Sustainable Advantage for the Investor Relations Team through Semantic Content

Daniel Hladky and MBA
Ontos AG
Switzerland

1. Introduction

Semantic web technologies promise to bring companies closer to their customers and deliver to consumers more relevant content than ever before. Two technologies in particular will help build sustainable advantage for the investor relations team. The first is natural language processing and second content enhancement. Intuitively, semantic content should help establish a higher quality of communication between information providers and consumers. This chapter describes the state-of-the-art in digital text information extraction, specifically the application of semantic technology to confront the challenges of the investor relations department. We discuss the roots of human language technology and ontology-driven information extraction and how such extracted semantic metadata can be used for better decision making, market monitoring and competitor intelligence. We will consider ontology as a sound semantic platform for defining the meaning of content and consequently supporting the prudent access to data for business intelligence (Hladky, 2008). Examples are given on dynamic hypertext views (Staab et al. 2000), a solution that links different web pages together based on their semantic meaning. The foundation of the proposed solution relies on an ontology-driven information extraction approach, a framework that merges same entities and stores the semantic metadata in a knowledge base. This framework supports the complete transformation process, including web page crawling, the extraction of knowledge, the creation of unique identifiers and presentations offering access to the portal. In this context, we describe how these technologies are being used in real customer scenarios and compare the classical search approach to a more intelligent approach based on ontology and information extraction. In particular, we describe semantic indexing (Khoroshevsky, 2005), building a knowledge base from various sources and give an introduction on how to create domain ontology based on customer queries. Then we tackle issues of merging information from text with semi-structured information from the Web, highlighting the relation to Linked Data (Berners-Lee at al. 2006) using standards like RDF/XML. Finally, we present possible user interfaces which display the aggregated semantic metadata inside a portal and other third party software tools. The chapter concludes by looking beyond the current solution to how semantic technology will add more information in the near future, including a short survey of recent thinking that offers potential extensions to today's model.

2. Profile of Investor Relations

A company, public or private, has a duty to inform its shareholders about ongoing activities, and strategies as well as secure current investments or even raise new money. A large number of companies have introduced a new organizational unit called the investor relations (IR) department led by an Investor Relations Officer (IRO). While execution is key to a company's success, the way in which a public or private company communicates its value to potential investors or purchasers is central to the price individuals are willing to pay. If vital elements are omitted from a company's story, or if this story is not told convincingly, true value will most likely go unrecognized. The art of positioning and communicating a company's story and investment proposition to investors is at the heart of attaining fair value. The list below offers a cursory view of the IR team's major activities:

1. Determine the company's investment proposition.
2. Identify and target the appropriate investor audience.
3. Screen the competition and identify your company's position in relation to it.
4. Develop communications platforms for presentation to the targeted investor audience.
5. Build relationships with the targeted investor audience and maintain consistent communications

The IR team is at the centre of communications, building a company story and brand (Kotler, 2006). Imagine a scene where company executives explain the organization's actual position, value and strategy at the general meeting (GM) to the shareholders. It's natural that shareholders will ask critical questions or question elements of the presentation. It seems obvious that the IR team does their homework by analyzing all major forces (Porter, 1998) that can influence the company and be prepared to answer these questions. However, in order to prepare effectively, the team needs to understand the competitive environment. We will focus on the task of screening the environment and will not investigate other elements of the communication chain described in Fig. 1. In order to properly evaluate the competitive environment, the IR team needs to classify various areas and determine which parts require deeper analysis.

Fig. 1. The art of communicating value

Central to any company's success is the ability to sell products or services in a highly competitive marketplace. Achieving this objective leads to more revenue, better margins and a higher company value for shareholders. Consequently, the primary purpose of the IR team is to screen day-to-day information from external and internal sources and create an aggregate picture. The results of the screening are not only useful for IR but serve as input for the management team's daily tactical decisions.

Table 1 shows a list of variables that should be considered when screening the environment.

About the company	• Sell-side analyst opinions
	• Factors driving investor interest
	• Stock information
	• Media coverage
	• Blogs and chat rooms
	• Shareholders
	   o Who are they
	   o Where are they from
	   o Investments in other companies

About competitors	• Shareholder information
	• Analyst coverage and comments
	• Stock information
	• Market share
	• Financial metrics
	• Changes in competitors
	   o Key management
	   o Business alliances
	   o New product launches
	   o Intellectual property
	• Conference participation
	• Lawsuits and litigation

About the industry	• General news • Trends • Identify new potential customers

Table 1. Competitive environment information

Having identified the main environmental factors requiring screening, the next task is identifying where such information can be retrieved. Obvious resources are search engines on the public internet. Additional special reports from analysts and company annual reports can be used for deeper analysis. However, most of the data that the IR team would like to investigate is unstructured. Gartner reported in 2002 that for at least the next decade more than 95% of human-to-computer information input will involve textual language. This large amount of information is growing, leading to the problem of keeping up with its pace. IR would need a huge number of people doing internet-based research to sort relevant from irrelevant information and compile findings into a complete knowledge database. All of this work is manual labour and time consuming. Knowledge Management (KM) and several other IT technologies have evolved that promise to facilitate this work, the latest being based on the semantic web (Berners-Lee, 2001) - a web of data that can be processed directly

or indirectly by machines (computers). Semantic web technology promises to reduce people-hours and costs while accelerating decision-making.

3. Knowledge Management and Semantic Web Technology

Knowledge Management (Fensel, 2005) is concerned with acquiring, maintaining, and accessing the knowledge of an organization (Fig. 2). Knowledge is now seen as a basic resource (Drucker, 1993) for an organization and consequently we need to understand its main building blocks and how they can be put to use. The aim is to galvanize an organization's intellectual assets for greater productivity, increased competitiveness, and better decision-making support. Given the large number of documents put online by organizations and the public, classical Knowledge Management systems have severe weaknesses:

- Searching information is based typically on keyword-based approaches that return irrelevant information rather than desired content.
- Extracting information through meaning-based searches requires human browsing and reading in order to garner relevant information.
- Automatic aggregation from various sources requires fast crawlers and human language technology to understand texts, allowing knowledge acquisition based on semantic annotation.

The IR team is mainly interested in a set of information as shown in Table 1. In order to have this up-to-date information, the system should be able to extract a large volume of information.
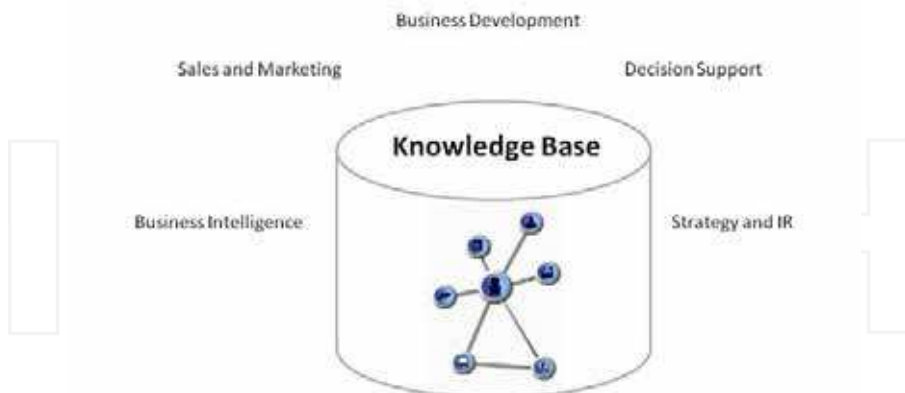

Fig. 2. Using the Knowledge Base within the organization

Thus, our Knowledge Management system needs to integrate heterogeneous, distributed and mostly unstructured or semi-structured information sources (Fensel, 2004). The core concern of our platform is steering clear of nonsense and maximizing computer support of the IR team in the aggregation and acquisition process. Before introducing the architecture's components, let's compare the keyword based search approach to the ontology driven information extraction approach.

3.1 Keyword based search approach

Imagine our IR team works for an automotive company and wants to analyze car news. A researcher might query the car brand "Jaguar" with a search engine (e.g. Google, Yahoo! or Live). The very rapid but imprecise result will be a huge set of pages containing the keyword "Jaguar" (Fig. 3). In many cases the user has to open the page in order to identify if the content is relevant. Recall is another weakness – it's unclear whether all relevant information has been found. To increase relevancy the user can extend his search expression by "jaguar+car" and reduce the result list dramatically. Let's take the example a step further and assume the researcher would like news about all car manufacturers that have reduced their working hours. What kind of search expression would the user need to find the relevant information? In this hunt, they would have to enter all relevant keywords that represent the meaning of reduced working hours.
¶

Fig. 3. Precision and recall in keyword based search

The user would then read the documents and extract the requisite information themselves. Then they might enter the data into the knowledge base system for additional analysis. In contrast to this information retrieval process (Davis, 2006) an ontology driven information extraction system would automatically populate the KM with the extracted data. Advantages of the information retrieval approach include speed and language independency. However, in situations where the user needs more precision and more complete recall without spending time reading text, another technology from the semantic

web space offers relief. Information extraction systems based on natural language processing will not only improve the search results but also contribute to knowledge capturing and sharing (Davenport, 2008). Additional value is generated when the new acquired knowledge is merged. The system will be able to identify hidden links and relationships, contributing to company knowledge.

3.2 Knowledge generation from text

We have not yet described the complete search engine transformation made possible by the Semantic Web or the Web of data (Berners-Lee, 2007). Most web pages are not using the suggested standards by the W3C (World Wide Web Consortium) in order for people or machines to find and correlate the information they need. Web pages are still mainly based on HTML coding which allows Web browsers (e.g. Internet Explorer, Opera, Firefox) to display a given string. Web pages or documents primarily use natural language text that represents content's meaning in a manner understood only by humans. A person reading a web page understands that the content has semantic data which describes, for example, a person, a company or the relationship between them. The IR team pursues the creation of knowledge in order to answer the queries listed in Table 1. Therefore we need to extend the information retrieval approach to a strategy that can extract the semantic metadata from text. The solution proposed is ontology-driven information extraction using Natural Language Processing (NLP). NLP is a computerized approach to text analysis based on both a set of theories and a set of technologies. This application analyzes text and presents only the specific information in which the user is interested. The system identifies concepts and relationships based on a given domain ontology (Gruber, 1993). In order to demonstrate what an ontology-driven NLP system can extract let's consider the following text:

"Google's founders Larry Page and Sergey Brin developed a new approach for online search."
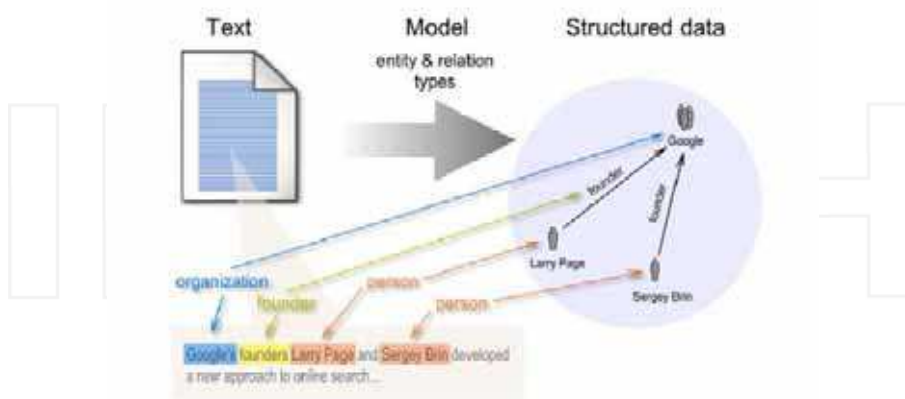

Fig. 4. Ontology-driven NLP

If a user wants to formulate a query like "Who founded Google?" than we need to understand the meaning of the text. Extracting the entities and the relationships (Fig. 4) from the text will help answer the query.

Based on a given domain ontology and set of rules the system will analyze the text and extract the information. The ontology-driven NLP will also create attributes like "FirstName", "FamilyName" and gender. Good NLP systems will also do reference resolution (e.g. discover that "he" in the following sentence is a reference to Larry Page: "Larry Page is working for Google. He also studied at Stanford University."). A substantial benefit of ontology-driven information extraction is that the system will automatically populate the ontology with new specific "instances" as they come up. Within the ontology, the system will describe concepts (e.g. person, location, company) and the relationships between the concepts. The semantic information extraction will then recognize these instances as persons (e.g. Larry Page, Sergey Brin) and populate the concept "person" with these names. Such an approach will allow queries to be formulated on an ontological level, for example "show all employees of company Google". The system will understand the query by recognizing that for the concept "company" and instance value "Google" we are looking for the concept "person" that is related via the semantic relationship "employees." Based on the extracted instances Larry Page and Sergey Brin, the query result will show those two values. Another crucial piece NLP can provide is a semantic graph (Khoroshevsky, 2005) or cognitive map for a single document at the end of the process. Such a process will merge "the same" instances together in order to grow the knowledge of connected information based on the semantic relations between the different ontology concepts. The same merging process can be executed for all documents in order to create a complete picture (Fig. 5).
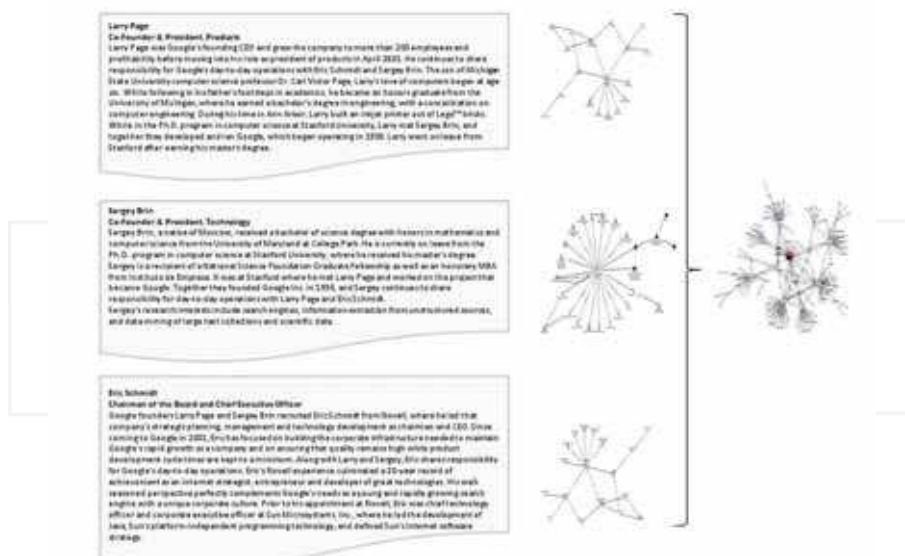

Fig. 5. Creating a Knowledge Base from a semantic graph

The final result of the merging process can be also described as the semantic index to all acquired information. NLP can be seen as the core process but in order to deliver a full solution we need another set of components in our framework.

4. Building the framework for IR

Because most web pages are still built with HTML, we need to define a framework that will allow the IR team to crawl, aggregate, merge and store semantic metadata. Besides analyzing text, we need a platform that can incorporate other data and convert it into new knowledge. The framework should work autonomously as much as possible and deliver the required background information to the IR user in a simple form through convenient user interfaces. As a first part of the framework, we need to define the knowledge model or domain ontology needed.

4.1 Building the domain ontology

Table 1's "competitive environment" assembled our main objects of interest. During the ontology engineering process (Fig. 6) we need to specify the domain ontology that we need in order to cover all possible user queries. Within the kick-off phase the main purpose is to define the general ontology application. This includes a list of possible dictionaries or a taxonomy related to the domain. After analyzing the input sources, the ontology engineer can decide which of them will be included. The lexical entries will be linked to the concepts/relations later on within the ontology. In this early stage one should also look for existing ontologies that can potentially be reused. The goal of the next phase is to develop the target ontology. This refinement takes the results of the kick-off phase as input, including the list of user queries, reusable ontologies and lexical data/dictionaries. The goal is to define the concepts and the semantic relationships between the concepts. Fig. 7 shows an extract of the domain ontology for competitive intelligence.

Fig. 6. Ontology engineering

The next step is the evaluation phase where we test the usefulness of the developed ontology. It is important to test the linguistic rules extracting objects/relationships from natural text, particularly when using NLP. During this phase it is also valuable to increase the quality of the NLP process. The main drivers for measurement are precision and recall (Porter & Cunningham, 2005). The goal is to tune the NLP system so both measures are high.
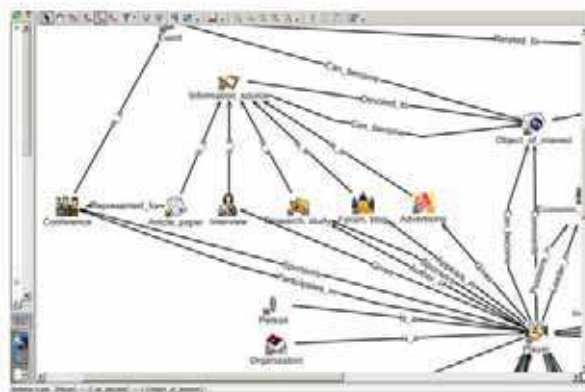

Fig. 7. CI domain ontology (extract)

In the case of this project, the testing of the NLP process was done using the environment of OntosMiner, the NLP engine from Ontos AG (see Fig. 8) and a simple web interface. The system shows the semantic annotation with mark-ups for the document. The left side panel shows the named entities that have been extracted and by clicking on marked text, the system shows the relevant facts on the right. Those facts are the extracted semantic relations of the text.

Fig. 8. NLP evaluation phase

After achieving a good level of quality with NLP, the next step of the evaluation phase is to link all components together by combining the framework's modules. The framework will then process the data and create valuable knowledge for the IR team.

4.2 Architecture

The Ontos Framework (Ontos Framework is provided by Ontos AG) that supports the solution for monitoring the competitive environment is based on a multi-layer architecture (see Fig. 9). Starting from the right, there is the presentation layer, the application layer (Web Services), the knowledge base, the semantic layer (OntosMiner) and the data source mapping section. This architecture provides the user functionality at run-time. A certain set of functions are used at configuration time, for example defining the external link to sources, creating the ontology engineering and the mapping of ontology to external databases.

Fig. 9. Framework-Architecture

Access to the framework is based on web widgets. This allows embedding the user interface for web pages, portals, smart phones, and desktop tools like Outlook.

4.2.1 The data source mapping layer

This layer comprises three levels of information gathering – the list of text documents, the internal databases and the external sources that have semi-structured or structured data. The first category of text files can be further divided into a list of subscribed RSS items from trusted sources as well as a list of internal directories/folders with text files. The crawler that handles the RSS and text files passes the associated metadata to a tool that creates plain text, focusing on RSS and linked web pages. The plain text is then sent to an agent who allocates a task to the next available OntosMiner for semantic information extraction. Internal databases are mapped using the External Data Source (EDS) tool. The EDS system uses an ontology that describes concepts and relations, mapping them to external databases. This mechanism allows the system to read instances from the external database. The extracted information is then passed to the knowledge base directly without using the OntosMiner. In order to enhance the information, the framework can use an API (Application Programming Interface) to access other Web sources. Examples of such sources include "dbpedia", "freebase", "geonames" and other trusted sources. Most of them follow the guidelines of the W3C and provide information in RDF/XML and similar formats.

4.2.2 The semantic layer

The semantic layer core function is the OntosMiner engine. The majority of information on the Web is in unstructured format. The OntosMiner engine processes plain text and extracts the entities and facts based on the defined ontology (see section 3.2). Before the extracted metadata can be imported into the knowledge base, the semantic layer process needs to merge same entities and clean-up the semantic graph. By using this developing set of guiding parameters, the merger also resolves the problem of unique identifiers. For example, if the OntosMiner has extracted from the concept "company" the instance value "IBM" then the next step is to determine if such a value already exists. It is possible that in the knowledge base we have "International Business Machines" and the engine needs to merge "IBM" with the existing instance. This is how the semantic graph grows and with it our knowledge about a specific object.

4.2.3 The knowledge base layer

The knowledge base is a database which stores the extracted semantic information. Following the rules and standardization of the W3C, the metadata are stored as triples. These triples (subject predicate object) will be stored according to the Resource Description Framework (RDF). RDF storage benefits the user through scalability and performance.

4.2.4 The application layer

The application layer is mainly connected to the knowledge base through SPARQL access. The type of application depends on the user scenario but in most cases standard functions like search or query are included. The architecture also allows for applications to interface directly with semantic layer components when necessary.

4.2.5 The presentation layer

Client devices interact with the presentation layer of the architecture. The device's independent components show results in a suitable format according to the capabilities of the device. In the Ontos Framework, the presentation layer is based on web widget and JavaScript technology which has the flexibility integrating with different devices. In the IR solution, the user interface consists of a set of web widgets that are integrated with existing web pages. Microsoft Outlook is used as an additional desktop device.

4.3 User scenarios

The current IR solution consists of two distinct scenarios. The first solution is an intelligent way of collecting media information with the ability to search and filter according to desired topic. The underlying semantic technology allows the IR user to receive only relevant articles. This solution replaces the classical clipping service where external agencies read printed news and collect articles according to a set of key words.

4.3.1 News clipping for IR

The primary objective of this service is to collect all news relevant to IR. In order to facilitate this automated clipping, the IR team decides which trusted sources should be crawled. The sources could be RSS, URL or Blogs. A list of words is generated to define a given concept so

irrelevant articles are filtered out. The system provides further possibilities for search and navigation within the user interface.
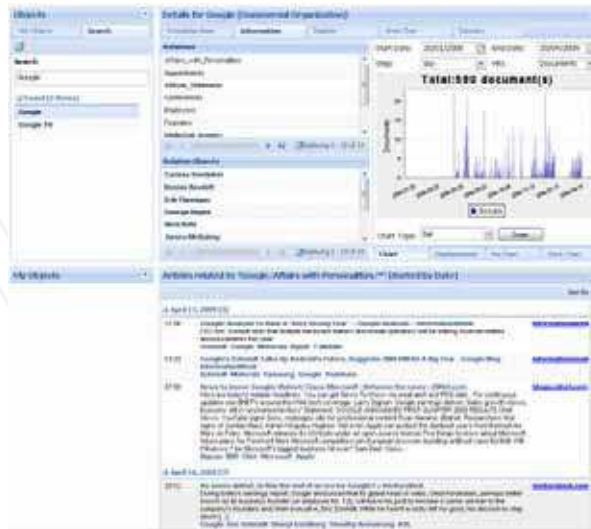

Fig. 10. News Clipping and Browsing

The described solution is a first step towards providing the IR team with more relevant information. The next level is to provide additional analytical functions to assist the team in the decision building process as well as supporting them in fact analysis.

4.3.2 Analytics for IR

Fig. 11. Analytics

While using the clipping service, the user still makes their own conclusions about the content while browsing articles. However, analytical functions provide an additional option for examining the collected information. This analysis could involve the monitoring of media coverage, brand measurement, the analysis of individual sentiments or identifying management change at a specific company. The objective is for the system to create different graphs about situations of interest from the metadata. The competitive intelligence analysis is presented in a dashboard where each user can configure their own needs (see Fig. 11).

The web provides large amounts of data and it is impossible for a human to aggregate and analyze this kind of volume. The analytical functions described above provide a great support in creating a real-time situation snapshot and comparison between companies and their competitive environment. They also represent an improved way of working with multilingual texts – the ontology simplifies analysis by connecting the same concepts from various languages into one graph.

4.4 Lessons learned and future directions

Lessons have been learned by deploying the first IR solution. There is a strong need to analyze financial data, ideally extracted from annual reports. Once a competitor is identified, it is probably possible to receive their annual report and process the content. This scenario would involve the reading of a PDF document, transforming it into XBRL (eXtensible Business Reporting Language) and extracting the financial data of interest.

These days there is a lot of audio and video information on the Internet. A next point of interest is the creation of tools allowing the transformation of speech to text in order to use the NLP system to analyze the content and create semantic metadata for the knowledge base. In order to provide more flexibility, the NLP engine will be expanded to allow users to add their own personal concepts and instances. This process will contribute to the flexibility of extended domain ontology creation and facilitate keeping pace with a changing environment.

5. Conclusion

Semantic web technologies can be effectively employed in the area of investor relations and knowledge management. It is not enough to have tools for business intelligence that focus only on structured data. Increasing economic pressure, the growing weight of unstructured content and the difficulty of accessing relevant information underscore the need for a broader conception of business intelligence. The success of companies is correlated to the speed and and depth of their analysis of the changing environment. This involves the scanning of all trusted sources for a better understanding of the competitive landscape, serving not only to attract and retain investors but also for the growth and competitive advantage of the company. The same technology can benefit other departments like the sales and marketing division, management and the research team. Semantic web technology is concerned with the acquisition of information and its transformation into metadata that can be processed by computers in order to create sustainable advantage. Ontology-driven information extraction improves precision and recall while generating knowledge for the IR department. The World Wide Web provides a huge amount of data that needs to be processed effectively if companies want to benefit. Only companies that manage this process will create sustainable advantage over their competitors.

6. References

Berners-Lee, Tim (2001). *Weaving the Web,* Texere Publishing, ISBN 978-0752820903

Berners-Lee, T.; Chen, Y.; Chilton, L.; Connoly, D.; Dhanaraj, R.; Hollenbach, J.; Lerer, A. and Sheets, D. (2006). *Tabulator: Exploring and Analyzing linked data on the Semantic Web,* http://swui.semanticweb.org/swui06/papers/Berners-Lee/Berners-Lee.pdf, MIT, Cambridge, MA, USA

Berners-Lee, Tim (2007). *Q&A with Tim Berners-Lee,* BusinessWeek Online, April 9, 2007

Daconta, Michael C.; Obrst, Leo J.; Smith, Kevin T. (2003). *The Semantic Web,* Wiley Publishing Inc, ISBN 0-471-43257-1, Indianapolis, Indiana

Davenports, Tom (2008). *Enterprise 2.0: The New, New Knowledge Management,* Harvard Business Online, Feb. 19, 2008

Davis, John; Studer, Rudi; Warren, Paul (2006). *Semantic Web Technologies,* John Wiley & Sons Ltd, ISBN 978-0-470-02596-3, England

Drucker, Peter F. (1993). *Post-Capitalist Society,* HarperCollins Publishers Inc., ISBN 0-88730-620-9, New York, USA

Efimenko, Irina; Hladky, Daniel; Khoroshevsky V.F.; Klintsov, V. (2008). Semantic Technologies and Information Integration. *Proceedings of 2nd European Semantic Technology Conference (ESTC2008)*, Vienna, 2008

Fensel, Dieter, (2004). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce,* Springer, ISBN 3-540-00302-9, Germany

Fensel, Dieter; Hendler, James; Liebermann, Henry and Wahlster, Wolfgang (2005). *Spinning The Semantic Web,* First MIT Press, ISBN 0-262-06232-1, Cambridge, Massachusetts

Gruber, T. (1993). A translation approach to portable ontologies. Knowledge Acquisition 5(2):199-220. http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html

Heyer, Gerhard; Quasthoff, Uwe; Wittig, Thomas (2008). *Text Mining,* W3L-Verlag, ISBN 978-3-937137-30-8, Germany

Hladky, Daniel; Khoroshevsky V.F. (2007). Semantic Technologies for Real Commercial Applications. *Proceedings of Semantic Technology Conference 2007*, San Jose, California, 2007

Hladky, Daniel (2008). Integration of Semantic Resources for Business Intelligence. *Proceedings of Semantic Technology Conference 2008*, San Jose, California, 2008

Khoroshevsky V.F. (2003). Shallow Ontology-Driven Information Extraction from Russian Texts with GATE. *Proceedings of IESL-2003*, Borovec, Bulgaria, 2003

Khoroshevsky V.F. (2005). Semantic Indexing: Cognitive Maps Based Approach. *Proceedings of RANLP-2005*, Borovec, Bulgaria, 2005

Khoroshevsky V.F. (2008). Knowledge Spaces in Internet and Semantic Web (Part 1). *Artifical Intelligence and Decision Making*, N 1, 2008

Kotler, Philip; Pfoertsch Waldemar, (2006). *B2B Brand Management,* Springer, ISBN 978-3-540-25360-0, New York

Porter, Alan; Cunningham, Scott (2005). *Tech mining: Exploiting new technologies for competitive advantage,* Wiley and Sons, ISBN 0-471-47567-X

Porter, Michael E. (1998). *Competitive Strategy,* The Free Press, ISBN 0-684-84148-7, New York

Staab, S.; Angele, J.; (2000). *Semantic Community Web Portals,* Computer Networks 33 (1-6), pp 473-491

Semantic Web

Edited by Gang Wu

Having lived with the World Wide Web for twenty years, surfing the Web becomes a way of our life that cannot be separated. From latest news, photo sharing, social activities, to research collaborations and even commercial activities and government affairs, almost all kinds of information are available and processible via the Web. While people are appreciating the great invention, the father of the Web, Sir Tim Berners-Lee, has started the plan for the next generation of the Web, the Semantic Web.   Unlike the Web that was originally designed for reading, the Semantic Web aims at a more intelligent Web severing machines as well as people. The idea behind it is simple: machines can automatically process or "understand" the information, if explicit meanings are given to it. In this way, it facilitates sharing and reuse of data across applications, enterprises, and communities. According to the organisation of the book, the intended readers may come from two groups, i.e. those whose interests include Semantic Web and want to catch on the state-of-the-art research progress in this field; and those who urgently need or just intend to seek help from the Semantic Web. In this sense, readers are not limited to the computer science. Everyone is welcome to find their possible intersection of the Semantic Web.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Daniel Hladky (2010). Sustainable Advantage for the Investor Relations Team through Semantic Content, Semantic Web, Gang Wu (Ed.), ISBN: 978-953-7619-54-1, InTech, Available from:
http://www.intechopen.com/books/semantic-web/sustainable-advantage-for-the-investor-relations-team-through-semantic-content

INTECH

open science | open minds