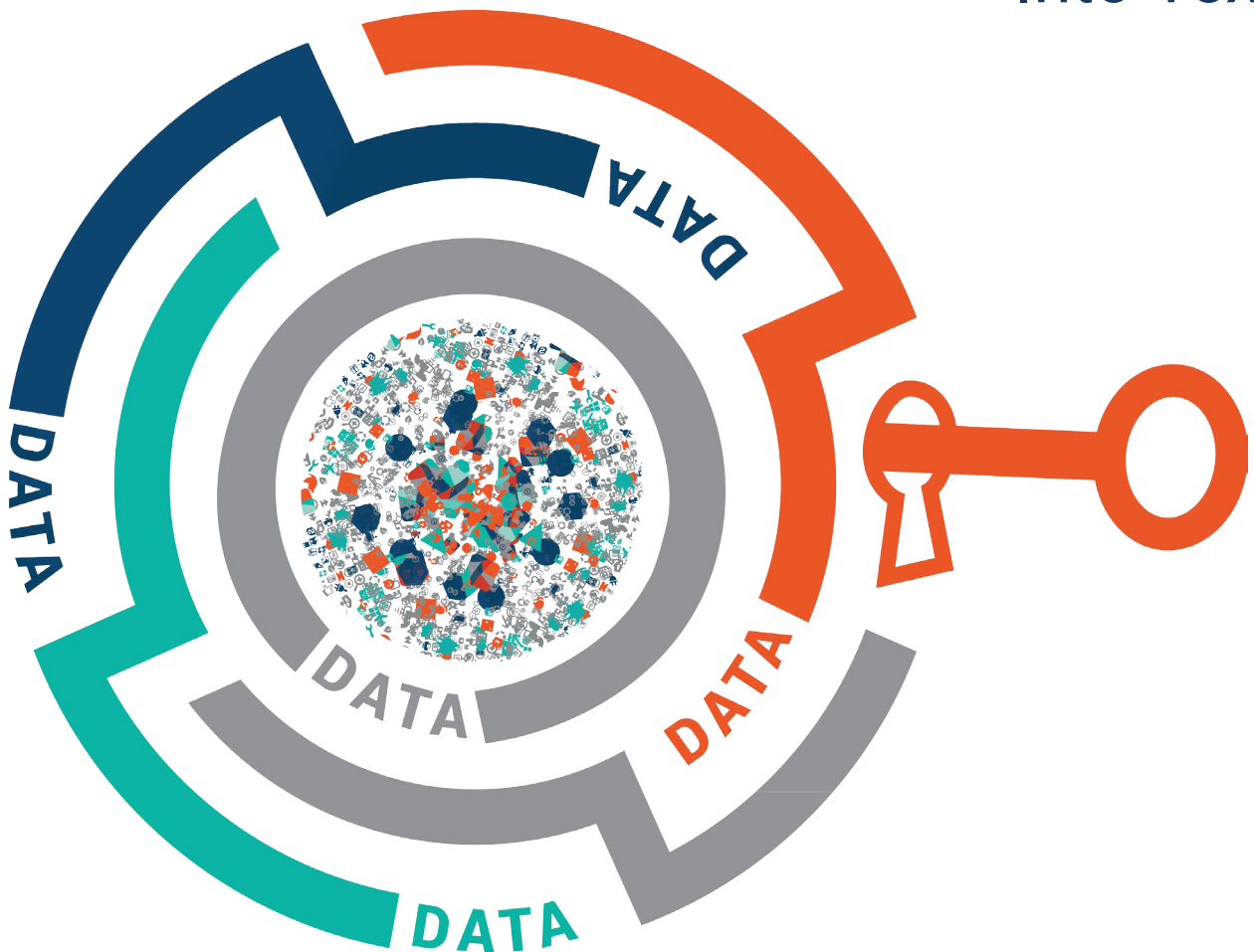




# Text Analytics for Enterprise Use

Why Interweave  
Semantic Data  
Into Texts



**INTRODUCTION**

**3**

**GAINING INSIGHTS THROUGH  
TEXT ANALYTICS**

**5**

**TEXT ANALYTICS AND  
SEMANTIC ANNOTATION**

**9**

**HOW TEXT ANALYTICS ENABLES  
THE EFFICIENT USE AND  
MANAGEMENT OF DOCUMENTS**

**14**

**CONCLUSION**

**19**

# INTRODUCTION

## THE RISING NEED FOR TEXT ANALYTICS

Today an overwhelmingly big part of the world's information exists in a textual form: business records, government documents, legal acts, social media streams, clinical trials, medical archives, emails, blogs. Such rapid increase of digital texts (across Internet and on Intranets) causes the rising need for text analytics and brings forward the question of finding a smarter way for reading and understanding texts, ultimately for deriving knowledge out of them.

With the many transformations the written word has gone through - from the oldest preserved inscriptions on clay tablets to the present astounding amount of documentation, stored in cloud systems (or other repositories), one thing remained unchanged: **the information our textual sources contain is only as good as our ability and tools to extract and interpret it.**

## THE NEW READERS ON THE BLOCK

In the increasingly textual environment we live and work text analytics is of critical importance. If we are to use texts with maximum productivity and minimum wasted effort, we should consider encoding machine-readable information in them. Only then computers can assist us with the huge variety and velocity of the textual flows.

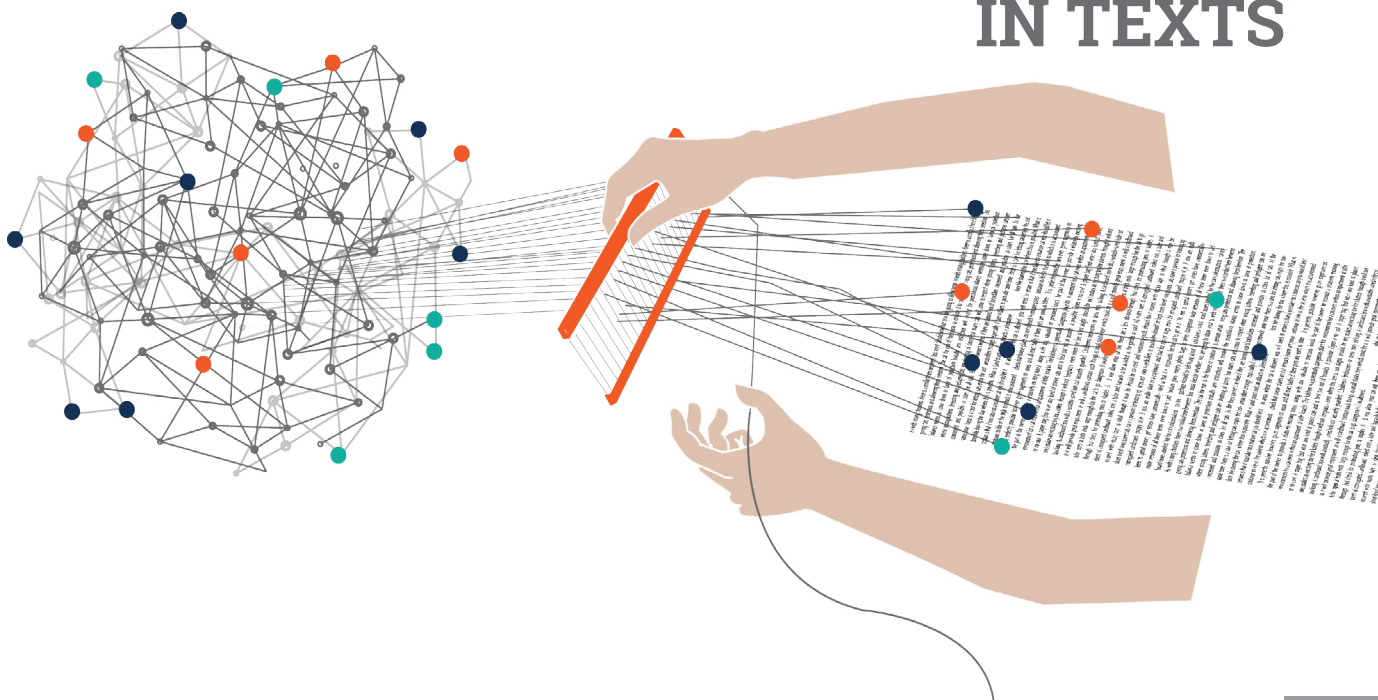
Software agents reading our texts might not be among the common readership we picture when thinking about presenting, storing and reusing information but they should be.

Efficient research, quick sifting through information and facts discovery in the age of data are unthinkable without the help of algorithms. On their analytical powers to process vast bodies of textual sources and present us with results, pertinent to our needs, depends a huge part of the information we extract, the knowledge we discover and most importantly the insights we arrive at.

**Deriving high-quality, structured and machine-manageable information from texts is what text analytics is all about.**

**To unearth relevant information we are now to look at reading anew, from the vantage point of its very essence - as extracting information from any encoding system.**

# ENTER DATA INTERWOVEN IN TEXTS





## WHY WEAVE DATA INTO TEXTS?

Weaving text and data together comes naturally to meet the need for deriving well-organized actionable information from unstructured, heterogeneous textual sources.

Applied properly, using semantic web technologies, the process of text analytics interconnects data and documents elements to help organizations meet one of the biggest challenge in today's digital environment: unstructured data. When data and text are interlinked, content becomes highly-manageable, easy-to-use and properly structured.

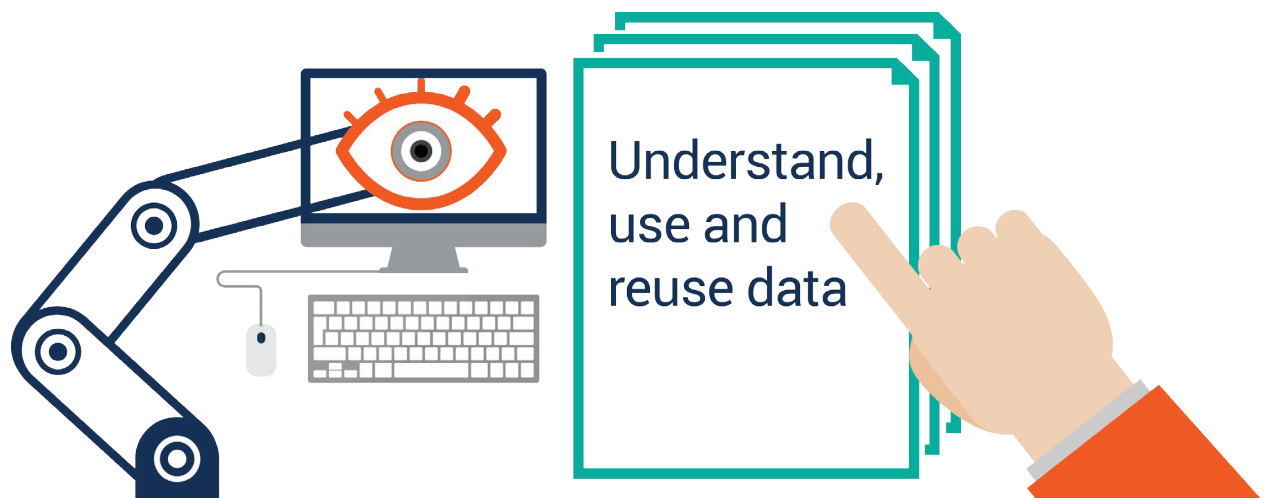
## SEMANTIC WEB TECHNOLOGY

The Semantic Technology defines and links data on the web or within an enterprise by developing languages to express rich, self-describing interrelations of data in a form that machines can process. Thus, machines are not only able to process long computing strings of characters and index tons of data, but they are also able to store, manage and retrieve information based on meaning and logical relations. Semantics adds another layer to the web and is able to show related facts and items instead of just word matching.

## SEMANTIC TECHNOLOGY AT A GLANCE

Semantic technology is used to define and link data (on the web or within an enterprise) by developing languages to express rich, self-describing interrelations of data in a form that machines can process. Thus, machines are not only able to process long computing strings of characters and index tons of data, but they are also able to store, manage and retrieve information based on meaning and logical relations.

The core difference between semantic technologies and other technologies for data, the relational database for instance, is that the semantic technology deals with the meaning rather than the structure of the data.



Text analytics, or turning texts into data pieces and further interlinking them, consists of a number of methods and processes, semantic information extraction and semantic annotation being the key ones. Using various algorithms to analyze the free flowing text, chunks of it are transformed into structured interconnected data elements. This enables organizations to easily and effectively use information, track and understand relationships in disparate textual sources, find topical information, discover facts.

## TEXT ANALYTICS IN ONE SENTENCE

Extracting structured data from unstructured texts.

Weaving data into texts through text analytics techniques results in:

- texts readable for machines
- improved information retrieval
- connectivity on multiple levels
- highly-manageable chunks of knowledge



## CONNECTIVITY MATTERS: FEW WORDS ON LINKED DATA

As a substantial part of the text analytics process, linked data techniques not only enrich the text and its data but also forge the creation of future-proof, highly interconnected textual assets.

## Linked Data Explained

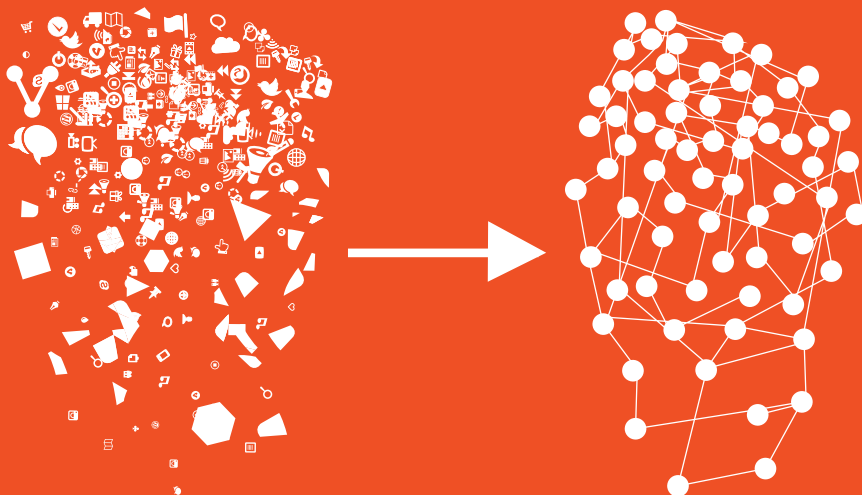
Linked data is a term denoting structured data that is interlinked and connected through W3C open standards. Linked Data technologies (for more details see Tim-Berners Lee note Linked Data) are unique for they allow us (via our software agents) to travel across diverse data sets, to find, share and integrate information easily and effectively.

Linked Data breaks down the information silos that exist between various formats and brings down the fences between various sources. Linked Data makes data integration and browsing through complex data easier, due to the standards it adheres to. Once standardized, enterprise data is then turned into smart data pieces that seamlessly connect with numerous other data pieces.

Linking data to selected chunks of texts is key to opening content up to more efficient access, use and management. Think of linked data as “machine-readable pieces” floating around the outer space of a text making it ready to connect automatically to other related texts or chunks of texts.

When everything is connected and disparate sources are integrated, a vantage point of a unified, 360-degree view emerges. One can see relations between all kinds of business records and enterprise-related content: from marketing and sales documentation to internal business records, service, supply chains and product information.

With data integrated in documentation textual sources are given yet another layer of meaning and assembled into a collection of machine-readable enterprise content to the point where



**CONNECTING  
THE DOTS  
BECOMES  
CONNECTING  
THE NODES**

# How Text Analytics Enables the Efficient Use and Management of Documents Across Enterprises

Semantic technology is of immense help when it comes to creating, curating and using textual sources. A set of universal standards, developed and agreed upon by the World Wide Web Consortium (W3C) international community, semantic technologies help enterprises discover data, infer links and extract knowledge from enormous sets of raw data in various formats and from various sources.

The use of semantic technologies for text analytics brings content creation and delivery to the next level where information discovery is done smoothly and effortlessly. Organizations of various size and domains use this approach to effectively meet the challenges that volume, diversity and incoherency in documentation pose before them and their content management systems.

**Through text analytics the tethered threads of fragmented, siloed and partially viewed information are carefully untangled and woven into an opportunity for integrated, consistent and efficient content management.**

With text analytics enterprise knowledge management and discovery processes are streamlined and texts unchained from the tyranny of inefficiently stored, managed and used (or rather not used) content. Handled and interwoven smartly in our human readable texts, data pieces help organizations know content (both internal and external) at a granular level of detail and thus be better at what they do, more knowledgeable to take data-driven decisions and informed enough to thrive in a hyperconnected environment.

# How Text Analytics Enables Use of Documents Across Enterprises

Text Analytics for Enterprise Use

Well-organized, systematically arranged and linked sets of machine-processable information can be leveraged for all kinds of purposes: from semantic search for documents retrieval, through automated tagging and dynamic presentation to automated topical clustering and questions answering.

## Typical applications of text analytics in an enterprise context are:

- Business intelligence
- Scientific research
- Content classification
- Personalised recommendation
- Risk analysis
- Fraud detection
- Customer service records analysis
- Regulation compliance
- Drug safety compliance
- News production and delivery

Turning texts into data coupled with linking these data to other sources adds value to content at a comparatively low cost. Text analytics methods allow information to be presented, recorded and retrieved in a common format and thus made ready for cost-effective management and seamless integration.

There are many industries that benefit immensely from “tidying up” their data and adding a data layer to textual sources. Among the domains that mostly benefit from the advances in the field of text analytics the ones where knowledge is mostly formal:

- Life Science
- Scientific publishing
- Media and content publishing
- Pharma and Healthcare
- Banking
- Financial Services
- Insurance
- Legal and Compliance
- Digital Humanities

Examples of companies that took the leap into creating smart content are:

**FT, Bloomberg, Euromoney, John Wiley & Sons, Oxford University Press, IET BBC, DK, - AstraZeneca, Foundation Medicine.**

# When connecting the dots is about connecting the nodes:

## How high-quality data forges pharmaceutical research

Forward-thinking companies understand the need for concrete steps towards integrating data from diverse sources and using the way these data relate to each other in order to transform them into value.

Case in point, facing the need to gather information from a broad range of biomedical data sources in an iterative way, [Astra Zeneca](#) found a solution in Linked Data. The researchers of the company needed a mechanism which will allow them to mine all data scattered among different relevant resources and to identify visible (direct) and invisible (distant) relations between biomedical entities studied along the pharmaceutical research and discovery process.

### **Linked Data is what provided AstraZeneca's researchers with such a mechanism.**

Creating a platform for interactive relationships discovery, called Linked Life Data, the bio-pharmaceutical company were able to obtain high quality research data out of their unstructured documents. The platform integrated over 25 data resources and aligned more than 17 different biomedical objects: genes, proteins, molecular functions, biological processes/pathways, molecular interactions, cell localization, organisms, organs/tissues, cell lines, cell types, diseases, symptoms, drugs, drug side effects, small chemical compounds, clinical trials, scientific publications, etc. The Linked Life Data was capable of identifying explicit relationships between entities, categorizing them to causality relation ontology. It also mined unstructured data to identify relations hidden within text.

This Linked Data solution was of immense help to researchers assisting them to get an overview on the existing relationships within the significant volume of diverse scientific and clinical data and connect the dots as to generate or expand a certain theory, test hypotheses, and make educated, informed assertions about which relationships are causal, and about exactly how they are causal.



# Text Analytics Benefits at a Glance

Data extracted from documents and linked to selected chunks of texts is beneficial to both creating and reusing all kinds of written content. Applying text analytics techniques to create machine-readable textual sources substantially enhances:

- Information access (through semantic search)
- Decision making (through integration of disparate and seemingly unrelated sources)
- Research and Development (through uncovering hidden relationships)
- Knowledge management (through aggregating all relevant information)
- Knowledge discovery (through automatically discovering references to concepts and entities)
- Content production and delivery (through interlinking text with big data)

## The Truth About Text Analytics: Formal Knowledge Wanted

Very often, busy chasing meaning per se, a major point in interweaving data into texts is missed and it is making information from textual sources manageable, through data.

When it comes to text analytics, the most important understanding not to be swept under the rug, is that this approach is not a silver bullet for discovering knowledge and teasing meaning out of data. Algorithms still have really hard time understanding texts the way we human readers do. However, with time our machines are getting better at completing well-defined, measurable, widely understood tasks, in which interpretation is a matter of computation. This is why **turning textual sources into data assets is best applied in the areas where knowledge is explicit and multiple, and ambiguous interpretations are rare.**

# How Text Analytics Enables Use of Documents Across Enterprises

Text Analytics for Enterprise Use

“ In the context of texts for enterprise and organizations’ use the interpretations should be performed automatically by machines in a strict and predictable fashion. This requires a formal definition of the interpretation and, because of this, a formal definition of the context. ”

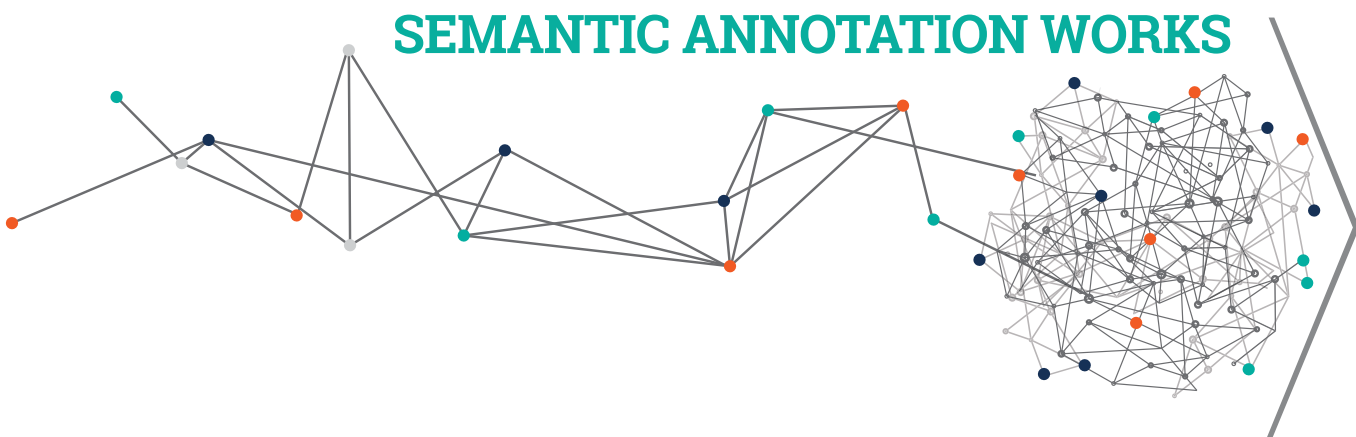
**Wiley**

Text analytics is nothing but a specific tool for taking informed decisions based on large scale textual resources that leverages semantic technology. It is as simple and straightforward as that. At the end of the many and complex set of practices and processes behind text mining and information extraction there is a key process, called semantic annotation.

**IN THE NEXT SECTION**

**YOU WILL LEARN HOW**

**SEMANTIC ANNOTATION WORKS**



# TEXT ANALYTICS AND SEMANTIC ANNOTATION

One of the main technologies through which text analytics reaches its overarching goal - turning text into data, is semantic annotation. Annotating texts with machine-readable and manageable information in the form of linked data pieces is what transforms textual content into high-quality information.

## But what exactly is semantic annotation and how does it work?

What is Semantic Annotation? **(A Metaphor to Take Away)**

Semantic annotation is a tool that gives us the ability to express, refer to and thus make documents and parts of texts machine-processable. As we saw in the previous chapters, our machines, the new readers on the block, need context to understand texts. Just as we do. Just as medieval readers did.

## THINK MARGINALIA.

Most of the medieval manuscripts abound with explanatory notes for the sake of clarity. These marginal notes were used to add more meaning and context to a particular word or phrase in a text and were usually written in the language of the text, or in the reader's language if that was different.

**"Marginalia (or apostils) are marks made in the margins of a book or other document. They may be scribbles, comments, glosses (annotations), critiques, doodles, or illuminations."**

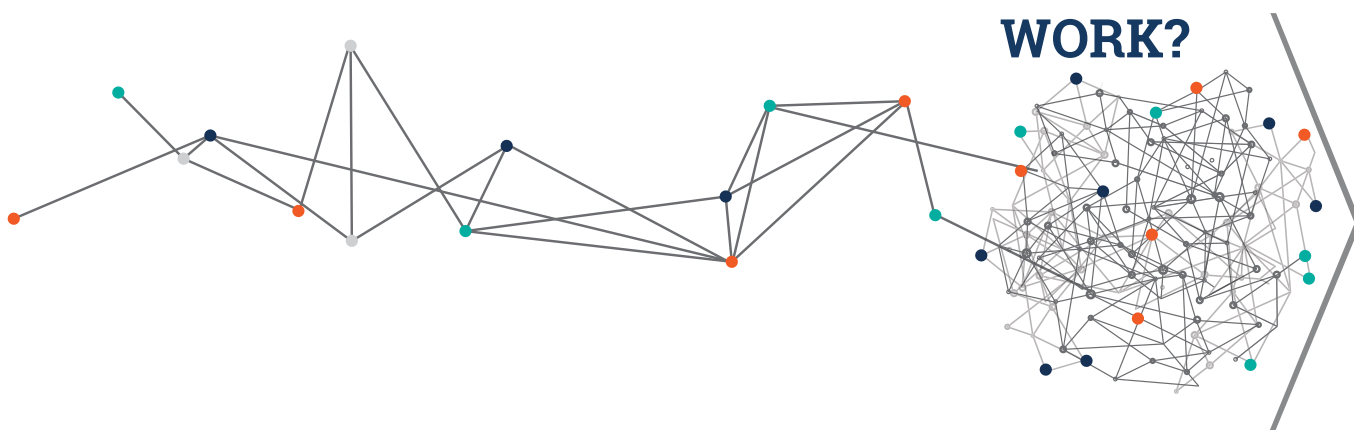
In the paradigm of semantic technology, creating digital marginalia for machines to read and process is about adding semantic metadata, linked to the whole document or to particular parts within it. Semantic descriptions can be added to any sort of text – web pages, regular (non-web) documents, text fields in databases etc.

**Leaving digital “margin” notes for the new readers on the block to use adds another level of meaning to our textual sources significantly improving our efficiency in accessing, using and reusing them.**

Semantically annotated texts are rich in machine-processable connections, that is in context and references, supplied in a readable by computers form. For if we want algorithms to understand textual sources we need to provide these machine readers with links to concepts with unambiguously defined meaning.

With that conceptual understanding of semantic annotation in mind, let’s now explore the details behind its processes.

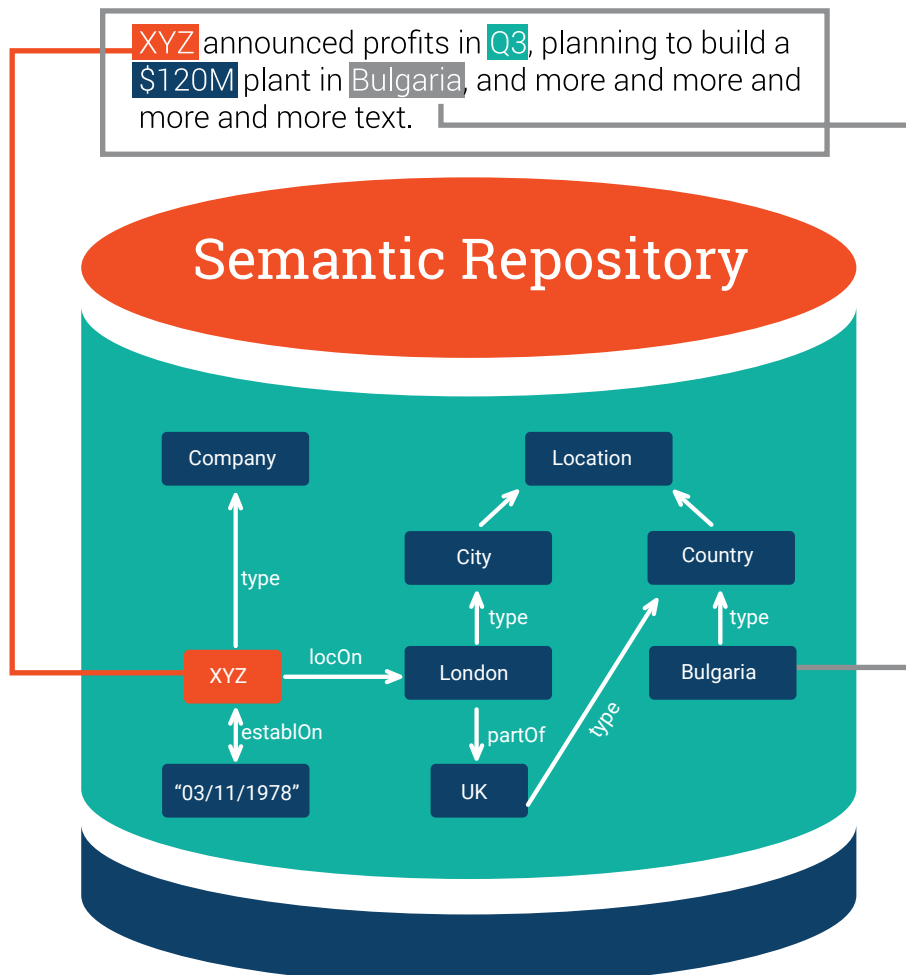
## HOW DOES SEMANTIC ANNOTATION WORK?



# SEMANTIC ANNOTATION UNPACKED

To add some technical details to the above conceptual understanding, semantic annotation, or semantic tagging, is about attaching names, attributes, comments, descriptions etc., to a whole document, document snippets, phrases or words. It provides additional information (metadata) about an existing piece of text. Compared to tagging, which adds relevance and precision to the retrieved information, semantic annotation goes one level deeper.

- It enriches the unstructured or semi-structured data with a context that is further linked to the domain structured knowledge.
- It allows results that are not explicitly related to the original search.

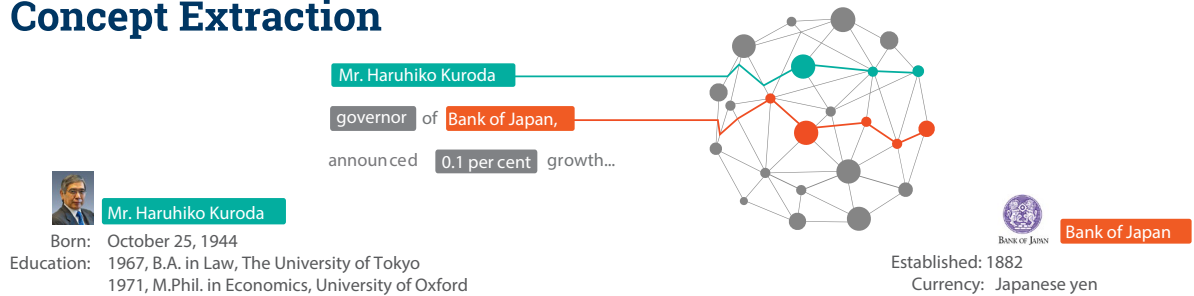


# 1 Text Analysis

Mr. Haruhiko Kuroda governor of Bank of Japan, announced 0.1 percent growth...

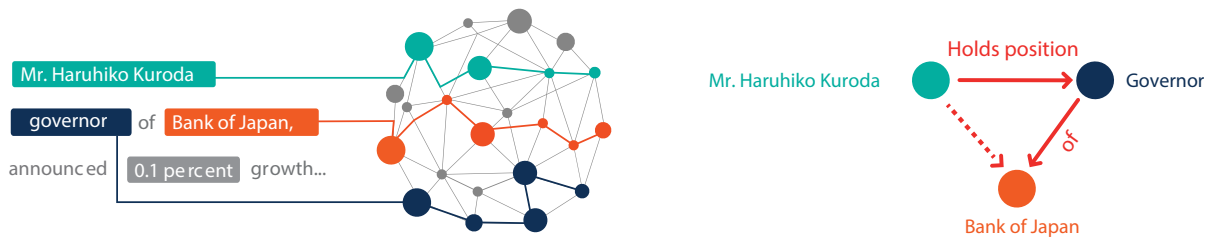
Once tagged, entities can then be recognised and have information from various sources associated with them. In this case the system knows that “Haruhiko Kuroda” is a type of ‘person’.

# 2 Concept Extraction



In order for the system to understand that ‘governor’ is a ‘job’ which exists within the entity of ‘Bank of Japan’, a rule must exist which states this as an abstraction. This is called an ontology (think of an ontology as the rule-book: it describes the world in which the source material exists). By telling a computer how data items are related and how these relations can be evaluated automatically, it becomes possible to process complex filter and search operations. Using the same example, the system is able to create a formal, machine-readable relationship between Haruhiko Kuroda, his role as the governor, and the Bank of Japan.

# 3 Relationship Extraction



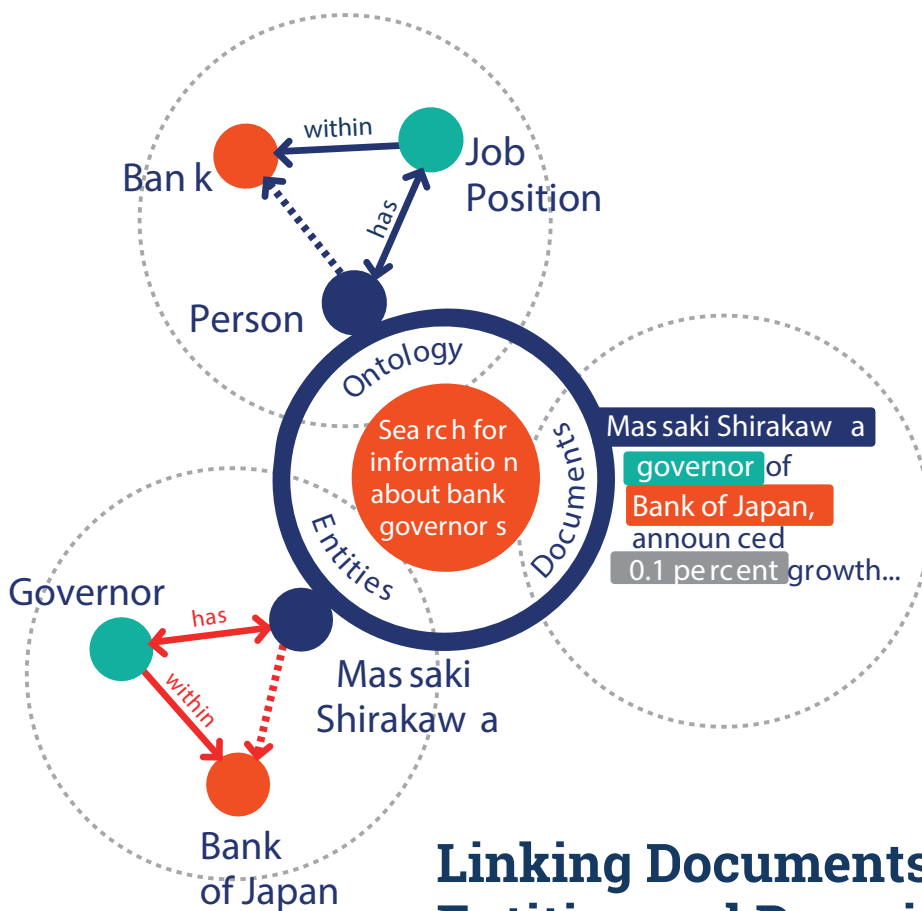
Once we have the annotations linked to the ontology and the background knowledge, we can begin to infer relationships between entities in a system that have not been directly linked by human action.

# 4 Indexing and Storing



Another example: a visitor arrives on the website of a newspaper and would like information about bank governors in Asia. Semantic technology allows the website to return a much more sophisticated set of results from the initial search query. Because the system has an understanding of the relationships defining bank governors generally (via the the ontology), it is able to leverage the entire database of published text content in a more sophisticated way, capturing relationships that would have been overlooked by computer analysis alone.

The technology allows the inference of relationships that are not specifically stated within the source material: because the system knows that Haruhiko Kuroda is the governor of the Bank of Japan, it is able to infer that he works with other employees of the Bank of Japan, that he works in Tokyo, which is in Japan, which is a set of islands in the Pacific.



## Linking Documents, Entities and Domain Models



# Letting Content Become the Knowledge It's Meant To Be

**now** news on the web by ontotext

## A broad and in-depth perspective of the world NOW

- > Discover new and interesting news, aggregated from various sources in categories.
- > Enjoy their semantically enriched content and get the advantage of the knowledge behind.
- > Focus on the trends to see the top mentioned concepts for a particular period of time.
- > Browse to see articles with similar semantic fingerprints.

Create account

Sports Science and Technology Lifestyle Business International

**Jeet and Jeetan show up for Williamson**  
Test cricket, as Faf du Plessis would explain, is most often won or lost on the extent to which a team can call on its experience. The senior players in the side – those who have seen similar situations

**Sydney Roosters vs Canterbury Bulldogs: NRL live scores, blog**  
The Sydney Roosters will look to continue their impressive start to the new NRL season when they play their first home game against the Canterbury Bulldogs who need to press the ignition button. Join

**Top trends** Today Week Month

Premier League	11 mentions
FA Cup	10 mentions
UEFA Champions League	10 mentions
Arsenal F.C	8 mentions
FC Bayern Munich	7 mentions

Semantically enhanced texts inaugurate a new era of content use. Using text analytics and semantic technologies, vast quantities of content scattered in various forms across documents of diverse formats are extracted, enriched and converted into manageable information pieces.

Activities, involving sifting through large bodies of writings (i.e. legal acts, enterprise documentation, regulations, scientific research) can now be scaled and automated (to the extent acceptable). Algorithms enter the processes of risk management, fraud detection, retrieving of facts and statistics, investigating connections, keeping up with compliance standards, tracking consumer behaviour and many more. Broadening the scope of business insight, machine-processable textual sources allow for efficient spotting of trends, revealing patterns, unearthing relationships.

## SEMANTIC ENRICHMENT

Text analytics removes the artificial divide between text and data and welcomes new assistants in our reading and writing processes - our machines. With text analytics, we open up large scale textual content coming from heterogeneous sources to machine processing in order to

expand our capacity for deep understanding, analysis and decision making. Interweaving data into texts, we give algorithms a key to the code of our communication system to enable computational powers to become a springboard for our creativity and synthesis.

Ready to learn more about how text analytics for enterprise use can help your business grow smarter?

**Contact us for  
a Free Consultation**

