direct relationships extracted are; BEA Systems, Microsoft and PeopleSoft all engage in the "competes with" relationship with Oracle. Depending on the internal configuration, the expert domain agents can further enhance the extracted entities with semantically associated entities from the ontology. An example of this would be; HPQ and HD are traded on the NYSE and BEAS, MSFT, ORCL and PSFT are components of the NASDAQ 100 Index.

The next stage is to identify the 'indirect relationships’. The use of semantic associations allows entities not explicitly mentioned in the text to be inferred or linked to a document, by incorporating such associated entities in the tagging of the document. This one–step–removed
linking is referred to as "indirect relationships." The relationships that are retained are application specific and are completely customisable, and whose inclusion makes it possible to traverse `relationship chains` to more than one level from within the document.

It is very important to keep in mind that the above-mentioned semantic metadata is only useful because it is specific to the domain of ‘Business’. This means that entities that fall outside of this domain are not contextually relevant to the overall document and are therefore not included. In other words, the domain-specificity of the semantic metadata elements is key to establishing the right context and relevance.

CONCLUSION

The value of metadata has long been recognized, right from data integration to application integration. It is however only with the progression from concentrating on syntax and structure to semantics (see Figure 2 above), that an increase in the control and the creation of business insight from information will occur. This is achieved through the creation and use of accurate semantic metadata, through the use of an ontological approach, where meaning and context are associated with information. It is this focus on semantic metadata that is undoubtedly the key to overcoming the significant information challenges in the provision of actionable information and business insight within the framework of Semantic Enterprise Information Integration.

ABOUT SEMAGIX

Semagix has developed patented, semantics-based enterprise information integration (EII) and knowledge discovery technologies. Using its semantic metadata approach, Semagix lets enterprise customers integrate and extract insights from their structured and unstructured information assets in order to conceive and develop smarter business processes and applications. The Semagix approach has proven itself in solutions as diverse as airline passenger

Semagix Freedom™ – “Semantic EII”

Based on its patented semantic technology, Semagix’s platform - Freedom, defines a new semantic approach to Enterprise Information Integration (EII) that allows enterprises to integrate and extract insights from their unstructured and structured information assets, in order to conceive and develop smarter business processes and applications.

Freedom manages the lifecycle of enterprise semantic metadata and exploits its structural efficiency. This lifecycle is defined as:

1. Aggregation of structured and unstructured content. Freedom uses intelligent content agents to extract, normalize and disambiguate data from either structured or unstructured content. This process is automated and the intelligent agents can be scheduled to update the content and knowledge at prescribed intervals.

2. Semantic enhancement of content through the use of a domain-specific ontology. The content is then classified and semantically enhanced through the use of a domain-specific ontology. Freedom provides the flexibility to build the ontologies, either from scratch (using powerful, easy-to-use toolkits), or by importing an existing taxonomy. Non-programmers (usually Knowledge Engineers or Domain Experts) can easily use the Freedom toolkits to address their information modeling requirements, down to a very granular level of detail. Included are sophisticated techniques for entity disambiguation that identify distinct entities, regardless of syntax, and the ability to create synonyms.

3. Knowledge Discovery through the Semantic Query Engine. The Semantic Query Engine generates and presents a unified view of knowledge and content about a particular query, inclusive of the obvious and non-obvious relationships held between different entities. These insights can be revealed through either browsing or through a visualization front-end, and whose results are returned in near real-time.

4. Metadata integration. Many applications throughout an enterprise can be enhanced through the use of semantic metadata. Freedom facilitates this integration through a series of tool sets for a broad range of adaptors, allowing the functionality of existing applications to be extended.
Semagix White Paper

risk-assessment for airports, anti-money laundering (AML) solutions for the financial services and legal sectors, and a market segmentation and resource allocation model for the music industry.


FURTHER INFORMATION


