UNIVERSITY OF GRANADA

University School of Computer
and Telecommunications Engineering

Department of Computer Science
and Artificial Intelligence

SUPERVISED RESEARCH PROJECT

SEMANTIC WEB MEETS COMPETITIVE INTELLIGENCE

Juan Antonio Bernabé Moreno

SUPERVISOR: DR. ENRIQUE HERRERA VIEDMA

Munich, December 2009

Semantic Web meets Competitive Intelligence

Abstract— In the new economy set up, where the globalization and the shift from the offline to the online world are consolidated trends, competitive intelligence is no longer a choice but a must for an enterprise to remain competitive and survive. The amount of information about the competition and the industrial environment is increasing to get to quality and volumes with no precedents. The technology supporting the practice of competitive intelligence has not succeeded in keeping the same transformational pace and cannot manage the uprising complexity, which results in a significant amount of relevant information remaining unexploited. A new approach is required to address this complexity and to bring the full potential the competitive intelligence. The semantic web is based on the premise that machines will be able to understand the content and information in the web documents in a similar way as human beings would make sense of. In this work we examine how the semantic web techniques can be leveraged to bring the competitive intelligence to the next level. After an introduction and a deep-dive into the competitive intelligence world, we take a detailed look at its enabling and supporting technologies. After that, we introduce the semantic web, its core application scenarios and briefly explain how to bridge the gap from today's web to a semantic enabled web. At the end, we present a multi-dimensional model to explain the benefits and transformation from the semantic web in the competitive intelligence context and we summarize our conclusions and outlook

***Index Terms*— Competitive Intelligence, Semantic Web, Semantic Integration, Semantic Information Retrieval**

INDEX OF FIGURES

INDEX OF TABLES

- The top 10 in-demand jobs in 2010 did not exist in 2004. We are currently preparing students for jobs that don't exist yet.... using technologies that haven't been invented yet... in order to solve problems we don't even know are problems yet.
- 1 out of 8 couples married in the US last year met online
- There are over 200 million registered users on MySpace. If MS were a country, it would be the 5th largest in the world (between Indonesia and Brazil)
- There are 31 billion searches on Google every month. In 2006, this number was 2.7 billion. To whom were these questions addressed B.G (before Google)?
- The first commercial text message was sent in December of 1992. Today, the number of text messages sent and received everyday exceeds the total population of the planet
- For the Radio, it took 38 years to reach a market audience of 50 million. For the TV, 13 years. For the Internet, 4 years. For the iPod, only 3 years. For Facebook, it took only 2 years to reach 50 million registered users
- The number of internet devices in 1984 was 1,000. In 1992 it was 1,000,000… In 2008 it was 1,000,000,000
- There are about 540,000 words in the English language, about 5 times as many as during Shakespeare's time
- It is estimated that a week's work of the New York Times contains more information than a person was likely to come across in a lifetime in the 18th century
- It is estimated that 4 exabytes (4 * 10^19) of unique information will be generated this year. That is more than the previous 5,000 years
- The amount of technical information is doubling every 2 years. For students starting a 4 years technical degree this means that half of what they learn in the first year of study will be outdated by their third year of study
- NTT Japan has successfully tested a fiber optic cable... that pushes 14 trillion bits per second down a single strand of fiber. That is 2,660 CDs or 210 million phone calls every second. It is currently tripling every six months and is expected to do so for the next 20 years
- By 2013 a supercomputer will be built that exceeds the computational capabilities of the human brain
- Predictions are that by 2049, a $1000 computer will exceed the computational capabilities of the entire human species

Table 1 The era of exponential changes in facts

ACKNOWLEDGMENTS

This work has been written in quite special conditions and therefore I would like to express my sincere appreciation to those who have made it possible. I observed the same 2 patterns in countless situations where I have asked for help: those that are really willing to help, those that do whatever it takes to help, even finding and proposing solutions that you haven't even thought about, even if the *"Sorry, but I can't do anything"* answer would have been the easiest and quickest way out… and those who just don't want to be bothered…

Fortunately, I came to work with plenty of innate helpers and I hope that they find my current work as a pay-off. To mention some of them, Dr. Enrique Herrera Viedma, the one who patiently coordinated my work and Dr. Francisco Herrera, who took out of my way superfluous things to allow me to focus on the real work

Out of all the people who made it possible, I would like to highlight Kathrin and my family, and especially my father, who has been always supporting me and solving for me the on-site issues.

To them all, a big thanks.

FOREWORD

Competitive Intelligence (CI) is basically a process involving a set of information-related activities. This process relies on the available technologies to access and process all the required information and after applying knowledge, derive intelligence to support a decision making process, assist one company's strategy, etc.

The semantic web promises the opening of new ways to access the information as well as the added capacity of reasoning and thereby knowledge inference.

In this work, we will be exploring the new possibilities that the semantic web opens for the Competitive Intelligence

I. VISION STATEMENT

The vision represents the place where one wants to be in the long term, the desired future state. The Competitive Intelligence world is a reality, a more and more usual business practice taking place on the market, an activity involving many people and more and more present in the business routine… Semantic Web is on the other hand still in its infancy, growing at a constant pace and invading slowly business scenarios… that's why it is appropriate formulating a vision statement:

With the progressive introduction of the semantic web the competitive intelligence will be brought to a next level

II. OBJECTIVES

Before starting the reading, we should get clear on the goals we are setting, to assess in advance if the work is a reading worth it, or rather a waste of time. This section will set clear expectations on the upcoming pages.

The first aim of this work is providing a clear understanding on the confusing Competitive Intelligence term and the *must-know* topics around it.

In order to assess the transformation that the Semantic Web brings to the CI world, it is necessary to understand the current stand of the CI applications.

Quantifying the improvement/impact of introducing semantics into the CI world is maybe too ambitious, but we will pave the way to that by establishing a framework to assess CI systems and set the basis for the definition of metrics for the evaluation criteria.

When we talk about Semantic Web the question *"by when"* usually arises. It points to review the semantic web state-of-the-art. This topic can be discussed in several books, but we will follow a very schematic, yet complete approach.

Another very obvious question that frequently arises relates to the "internet status quo" and "internet-to-be" from the semantic perspective… We will explain how to bridge this gap.

We will also consider the potential semantic web improvements from every single dimension of interest and will present potential adoption suggestions from the visionary yet pragmatic perspective.

III. ORGANIZATION OF THIS WORK

This work is divided into four parts:

The *first part* introduces the topic competitive intelligence in a very systematic way, as well as the competitive intelligence cycle and the compounding processes

The *second part* recaps on the techniques that have been applied so far to the different processes involved in the competitive intelligence cycle

The *third part* takes up the semantic webs

The *fourth part* focuses on the impact and possibilities the semantic web shall bring to the competitive intelligence world

PART I: DEEP-DIVE INTO THE COMPETITIVE INTELLIGENCE WORLD

Nowadays, no company can take the risk of not being aware of the changes in the business environment, because the globalization widens it enormously and the rapid changes in technology foster tremendously the competitiveness. The information society supported by the new allows for a dislocation of the competitors… A little company in China might push a western giant out of the market, and Indian engineer can be selected for a job in London over a local engineer in UK, etc.
It makes CI increasingly attractive for companies aiming at remaining competitive and at the same time introduces the need to keep track on the last technology advances that open the door to new ways of handling information, as the semantic web

IV. INTRODUCTION

A. A historical note to Competitive Intelligence

What enables the wise sovereign and the good general to strike and conquer,
and achieve things beyond the reach of ordinary men, is foreknowledge
–General Sun Tzu

Since the man is man and the man combats others to survive, the intelligence has been playing a prominent role. If we look back to the past, we found the Sun Tzu's military masterpiece *The Art of War*, that can be considered as the origin of the CI usage. [6] In this book the following statement is made: "One will not be in danger in hundred battles if one knows his enemy and himself" Nowadays, the statement is still valid, and could be formulated as follows: competition unawareness is a big competitive disadvantage.

Competitive Intelligence, either with military or economic/industrial background, has always been part of the society. In Chine there are some written evidences in the military field which testifies, that it has existed for at least 5000 years[29].

1) Before 1500 A.D.

We have encountered countless ancient text about the role of military intelligence and espionage in the war. Those who possessed highly-developed military intelligence techniques were able to succeed in commerce and economics by applying them Especially in Asia these intelligence techniques have been applied far before the Christian era. For example there is constancy that Chandragupta Maurya, founder of the Maurya Empire in India, made use of assassinations, spies and secret agents, as described in an ancient Indian treatise on statecraft, economic policy and military strategy named Arthasastra.

The ancient Egyptians had a thoroughly developed system for the acquisition of intelligence. Egyptian hieroglyphs reveal the presence of court spies, as do papyri describing ancient Egypt's extensive military and slave trade operations. Early Egyptian pharos employed agents of espionage to ferret-out disloyal subject and to locate tribes that could be conquered and enslaved. From 1,000 B.C. onwards, Egyptian espionage operations focused on foreign intelligence about the political and military strength of rivals Greece and Rome. Hebrews used spies as well, as in the story of Rahab.

Between 1500 B.C. and 1200 B.C., Greece's many wars with its regional rivals led to the development of new military and intelligence strategies. The early Greeks relied on deception as a primary means of achieving surprise attacks on their enemies. In the era of democratic Greek city-states, espionage was chiefly employed as a political tool. Agents of espionage spied on rival city-states, providing rulers with information on military strength and defenses. The most farsighted contribution of the ancient Greek intelligence community, however, was its creation of a complex and efficient means of communication between cities. Couriers delivered messages between cities, but important messages were also relayed between a series of outposts or towers using semaphore, a form of communication that utilized signals to convey messages.

Unlike most of the Roman writers claim, Rome did not defeat its enemies by trickery or deceit but by superior force of arms, and for the most part they were right. The Roman legions could outstrip almost any opponent in manoeuvrability and discipline.

Rome certainly did not lack enemies to target. Neighbouring clans like the Aequi and Volsci, and later the Etruscans, Samnites, and Gauls, kept the Romans constantly at war during the early and middle republics. Collecting intelligence about these surrounding tribes and discerning whether they would be friendly or hostile in a given situation was probably a full-time job, and instances of such intelligence gathering appear in Livy's stories.

The Romans were heavily involved in espionage, but it cannot be said that they ever established a formal intelligence service. The closest they came was in using groups like the *frumentarii* and the *agentes in rebus* for various internal security tasks. Protecting the emperor as well as keeping him on the throne became so crucial after the third century that most of Rome's intelligence activities were focused inward. Ironically, for all their reputation as empire builders, the Romans were never as good at watching their enemies as they were at watching each other.

In the Middle Age, the birth of large nation-states, such as France and England, in the ninth and tenth centuries drove the need for intelligence in a diplomatic setting. Systems of couriers, translators, and royal messengers carried diplomatic messages between monarchs or feudal lords.

The Crusades also changed the tenor of espionage and intelligence work within Europe itself. Religious fervor, and the desire for political consolidation, prompted thirteenth century church councils to establish laws regarding the prosecution of heretics and anti-clerical political leaders. The ensuing movement became known as the Inquisition. Although the Church used its political force as impetus for the Inquisition, enforcement of religious edicts and prosecution of violators fell to local clergy and secular authorities. For this reason, the Inquisition took many forms throughout Europe. The same movement that was terror-filled and brutal in Spain, had little impact in England and Scandinavia.

Espionage was an essential component of the Inquisition. The Church relied on vast networks of informants to find and denounce suspected heretics and political dissidents. By the early fourteenth century, Rome and the Spanish monarchs both employed sizable secret police forces to carry out mass trials and public executions. In southern France, heretical groups relied on intelligence gathered from their own resistance networks to gauge the surrounding political climate, and assist in hiding refugees [32] (a kind of early Résistance)

2) The Renaissance

Once the Church started loosing its complete domination of the world, there was a transition within Europe to a more localized, yet nationalistic model of government. With the wealth increase of each nation, the competitive activity gained more power. The

over trade dominance and exploration of the New World was the first target of the competitive intelligence and espionage, and forced regimes to adopt sophisticated measures to protect their political, economical and military interests.

Niccolo Machiavelli inspired by Aristotle and Cicero, recommended in his "The Art of War", that rulers routinely make use of espionage tradecraft, engaging in deception and spying to insure protection of their power and interests.

In the context of the emerging schism, Henry VIII created a secret police to locate infiltrate Catholic cells that might put at risk the English monarchy. This tradition was followed by her daughter Elizabeth I, who could save so her throne in many occasions discovering several conspiracy plots by using the intelligence services.

The main contribution of the Elizabethan espionage system relies on the employment of highly educated well-trained scholars, engineers and scientists to seek and analyze intelligence information, instead of haphazard, ill-trained, unspecialized volunteers. Thus, the intelligence became a highly-skilled job

Along with the institutional revolutions, the renaissance gave birth to countless technology artifacts and developments that altered the practice of espionage: new telescopes and magnifying glasses, rebirth of complex mathematics to enable the encryption of information, invisible inks, transport facilities that enabled better communication and therefore increased information exchange, etc.

3) The Age of Empires, Industrial Revolution, and the Nineteenth Century

"To be beaten is excusable; to be taken by surprise is unforgivable"

Napoleon Bonaparte

During the Age of Empires the intelligence saw its greatest development so far. Europe was scenario of numerous conflicts that quickly expanded to the colonial worlds. The industrial revolution, the territorial expansion and commercial growth, the emerging new political philosophies and regimes, the immigration drove an unprecedented transformation of the intelligence communities

The French Revolution in the late 1790s is the most prominent example of a social and political change, where all fractions relied to certain extent on the espionage. The Jacobian Club leaded by Robespierre developed a network of informants in charge of denouncing all potential republic enemies. Beside this network, there was a new set of treason laws to ensure the execution of aristocrats and clergy that might be potential traitors. It has unquestionably been one of the greatest abuses of the intelligence powers in the modern era.

The American Revolution (1776-1783) together with the independence wars in South America in the first third of the 19th Century led to the dismantling of the Europe's Colonial Empires. Europe turned its attention to Africa and Asia. The competition for the Europe hegemony meant the creation of a very complex alliance system. The use of the intelligence was in this constellation crucial for the colonial rulers to detect separatist movements and other rebellions.

The industrial revolutions of 1848 that can be seen as the Imperialism motor motivated the birth of the modern industrial espionage. England, France and Prussia employed infiltrated spies into political and labor organizations to detect any antigovernment activities. At the same time, labor organizations often spied each other to report on working conditions, factory operations, mining productivity, etc. Sporadically, this intelligence was used to carry out sabotage acts, destroying factories, mines and government property. Once the conflicts situation stabilized, the governments continued employing the same industrial espionage practices increasingly against foreign economic competitors.

In 1837 the daguerreotype was invented and with it the first practical form of photography. It changed forever the collection of intelligence information, as the agents could for the first time portray targets, documents, etc as they actually were. Cameras were made smaller, disguised and placed in mundane items for espionage purposes. Until the invention of the electronic data storage, in the 20th century, the photography was irrefutably the best medium for storing and copying information.

Another very important date in the intelligence history is the May 24 1844, when Samuel Morse sent the first message via telegraph. Within minutes it was possible to send a message over lines requiring only operational code knowledge. Also government relevant information was going through these lines and it was matter of months till the rival intelligence services learned to tap the line to get access to the communication contents. To overcome this situation, governments created specialized departments that coded/decoded the transmitting messages using highly complex cryptographic algorithms. A new intelligence era was born in which the competing nations in Europe and the Unites States systematically kept under surveillance the wired and unwired communications.

The advances in transportation (i.e.: the invention of the locomotive and the construction of railroads) permitted agents to travel to foreign destinations disguised as a tourist. It explained also the immigration increase and led to the creation of special intelligence department with foreign cultural and idiomatic background.

By the down of the 20th century the espionage developed as a highly specialized technical field, involving more research and analysis than field operations and leaving behind political intrigues and the battlefield.

4) Modern Competitive Intelligence

From the 1960s onwards, the Competitive Intelligence emerges as industrial practice separating from military purposes. Prescott [35] introduced an evolution model based on the predominant intelligence activity that we have expanded:

a) Competitive Data Gathering (1960-1980)

In the United States and other developed nations, the 1960s experienced a period of prosperity never before experienced. Equally, in some developing nations such Brazil, Mexico and South Korea, manufacturing and international trade had profound positive effects on gross national products. Funding of applied research by manufacturers was breaking records as well, with the resulting increase in the amount of technical information available

In 1971 a group of Swedish banks cooperated to form their own CI research organization to provide information about their competitors in foreign countries.

The German pharmaceutical corporation Bayer, that has been systematically analyzing the patents of its competitors as early as 1886, increased the competitive data collection activities.

In 1973 the economic downturn following the worldwide had a dampening effect on the industry which resulted into an increased competition not only in sales, but also in costs of production. It led some individuals to industrial espionage as a solution to the problems they were facing [33].

The modern CI activities emerging between the 60s and the 80s were primarily focused on tactical, but informal, data gathering, although market research with an orientation towards customers was well established. It was developed on an ad hoc basis with limited to inexistent analysis and has a minimal impact on business decision making.

To put it simple, it was all about finding the information, simple role yet of great importance. Several firms established as service providers for information cataloging, training and information brokering, such as Washington Researchers, Fuld and Company or Find/SVP.

The ground principle that governed this époque stated that the information is only as good as the data you can gather. The publishing of Porter's book [34] took the CI to the next level.

b) Industry and Competitor Analysis (1980-1990)

During the 80's, HITACHI and Mitsubishi deployed great means to procure itself the descriptive brochures of new computer chips elaborated by IBM, but ended up getting caught by the FBI and forced to pay more than 800 million dollars

Mitsubishi again and other Japanese giants got involved in the early 90s in a scandal when an employee of Applications International Corp., a California based high-tech company, stole computer program codes to sell them to the Japanese.

The American Society of Industrial Security conducted a survey in 1991 and 37 percent of the 165 U.S. firms responding said that they had been targets of espionage.

After the maturity of the competitive data collection, the 80's shifted the focus from the gathering to the analytical processing of information. The results of these analysis's were more and more considered in the decision making process and corporate planning. As a consequence, many corporations started incorporating CI as part of their organization, employing dedicated resources with the required skills set and defining processes and interfaces with the decision making management board, who was provided with rather quantitative summaries (the qualitative era was yet to come)

The *raison d'être* of the analysis was transforming the data into intelligence, which brought the planners to a more prominent role, as they have traditionally been interested in the relationships business-environment. For the first time, the CI activity experienced a specialization or division of labor: those who gathered the data, and those who by means of analysis transformed it into intelligence.

As Competitive Intelligence derived directly from espionage, practitioners had to deal with this "spy" image and there was no clear ethic code to follow. Still nowadays, many firms talk about their CI processes with certain reluctance and keeping it intentionally obscure, which can be seen as a major adoption barrier.

c) Competitive Intelligence for Strategic decision making (1990-)

CI promoted to a privileged role within the corporate strategic decision making process, but there wasn't a consensus on how/whether it shall influence the bottom line. A new uprising technique called *benchmarking* enabled CI analyst to address the bottom-line issues showing much more tangible outcomes that other contemporary techniques way more abstract.

A controversy started arising around the participation of the governments in the different countries in business intelligence activities and its impact on the competitiveness[1].

1 In FRANCE, cooperation between industrials and the secret service goes back to the 50's. Information from the SDECE (today the DGSE) played an important role in the growth of certain national industries by creating, approximately 10 years back, the "Y" service, in charge of economic and financial information. It is capable of furnishing confidential documents on multinational firms' projects as well as their research and marketing plans.
As far as the DST goes, it intervenes regularly to prevent leaks in the large government infrastructures.

In 1992, the French intelligence budget was increased nine percent to enable the hiring of 1000 new employees. Former French Spy Master publicly disclosed in 1991 that the French intelligence services had conducted extensive international spying to keep abreast of changes in the commercial and technological industries to support French commerce. He also confirmed that the French built a computer with stolen American technology and, as a joke, nicknamed it with the initials of the French Secret Service. The French intelligence services are not the only government-sponsored operations spying on U.S. businesses. U.S. intelligence and law enforcement groups have identified Israeli, British, Canadian, Japanese, South Korean, Argentine, Egyptian, Chinese, German, and Swedish Intelligence groups in recent years.

One of the branches of the KGB, dispersed in 1991, dedicates itself exclusively to industrial information and espionage, while a large number of its former agents selling their services and know-how to the private sector, in order to infiltrate companies and get closely guarded strategic information out.

At the same time, the information systems were in the target of debate. Although they were more and more sophisticated, the amount of information available increased unprecedently, so the diversity of sources.

Intensifying that, the transformation into a global economy extended the scope of the CI activity to the international playground. Competitors were no longer local, but global… with cultural and idiomatic issues.

Altogether led to the consideration of CI as cornerstone within the strategic decision making process.

Documented cases of U.S. industrial espionage by foreign competition have experienced a significant increase in the 90's. In a well publicized case between GM and Volkswagen, German prosecutors linked an ex-GM executive, Inaki Lopez, who had joined Volkswagen, to a cache of secret GM documents.

In 1992, Recon/Optical, a suburban Chicago military contractor, charged the Israeli Air Force with trying to steal the blueprints for a top secret airborne spy camera - the Israelis agreed to pay $3 million in damages.

The Federal Bureau of Investigation and industry experts have estimated the U.S. trade secret theft in 1992 cost U.S. companies more than $100 billion in lost revenues. If left unchecked, analysts estimate the losses could grow an additional 50% by the year 2003. Documented cases of industrial espionage are further supported by the numerous methods used by foreign companies, which include, bur are not limited to, theft of high-tech equipment from loading docks, warehouses and assembly lines, wire taps and microphones, fax interceptors, recruitment of competitor's employees, planting of moles, and hiring "hackers" to break into telephone and voice mail systems, computers and corporate networks. Cellular telephone usage has increased the threat to executive confidentiality. A high-tech threat which has recently been documented include EMI intrusion or electronic eavesdropping, of computer and telephone systems by sensitive electronic monitoring equipment.

Cellular telephones have been examined to extract programmed speed call numbers, and in some cases, programmed passwords. Portable personal computers often contain programmed passwords, and computer encryption software used by large firms has recently become the targets of industrial spies. Documented cases also exist of U.S. executives travelling to foreign countries and while having dinner or attending meetings, their briefcases, luggage and personal belongings have been searched in hotel rooms.

Industrial espionage by U.S. firms has taken place domestically and internationally, but as a general rule it is discouraged by corporate executives. Generally, competitive information of a classified nature moves from one U.S. firm to another through normal employee attrition and competition. Therefore, the average U.S. firm is naive about company-financed and government-advocated espionage. Historically, U.S. firms have viewed our government as a tax collection agency, which hampers business by imposing rules and regulations that minimize profits and production. The opposite holds true for international firms which are often subsidized by government to allow increased employment, or are financed by their governments to encourage competition in the international market place, thereby increasing exports and taxes.

Business philosophies outside the U.S. are also quite different in the international market place. While the U.S.

competed against the Soviet Union in the Cold War, other governments (many of which the U.S. considers friendly) viewed entire nations, such as the U.S., as competitors in the high tech, high export, worldwide market place. Foreign competitors have developed, and enforce, stringent security policies to protect confidential information from U.S. competitors. U.S. corporations have, over the past five years, begun to take the threat of industrial espionage seriously and to evaluate the costs of compromised research and development. This is a good beginning for U.S. firms but years behind their international competitors.

Table 2 Industrial Espionage in the late 80s

V. WHAT IS CI GOOD FOR? IMPORTANCE OF THE CI

Competitive intelligence can help you, the small to medium-sized enterprise (SME), make money — and avoid wasting and losing money — by knowing more about your company, competitors and industry.
Have you ever had the experience of developing a new product or service, only to see a competitor beat you to the market?
It's that sort of thing that makes competitive intelligence a necessity.

Like the majority of SMEs, you probably have a growth strategy focused on developing new customers, gaining market share, offering new products/services, and upgrading your equipment. All of these activities can be undermined, with disastrous waste of resources, if you do not have enough information to make accurate forecasts and implement effective competitive strategies [5]

A. Why playing in disadvantage?

An appropriate analogy is to consider this advantage the CI brings as akin to having a good idea of the next move that your opponent in a chess match will make. By staying one move ahead, checkmate is one step closer. Indeed, as in chess, a good offense is the best defense in the game of business as well[9]

B. The cost of CI's ownership

The nature of information is unique in two ways: its lifecycle is unique and not easy to predict, and the consumption of information does not lead to less information. As a result, the value of the information is not ready quantifiable. Although it is possible to measure the cost of its production, measuring its value remains a challenge. The CI outcome is value-added information in the form of one actionable result, but the process is really costly and its pay-off should be calculated against the outcomes.

VI. DEFINITION OF COMPETITIVE INTELLIGENCE

A. Basic definition of CI and scope

The Society of Competitive Intelligence Professionals (SCIP) has given the following definition for competitive intelligence as 'the process of ethically collecting, analyzing and disseminating accurate, relevant, specific, timely, foresighted and actionable intelligence regarding the implications of the business environment, competitors and the organization itself' [4]

B. Discussion about the scope

According to different authors, the scope of the CI varies, for example, for Fuld & Co, CI should focus on the marketplace and the rivals. For Miller, CI is applied only to the business environment. Kahaner put the emphasis on the competitors' activities and general business trends. Prescott & Gibbons consider an industry and the capabilities and behavior of its current and potential competitors

For the authors of the Canadian Industry Report on CI [5], Competitive intelligence isn't simply about finding information. It's not about snooping on competitors. It's about analyzing information in a continuous process that is tightly linked to your strategic planning. It's your company's circulatory system for knowledge. That knowledge encompasses the following:

- Competitors
- Technology
- Legal and regulatory changes
- Suppliers
- Materials
- Industry and market trends
- Political and economic changes

C. But who are the competitors? How to identify them?

Getting clear on the set of competitors that are relevant for the Competitive Intelligence is of vital importance before getting down into the business and starting the process.

Following factors are to be taken into account:

- Targeting a too wide range of competitors might unnecessarily complicate phases of the CI cycle and increase the time-to-results[2]
- Targeting a too restricted set of competitors might leave certain changes in the environment unidentified, increasing the vulnerability of the firm
- Targeting only direct competitors (those who sells the same product or service you are selling) might leave unwatched companies producing a subsidiary product –even if one's growth rate impacts negatively other's one-
- Targeting only competitor for which the amount of available relevant information is high can neglect important ones, or produce outcomes that don't reflect the environment realty.

At the end of the day, CI focuses only in a well-defined set of direct competitors, but the globalization of many markets bound to the rapid technology development have multiplied the risks of defining the competitive environment too narrowly[7]

VII. CI SOURCES

A critical success factor to implement the CI strategy is selecting the right information sources. This statement that may look obvious is tied to one of the most complex and time-consuming processes within the CI activity.

If you add the complexity related to the fact that people hardly ever exactly know what they are searching for, people rarely get to express their information needs in an understandable and precise way, to the variety of information sources and to the "specifics" on each one and the credibility.

[2] Time-to-results can be defined as the time spent from the process launch to the presentation of the CI results

1) Classification of Sources

The literature treats the valuation step more implicitly. Most of the time, it discusses distinctions regarding sources, such as open vs. closed sources, internal vs. external sources, or primary vs. secondary sources [11]

a) Internal vs. External

"The next best thing to knowing something is knowing where to find it."

Proverb

Depending on whether a source can be found within the enterprise confines, or rather has an external precedence we distinguish internal and external sources. A good mix of both internal and external sources is indispensable. While internal sources include available documents both on paper and electronic and the individuals within on organization, external sources include materials published or available from outside the organization as well as outsides themselves.

b) Open vs. Closed

These distinctions implicitly refer to different criteria used in the valuation of sources. The distinction open vs. closed sources implicitly refers, for instance, to criteria such as "ease in collection" or "relevance." The distinction primary vs. secondary sources implicitly refers to the criterion "reliability of the data." In our view, it is possible to value sources more precisely when the valuation criteria are stated explicitly and not implicitly in the form of these distinctions.

Geert Hofstede created a method to measure the cultural differences among countries. This method is developed upon 5 indicators:
Power Distance Index (PDI) or to which extent less powerful individuals and institutions accept and expect that power is distributed unequally.
Individualism (IDV), or the degree to which individuals are integrated into groups (one extreme would be societies with individuals that are only expected to look after him/herself and his/her immediate family vs. societies in which people from birth onwards belong to extended families which continue protecting them in exchange for unquestioning loyalty.
Masculinity (MAS) vs Feminity refers to the distribution of roles between the genders which is another fundamental issue for any society to which a range of solutions are found. Women's values differ less among societies than men's values; Men's values from one country to another contain a dimension from very assertive and competitive and maximally different from women's values on the one side, to modest and caring and similar to women's values on the other.
Uncertainty avoidance Index (UAI) or society's tolerance for uncertainty and ambiguity (in other words, to what extent a culture programs its members to feel either uncomfortable or comfortable in unstructured situations). Uncertainty avoiding cultures try to minimize the possibility of such situations by strict laws and rules, safety and security measures, and on the philosophical and religious level by a belief in absolute Truth; 'there can only be one Truth and we have it'. People in uncertainty avoiding countries are also more emotional, and motivated by inner nervous energy. The opposite type, uncertainty accepting cultures, are more tolerant of opinions different from what they are used to; they try to have as few rules as possible, and on the philosophical and religious level they

are relativist and allow many currents to flow side by side. *Long-term orientation (LTO)* deals with Virtue regardless of Truth. Values associated with Long Term Orientation are thrift and perseverance; values associated with Short Term Orientation are respect for tradition, fulfilling social obligations, and protecting one's 'face'.

Table 3 Hofstede's method to measure cultural differences among countries [24]

c) Opening up to the military intelligence

Just for reference purposes, the categorization of intelligence sources followed by the US Government and Military has been included in this document. They use a much wider range of sources available that falls into five broad categories [12].

1. *Human-source intelligence (HUMINT)*: human sources outside the company, including competitor personnel, customers and suppliers as well as subject-matter experts, industry analysts, etc. Human sources are targeted for their knowledge and 'referral' to other sources

2. *Imagery intelligence (IMINT)*: ground and overhead imagery (aerial photography, electro-optical imagery, imagery radar and infrared sensors). In the military, IMINT is the only discipline that allows commanders to 'see the battlefield' in real time as operations progress.

3. *Measurement and signature intelligence (MASINT)*: technically derived intelligence data other than IMINT or signals intelligence. It is commonly regarded in the US intelligence community as technically derived intelligence which, when collected, processed and analyzed, results in intelligence that detects, tracks, identifies or describes the signatures (distinctive characteristics) of fixed or dynamic target sources (e.g. electromagnetic energy, sound, chemical and material residues).

4. *Open-source information*: information available in the public domain. This can be anything from a newspaper article or web page to non-proprietary company documents. For the most part it comprises that which companies themselves are prepared to release, as well as assessments of public domain information from industry observers. Open-source information provides the necessary starting point for any intelligence investigation and will usually point to a wide array of potential human sources, but in and of itself does not constitute intelligence.

5. *Signals intelligence (SIGINT)*: information derived from intercepted communications, radar and telemetry. SIGINT can be regarded a 'bridge' between imagery's ability to observe activity and HUMINT's ability to gauge intentions. SIGINT, in short, provides a hedge against strategic deception and is useful in support of other collection assets.

The practice of competitive intelligence at industry level is limited mainly to open-source and HUMINT, although a company's counterintelligence and security staff should be especially concerned with the threats posed by SIGINT in terms of economic and industrial espionage. Our focus here, however, is on open-source and HUMINT collection disciplines.

B. Choosing the right sources

Cast a cold eye on life, on death. Horseman, pass by!

Poet William Butler Yeats' Epitaph

Assessing quality of sources ties back to four major criteria that have been derived from the interaction activities that the information researcher performs regarding the source:

- Discovery: determining the exact location and approaching the source to prepare the retrieval
- Retrieval or extracting the data from the source
- Consuming and Exploiting or how to use the data for further processing to create intelligence

Pursuing this rationale the four core source selection criteria [23]:

1) Ease of access and easy of retrieval

Quantifying the "ease" is always challenging and often ends up entering the subjectivity and inaccuracy domain. Under these circumstances a Likert-scale [25] would do for example having 5 categories ranging between "none" to "very problematic"
The Table 4 recaps 6 barriers that can be evaluated to determine how easy a source can be accessed and retrieved.

Barrier	**Explanation**	**Retrieval**	**Access**
Language	Depending on the host and consumer language it can challenge multinational competitive intelligence activities	Y	Y
Culture	If the source and collector cultural background differs significantly, accessing and interacting with the source might be difficult. Hofstede develop a method to measure cultural distance between countries (see Table 3) that can be helpful to quantify the challenge.	Y	Y
Bureaucracy	Certain institutions might impede or difficult approaching the source	Y	Y
Character	Certain personal characteristics might make difficult the interaction with certain individuals	Y	Y
Geography	Certain resources might need to be acquired on-site (conference, trade fairs, etc)		Y
Technology	Accessing and retrieving information from some resources might required very special technological skills and specific knowledge on the source implementing technology	Y	Y
Fee	Sometimes there is a fee to be charged when accessing a source	Y	
Time	The response time is critical. By the time you get the information it might be already out of date	Y	Y
Stability	Sometimes the source is not available (server down) or the accessing mechanism stops working	Y	Y

Table 4 Source accessing and retrieving barriers [23]

2) Usefulness of content and processing ease

The criteria related to the content can be summarized in three major ones, namely completeness, reliability and timeliness [26]

a) Completeness

The data source can deliver the required data to gain insight into the data class the data source is meant to cover (i.e.: how often the source didn't deliver the required information and which/how many aspects of the expected information didn't get covered within the delivered result set)

b) Reliability (also credibility)

For example, in the particular case of the World Wide Web, information and data from all over the world can be accessed. Because so much information is available, and because that information can appear to be fairly "anonymous", it is necessary to develop skills to evaluate what you find. In the paper publications area, books, journals and other resources have already been evaluated in one or another way before you get it. When you are using the World Wide Web, none of this applies. There are no filters. Because anyone can write a Web page, documents of the widest range of quality, written by authors of the widest range of authority, are available on an even playing field. Excellent resources reside along side the most dubious. The Internet epitomizes the concept of *Caveat lector: Let the reader beware.*

There is a set of sub-criteria that helps determining how reliable a source it:

Authorship or who created the data source? when we look for information with some type of critical value, we want to know the basis of the authority with which the author speaks. When the author is a renowned one or someone that you consider trustworthy, it's easier, but when you find an author you do not recognize, you could consider following points:

- The author is mentioned in a positive fashion by another author or another person you trust as an authority;
- You found or linked to the author's Web/Internet document from another document you trust;
- The Web/Internet document you are reading gives biographical information, including the author's position, institutional affiliation and address;
- Biographical information is available by linking to another document; this enables you to judge whether the author's credentials allow him/her to speak with authority on a given topic;
- There is an address and telephone number as well as an e-mail address for the author in order to request further information on his or her work and professional background.

If none of the above mentioned points applies, there's nothing but taking the risk.

The *publishing body* gives further information about the source credibility. For printed media, you can assume that the source has gone through a review process. For online media, you can ask following questions:

- Is the name of any organization given on the document you are reading? Are there headers, footers, or a distinctive watermark that show the document to be part of an official academic or scholarly Web site? Can you contact the site Webmaster from this document?
- If not, can you link to a page where such information is listed? Can you tell that it's on the same server and in the same directory (by looking at the URL)?
- Is this organization recognized in the field in which you are studying?
- Is this organization suitable to address the topic at hand?
- Can you ascertain the relationship of the author and the publisher/server? Was the document that you are viewing prepared as part of the author's professional duties (and, by extension, within his/her area of expertise)? Or is the relationship of a casual or for-fee nature, telling you nothing about the author's credentials within an institution?

- Can you verify the identity of the server where the document resides? Internet programs such *dnslookup* and *whois* will be of help.
- Does this Web page actually reside in an individual's personal Internet account, rather than being part of an official Web site?

Objectivity: frequently the goals of the authors or sponsored are not clearly stated. Often the Web serves as a virtual "Hyde Part Corner", a soapbox. Because data is used in selective ways to form information, it generally represents a point of view. Every writer wants to prove his point, and will use the data and information that assists him in doing so. When evaluating information found on the Internet, it is important to examine *who* is providing the "information" you are viewing, and what might be their *point of view* or *bias*. The popularity of the Internet makes it the perfect venue for commercial and sociopolitical publishing. These areas in particular are open to highly "interpretative" uses of data.

Judging the objectivity of a source is one of the most complex tasks and typically it's sufficient extracting the main ideas by filtering out biased content, but this only is not always possible. There are some rules of thumb referred to the location of the source (i.e.: corporate sites tend to be over positive when talking of own products, or when you look for products you might be victim of advertising obscuring the real features of the products, etc), or the agenda of the information publisher (i.e.: political driven organizations often provide a biased background).

Referral or knowledge in the literature or scientific rigor. It can be typically checked by the presence of a bibliography in the source, related sources with proper attribution, the displaying of theories and techniques, etc

Accuracy especially when figures are provided… Verifiability of details is an important part of the evaluation process. It is important to verify that there is one explanation in the source on how the figures have been calculated or where they come from. In case the figures come from a presented methodology, there is enough information to replicate the methodology and come up to the same figures.

c) Timeliness

The data should be up-to-date. It refers to the currency of information. In printed documents, the date of publication is the first indicator of currency. For some types of information, currency is not an issue: authorship or place in the historical record is more important (e.g., essays on tradition in literature). For many other types of data, however, currency is extremely important, as is the regularity with which the data is updated. To ensure you have enough information about the currency of one document, check for the date(s) at which the information was gathered (e.g., US Census data). Beside that, if the document is updated on regular basis, make sure that every update is documented and dated (e.g.: "last updated" date is present), and there is a date of copyright available

3) Cost efficiency

"In medio virtus"

Aristotle

On the internet there are countless resources that are available for free but when it comes to competitive intelligence they might not be sufficient or the time you need to find what your search for is more costly than the fee you pay for accessing certain source.

The best resource collections contain a healthy and complementary mix of the two, since 1) fee-based content tends to be of higher quality and offer more robust search and output features and 2) there is much quality free content available online.

In general, there are 3 aspects that speak for fee-based data [28]:

- The authority and accuracy of web-based data are more often in question than those from a reputable fee-based source.
- Fee-based sources are often more easily - and quickly - searched, retrieved, and stored. And output can often be sorted or otherwise customized.
- Fee-based sources may be more up-to-date.

C. Internal vs External

Another categorization of the CI sources distinguishes between external and internal sources, depending on the origin of the information (inside or outside the company). In any case, the very first step in the information collection should be the sources identification (both internal and external).

Grzanka conducted several studies that showed that the majority of a company's information needs can be satisfied from within the organization [13]. Many other CI professional still put emphasis on external sources, but it leads us to point to an internal information audit as a sub-step in the overall information acquisition process.

In a typical information audit, the organization's existing collection of documentation –records, reports, databases and publications- are reviewed, as well as employees to determine what is already known about competitors.
Once all relevant sources have been identified, the next step consists of the acquisition itself. At this point there are two strategies to be followed. First, a targeted strategy will take care of retrieving already identified information pieces. Second, an active environment monitoring is performed to identify information pieces that might be relevant, but have not been identified yet by the targeted strategy.

The gathered information needs to be filtered to retain the relevant one and to discard unwanted information –or noise-. The filter is against the previously identified requirements, needs and topics and constitutes the forth sub-step in the acquisition process.

A mandatory last step consists of assessing the validity and value of the information (rejecting inconsistent, erroneous and redundant content). Notice that both filtering and validating are closely related to each other. Thus, filtering actions can be followed by validating procedures in an iterative approach.

Information validity check is usually performed throughout the sources to detect inconsistencies and misinformation. The failure to test and reject what others regard as an established truth can be a major source of error. Generally, internal information is harder to verify because it is often the result of a word of mouth process-.

D. Primary vs Secondary sources

When the information source material is closest to the person, information, period or idea being studied, we call it primary source. Secondary sources cite, comment on, or build upon primary sources, though the distinction is not a sharp one. Actually, the difference between these terms is relative, with sources judged primary or secondary according to specific contexts.

To assess the quality of a primary source, following questions are frequently asked:

- What is the tone?
- Who is the intended audience?
- What is the purpose of the publication?
- What assumptions does the author make?

- What are the bases of the author's conclusions?
- Does the author agree or disagree with other authors of the subject?
- Does the content agree with what you know or have learned about the issue?
- Where was the source made?

Secondary sources involve generalization, analysis, synthesis, interpretation, or evaluation of the original information

E. Online vs. Offline sources

The advent of the PC and internet era took the information gathering practice to break the “physical” barriers: information was no longer prisoner confined in a paper. In the late 1980s, the situation drastically changed with the invention of the Compact Disc – Read Only Memory CD-ROM, an ideal medium to hold an offline database, yet still presenting limitations in searching and information retrieving and in combining it with paper based literature in libraries and patent offices.

When Internet consolidated and become so popular, the existing offline database have been made available online and search engines such as Google, Alta Vista, Yahoo, etc became more and more powerful. The searching activity changed its nature from human driven to electronic driven and the time-to-information got drastically reduced. Thus, where a whole team of librarians worked for weeks to gather all relevant information about a subject, a single person with internet accessed required only a couple of hours to gather the same information. Along with information gathering, information analysis and transformation into knowledge substantially benefited from the upcoming analytical tools and increasing computing power.

1) Online Sources

There is a very good compendium of online sources offered in [28]. We have extended it to incorporate the newer online sources that have been born in the context of the web 2.0 wave.

a) Competitor Websites

That’s the first address when observing competitors. Carefully analyzing a competitor website will enable to portray a competitor by finding following kind of information: company and management descriptions, organization chart, locations and operations, product palette and service descriptions, marketing material and press releases, annual reports and security filings, management presentations and conference calls, partners, affiliates, suppliers, and key customers (often included as references), Channel and Competitor Affiliates' Websites

The competitor website research often requires analyzing competitor's suppliers, distributors, partners and clients’ corporate web sites to gain even more information missing on your competitor’s corporate website and double check particular entries. Also specialized market web sites can present information that your target company has not publicly shared.

b) Newsgroups

Usenet newsgroups are basically Internet discussion groups on a certain topic of interest posted from many users in different locations. There are more than 50,000 newsgroups, and more are added all the time. The first way students and scientists using ARPANet (early version of the Internet) started sharing their interests and hobbies was to create newsgroups. In newsgroups students and scientists placed information about their interest and as the number of newsgroups began to expand, the Internet administrators grouped all newsgroups together to form a category known as Usenet.

The usage of HTTP (hypertext transport protocol) changed the way computers transmit and receive information and meant for the newsgroups a truly step ahead towards in terms of adoption ease. Usenet.com state that there are currently over 80,000 discussion categories (known as newsgroups) available on Usenet. Usenet is, by nature, a text-based system; however, binary files such as movies, pictures, music files, and programs can also be shared among Usenet surfers, making it an excellent file exchange medium.

One word of caution: the quality of information may not be up to par so you may need to do additional research to verify.

Newsgroups are listed in a hierarchy that goes back to the early 1980's. The different types of newsgroups are shown by an extension (see Figure 1)

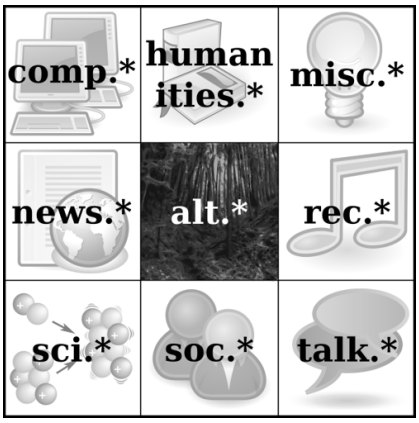


Figure 1 "Big Nine" Usenet categories

When you use a newsgroup, a forum, a blog or whatever platform where people are allow to publish their comments or make their contributions to a giving discussion subject, it's important that you keep an eye on following behaviour to judge on the relevance and quality of the information:

Trolling and trolls feeders in Internet slang, a troll is someone who posts controversial, inflammatory, irrelevant, or off-topic messages in an online community, such as an online discussion forum, chat room or blog, with the primary intent of provoking other users into an emotional response or of otherwise disrupting normal on-topic discussion. Those who response the troll's entry are "feeding the troll" and falling into the trap.

Hit and runner's comments: refers to a tactic where a poster at an internet forum enters, makes a post, only to disappear immediately after. The post often consists of a lengthy text making lots of claims that can be, but are not always, on topic. They follow the principle "throw enough in and some will stick".

Flamers: are those who write a flamebait with the intent of provoking an angry response (a flame). The motivation behind that might be reducing the popularity of the discussion group or just engaging into a violent conflict.

Leechers and lurkers: those that simply observe and benefit from the discussion group without making any kind of contribution. Bound to them exists the 1% rule that states that only 1% of the discussion group consumers actively participate in it.

Gadfly: are those who upset the status quo by posing upsetting or novel questions, or just being an irritant.

c) Press Release Wires

The majority of company related news is triggered by a company press released. Thus, the press release wires that distribute company press releases to the subscribing partners are essential for the CI activity.

A very well know press release wire is the PR Newswire that started in 1954 and provides information worldwide. There are also country-based and industry-based PR wires, like Business Wire or Voice of America, UK Wire.

d) Content aggregators

The diversity of news services motivated the creation of platforms that enable the aggregation of news from different publishers. The result is a searchable collection of research reports, etc. They are intended to make the CI searchers' lives easier.

It's important to take into consideration 3 important points when choosing an aggregator: content overlaps, content exclusivity and delay between original publisher's release and content being available within the aggregator.

e) Broadcast Media

TV, radio and broadcasted media are unquestionably a very reach source of information. Web 2.0 had brought these media into the web making this content available as streaming videos or downloadable files. Sometimes, there are transcripts available for them. Webinars and webcasts provide also valuable information about organizations and markets.

Many search engines start offering video, webcast and podcast content search like Youtube, Video Search, Blinkx or Google Video.

f) Online Advertisements

The usage of internet as advertisement location has significantly increased in last decade. Following the Ad Campaigns of your competitor will allow for gaining insight in their strategies, product palette and marketing initiatives. There are tools available to analyze the online banner campaigns done by your competition

g) Company earnings and conference calls

Many companies organize earning calls for the investors and analyst to report on financial performance. Usually they are available as streaming format or even as transcript.

h) Business and Industry Databases

They are online accessible and searchable data bases covering a given industry or business. Their content and their quality vary as well as the access price. For give some examples, just mentioning Hoovers, Thomas Register and CorpTech

i) Market and Industry Research

Professional analysts produce reports on the state of a business or a market putting emphasis on the current conditions and events. These reports are usually quite expensive.

j) Equity Research

Investment banks periodically publish analysis about the companies they invest in.

k) Government Websites, public filings and databases

As already mentioned in the previous section "A historical note to Competitive Intelligence", the different governments have pioneered the CI activity and during a long period of time, they were almost alone providing CI information to companies. When researching government sources, it is imperative to remember your civics lessons, allow for time, and consider the type of government (federal, state, local, or international). Also be prepared to apply different methods and approaches (and perhaps legwork).

When researching online material from governments or from international governmental bodies, learn how the government is structured and consider searching both English-language and other language content, when possible. Search country portals, business portals, as well as Internet directories (GoogleDirectory, Yahoo! Directory, Librarians' Index to the Internet) for leads and pointers.

l) Intellectual Property (IP) Databases

Patents, trademarks and service marks, copyright, and domain names all comprise the intellectual property of individuals and organizations. Most of the patent regulation organisms offer online access to their patent database. The European Patent Office, the US Patent and Trademark Office, and the World Intellectual Property Organization are the most important ones.

m) Industry/Trade association websites

These can be very good sources to get information about certain companies and access to subject matter experts from the competition.

n) Business and Industry portals

Portals presenting specialized content on certain businesses or industries are precious places to get good insights. There are plenty of them, differing in quality, scope, granularity and registration fees.

o) Academic and STM Content

Inspecting Scientific, Technological and Medical article allows for getting to valuable content and leading experts.

p) Employment posting and resumes

Getting access to the jobs posted by your competitor can give you an idea about its strategy (required skills that might show a technological focus, languages that might indicate expansion to other markets, etc). Employment portals provide access to their databases and even access to the resumes of registered people, but usually the terms and conditions prohibit the usage of this information for other purposes rather than hiring.

q) Web logs (blogs) and microblogs

Blogs are becoming more and more popular and valuable sources of information. A blog is a type of website, usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video. Entries are commonly displayed in reverse-chronological order.

Many blogs provide commentary or news on a particular subject; others function as more personal online diaries. A typical blog combines text, images, and links to other blogs, Web pages, and other media related to its topic. The ability for readers to

leave comments in an interactive format is an important part of many blogs. Most blogs are primarily textual, although some focus on art, photographs, videos, music, and audio (podcasting) [36].

Microblogging is another type of blogging, featuring very short posts. The content of a microblog differs from a traditional blog in that it is typically smaller in actual size and aggregate file size. A single entry could consist of a single sentence or fragment or an image or a brief, ten second video. But, still, its purpose is similar to that of a traditional blog. Users microblog about particular topics that can range from the simple, such as "what one is doing at a given moment," to the thematic, such as "sports cars," to business topics, such as particular products. Many microblogs provide short commentary on a person-to-person level, share news about a company's products and services, or provide logs of the events of one's life [37].

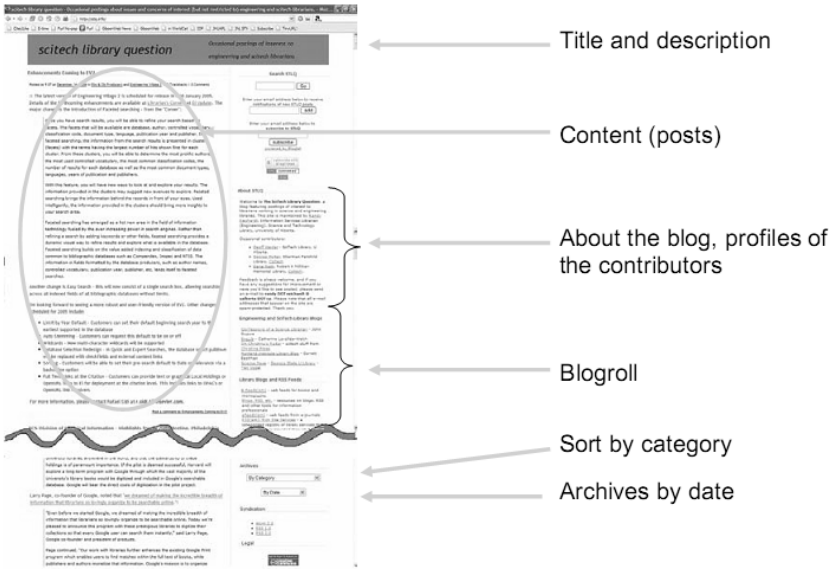


Figure 2 Anatomy of the main blog page[39]

There are a range of information services companies that monitor a select set of blogs and summarize them, or in the case of Intelliseek, their site BlogPulse.com has a set of tools that allow for enhanced conversation tracking and trend charting to compare yourself to your industry or to competitors (e.g.: comparing the last six months of mentions in blogs of your competitors and yours). Another company, Buzzmetrics, also tracks blog mentions of your company and also has a methodology for measuring the impact of word-of-mouth influences in purchase decisions.

What's most interesting about techdirt is their open door approach to having anyone submit stories for publication on the site. It's more like an opinionated news feed than a collection of press releases and shows what a blog can become. Techdirt CI provides competitive intelligence on the blog topics. The concept of Techdirt CI is to scan blogs, web pages, and any other form of publicly available electronic information and then deliver to your company a personalized blog of information on market information, competitive analysis and major news from your industry.

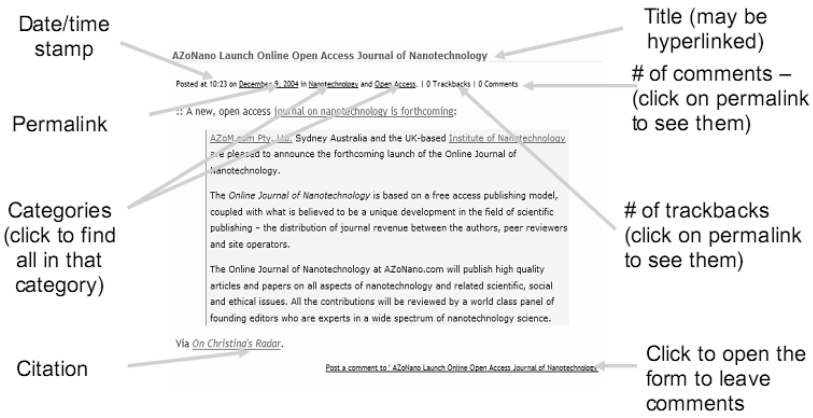


Figure 3 Anatomy of a Post[39]

r) Social Networks

Social Networks, as per the latest statistics, are tremendously growing in number of users and therefore in amount of content in form of text, photos, videos, etc). As a result, online researchers must consider online social networks as valuable sources of competitive and strategic information.

The nature of social networks is as complex as the society itself is. Thus, we found global multipurpose ones, like Facebook or MySpace, etc, or local ones (like team-ulm.de having more than 300k registered users, whereas the city of Ulm has less than 150k), segmented by age (e.g.: the Spanish Tuenti for teenies), targeted to business networking (e.g. XING or LinkedIn), etc.

The most valuable feature consists of getting the customer view of your competitors and therewith deriving their strengths and weaknesses. E.g.: you are trying to get information about the customers' perception of a market leading company –your competitor-. Just sign up in the Facebook search the name of the competitor and you will see plenty of communities –in FB language "user groups"- talking about it. Each group has a discussion board where people tell their experiences, troubles with the company products, etc… This is a big saving in money and in time compared to conducting a survey and people are much more honest in their opinions.

Another example can be performed using LinkedIn. There you have information about a company's new hires and analyzing their resumes –that are also available there-, you can derive which strategic direction they are taking (e.g.: planning the introduction of a new product, or heavily investing in a new technology, or expansion plans supported by certain idiomatic skills, etc)

In Twitter, many companies launch branding and sales campaigns that people can freely subscribe to. Herewith you can monitor any promotion or marketing messages that your competitors are sending to potential or existing buyers via tweets.

s) Social Bookmarking

Using bookmarking services such as Delicious, Mento, Diigo, Connotea, CiteUlike, etc can allow you to build up a picture of the market you are interested in. That data can then be tracked and tabulated in a spreadsheet if you need more detailed information. The bookmarks can also be clicked on, giving you easy access to information shared about the company, its products, etc. Overall, it's a simple and easy way to use web 2.0 and cloud computing tools to perform competitive intelligence gathering from the public domain behind the scenes [38]

t) Company rumors websites

Sites that are dedicated to company and product rumors must be used with discretion, but might provide insights to developing news and current issues the company might be facing. F-ckedCompany.com for example talks about big companies laying off people, or the ThinkSecret.com collates all kind of rumors around Apple Corp.

u) Discussion groups, lists and forums

There are countless industry discussion lists, which include discussions from consumers and users, industry professionals, employees, and other useful sources. They are typically hosted by professional and trade associations but also managed through Google Groups and Yahoo! Groups.

Non-Google or Yahoo! lists can be found by using Tile.net, listTool.com, or CataList (www.lsoft.com/catalist.html).

v) Tradeshow Calendars and Directories

The event calendar can help you preparing a good human source collection strategy. Your competitors might appear as exhibitors in trade fairs or key speakers in thematic conferences.

Tradeshow calendars are usually available on the web, like TSNN or Tradeshowweek

w) Search Engines: General and Specialized

Search engines are the mostly used tool for a CI researcher. They are indispensable to find information in the open web indeed. They are getting more sophisticated and more specialized (subject-areas, geographies, media and more). Search Engine Showdown and Search Engine Watch offer a collection of search engines.

x) Newsletters

Newsletters delivered electronically via email (e-Newsletters) have gained rapid acceptance and have become one of the most important online marketing enablers

2) Print and Other Sources

In spite of the increase of information available in electronic format, the role of the printed publications in the CI gathering shouldn't be underestimated. Certain trade and industry material are still published exclusively in printed media. Moreover, it is usual, that the printed version contains different content or features than the online edition of the same title.

a) Competitor Brochure

You can get a lot of information about products and services, marketing strategy, brand strategy, etc from a competitor brochure.

b) Competitor Advertisements

The field of paper-based ads outweighs the online marketing campaigns and therefore should be taken into consideration. There are plenty of companies offering news and profiles about company advertising, like Adweek, AdAge, etc.

c) Newsletters

A newsletter is a regularly distributed publication generally about one main topic that is of interest to its subscribers. Newspapers and leaflets are types of newsletters. They can provide coverage of specialized industrial topics that other sources may not cover.

d) Local and regional press

One of the best ways to get information about a competitor is just purchasing the local newspapers in the competitor's geographic location. They may also be closer to company management and staff. These publications use to be only printed or maybe offer a limited online content.

e) Special Issues of Business Publications

These publications often offer in-depth coverage of topics, industry surveys, industry outlooks, tradeshows, and company rankings. For example: Directory of Business Periodical Special Issues or specialissues.com

f) Government Records, Public Filings and Other Documents

A significant portion of material is still not available online. These include a lot of state and local documents, intellectual property records, etc.

g) Court Case Files

Litigation resources can contain information that can be very useful. There are though some obstacles, like the lack of searching procedures (e.g.: case file contents are not available for online searching) or bureaucracy asking for access, etc. Experience searcher can be of great help.

h) Grey Literature

Those documents on draft state such as technical reports, working papers, committee reports, conference papers, government documents, etc, are not widely published or distributed, and thus can offer competitive information for CI researchers. Grey literature is increasingly available on the Web, although they are not widely known and accessed.

VIII. WHO IS PLAYING CI ON MY ORGANIZATION

CI aims at providing all required information about the business environment that allow a company to remain competitive, but a company should also be aware of the revealed information that allow the competitors to remain competitive over itself.

Competitive Intelligence should be considered a bi-directional practice… Competitors are always watching you, so CI should also take care of protecting the company own knowledge assets.

The new technologies and the information society has multiplied the vulnerability of the privacy and firms aiming at remaining competitive should know the game rules and while scanning the environment seeking for other's information, putting in place the mechanisms to protect and defend the own knowledge assets.

So a merely offensive CI strategy is not enough and should also holistically integrate the defense component.

Additionally, a new dimension is getting more and more importance with the advent of the web 2.0 and the online collaborative spaces. Everybody can have an opinion that sometimes might be deliberatively bad. The bad news travels ten times faster than

good news! One dissatisfied customer will tell ten other people, while one satisfied customer may tell one or none. All it takes is for a few postings of dissatisfied customers to lose a lot of business.

Can you imagine how powerful it will be when you are right there to publicly take care of any customer service problems and respond to dissatisfied customers who criticize or complain about your product or company? Classical Competitive Intelligence mechanisms (such as early warning, etc) can be leveraged for that.

A. Protecting yourself

There's no rule of thumb to determine which knowledge assets need to be protected from competitors. Even if you consider one information item about your company as superfluous, competitors can derive potentially useful information about your organization by assembling the information puzzle, or in other words, putting it with other apparently irrelevant pieces of information together within the right context and giving it the right sense.

One a priori list of assets to be protected might include list of customers, IP items like technical designs and formulas, organizational items, marketing strategies, price structures, information about new products to be launched, location of factories, manufacturing processes, etc

There are three core keys for an effective cloaked competitor:

1. Determine the activities of greatest interest to your competitors and focus on protecting them,
2. understand the channels through which your competitors collect raw data on your firm and control what goes into them,
3. discern what techniques your competitors use to analyze the data and then deprive it of a few key pieces of data that are necessary to complete the analysis.

IX. ANALYTICAL TECHNIQUES

In this section the most common CI analytical techniques will be briefly described. These techniques aim at transforming the available information in intelligence to drive strategic decisions.

A. Competitor profiling

Superior knowledge of rivals is the key of offering superior customer value in the firm's market. In this way, this knowledge is a competitive advantage that leads one company to define the customer value relative to the offering making competitor knowledge.

Profiling the competitors facilitates this objective in three ways [9]:

- it can reveal strategic weaknesses in rivals that the firm may exploit
- the proactive stance of competitor profiling will allow the form to anticipate the strategic response of their rivals to the firm's planned strategies, the strategies of other competing firms, and changes in the environment.
- The proactive knowledge will give the firm proactive agility in two ways. Offensive strategy can be implemented more quickly in order to exploit opportunities and capitalize on strengths. Defensive strategy can be employed more deftly in order to counter the threat of rival firms from exploiting the firm's own weaknesses.

It highlights a comprehensive profiling capability as a core competence required for a successful competition.

1) Creation of a competitor profile

First question you should ask when asked for a profile is: "What decision needs to be made that will require this profile?" It the question to be answered is known on beforehand, the definition of the basic information needs can be more effective.[10]

TYPE OF DECISIONS	PROFILE INCLUDES
To help senior management prepare for a field trip	- Last news about the company (useful for chatting as an introduction) - Profile of the people who will be met - Key information about the company visited - Historic of relations with your company (speak to customer service also)
To investigate the opportunity to acquire a company	- Summary of activities and split by product and region - Financial information - List of customers - Strengths and weaknesses assessment - Estimate of company value based on previous offers and own analysis
To build awareness internally about a competitor	- Description of activities - Business system - Prices - Bidding policies
To understand opportunity for partnership	- Markets by product and region - Profile of key management and background - Historic behavior of management - Example of existing partnerships - Analysis of synergies or overlaps
To assess viability of a supplier	- Credit rating - Profile of key people and background - Historic of relation with customers - Product benchmarking

Table 5: Profile elements according to decision requirements

It is also recommended to be very specific about the level of detail the profile should be created, and it requires taking decisions before starting the information acquisition process (example: how many years back you want to go when analyzing financial information).

It is also highly recommended, that whenever considered relevant, insights about the competitor referred to your company are added (example: what does the collected information mean for your company or benchmarking with your company or with the industry average)

2) Personality profiling

This kind of profiling relies on that theory that the actions of a company will, to certain extent, depend on the personality of its top executives. Personality profiling focuses on three key aspects of a competitor's CEO

- past successes and failures –based on the theory that previous actions predict future actions

- behavioral traits of CEO –based on the theory, that people taking personal risks are prone to take business risks.
- Current business environment: the CEO is likely to act differently based on the current corporate culture

From this information, it is possible to better predicts future actions of a competitor

B. Conjoint analysis

The objective of conjoint analysis is to determine what combination of a limited number of attributes is most influential on respondent choice or decision making. A controlled set of potential products or services is shown to respondents and by analyzing how they make preferences between these products, the implicit valuation of the individual elements making up the product or service can be determined. These implicit valuations (utilities or part-worths) can be used to create market models that estimate market share, revenue and even profitability of new designs.

A product or service area is defined by a combination of p attributes. Each attribute can be broken down in a number of levels. Respondents would be shown a set of products, prototypes, mock-ups, or pictures created from a combination of levels from all or some of the constituent attributes and asked to choose from, rank or rate the products they are shown
Each example is similar enough that consumers will see them as close substitutes, but dissimilar enough that respondents can clearly determine a preference. Each example is composed of a unique combination of product features.

The gathered data may consist of individual ratings, rank orders, or preferences among alternative combinations.

The perfect product is generally unrealistic (like a car with high speed, comfort, security and low price.

A compensatory model instates the consumer to make a “Trade off” between attributes by putting into balance advantages and inconveniences.

The ultimate goal of the conjoint analysis is decomposing preferences according to an additive utility model, specific to each interviewee.

As the number of combinations of attributes and levels increases the number of potential profiles increases exponentially. Consequently, fractional factorial design is commonly used to reduce the number of profiles that have to be evaluated, while ensuring enough data is available for statistical analysis, resulting in a carefully controlled set of "profiles" for the respondent to consider.

1) Information collection

Data for conjoint analysis is most commonly gathered through a market research survey, although conjoint analysis can also be applied to a carefully designed configurator or data from an appropriately design test market experiment. Market research rules of thumb apply with regard to statistical sample size and accuracy when designing conjoint analysis interviews.

The length of the research questionnaire depends on the number of attributes to be assessed and the method of conjoint analysis in use. A typical Adaptive Conjoint questionnaire with 20-25 attributes may take more than 30 minutes to complete. Choice based conjoint, by using a smaller profile set distributed across the sample as a whole may be completed in less than 15 minutes. Choice exercises may be displayed as a store front type layout or in some other simulated shopping environment.

2) Analysis

Any number of algorithms may be used to estimate utility functions. These utility functions indicate the perceived value of the feature and how sensitive consumer perceptions and preferences are to changes in product features. The actual mode of analysis will depend on the design of the task and profiles for respondents. For full profile tasks, linear regression may be appropriate, for choice based tasks, maximum likelihood estimation, usually with logistic regression are typically used. The original methods were

monotonic analysis of variance or linear programming techniques, but these are largely obsolete in contemporary marketing research practice.

In addition, hierarchical Bayesian procedures that operate on choice data may be used to estimate individual level utilities from more limited choice-based designs.

3) Advantages

Conjoint analysis estimates the psychological tradeoffs that consumers make when evaluating several attributes together and measures preferences at the individual level.
Moreover, it uncovers real or hidden drivers which may not be apparent to the respondent themselves, reflecting at the same time the realistic choice in the shopping task.

4) Disadvantages

Even if the design of conjoint studies can be complex (especially if too many options or respondents are present) which points to seeking for simplification strategies. It is also difficult to use for product positioning research because there is no procedure for converting perceptions about actual features to perceptions about a reduced set of underlying features

- respondents are unable to articulate attitudes toward new categories
- poorly designed studies may over-value emotional/preference variables and undervalue concrete variables
- does not take into account the number items per purchase so it can give a poor reading of market share

C. SWOT analysis

1) Description and steps

This tool is intended to evaluate the strategic position of a company by making out its strengths, weaknesses, opportunities and threats.

After the exercise, key areas which might need further research are identified. Subsequent the firm tries to formulate its strategy by eliminating weakness and concentrating and extending its strengths. Alternatively threats can be tried to be turned into opportunities, e.g. a new entrant into the market can be turned into an opportunity by forming strategic alliances and make good use of the strengths of both partners.

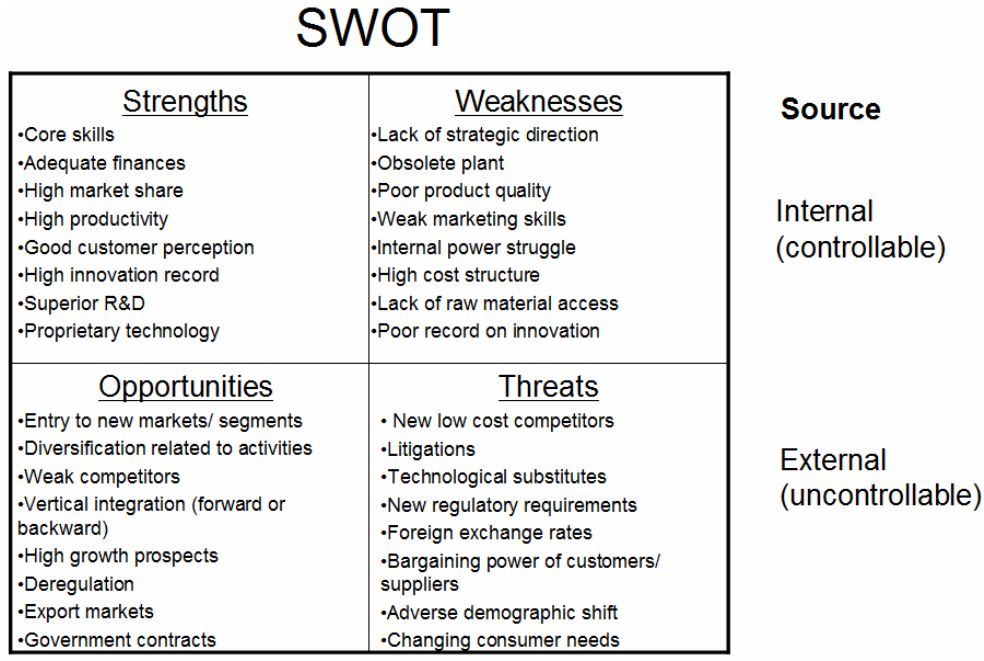


Table 6 SWOT analysis as per [74]

2) Advantages

The requirements for conducting SWOT analysis are relatively straightforward.

It is a simple yet structured approach for senior executives to conduct a reassessment of the business. The method illustrates the outcome of a marketing audit by resuming internal strengths and weaknesses (focussing on resources that would be valued by the costumer) and their relation to external opportunities and threats.

3) Disadvantages

In practice the distinctions to classify internal factors into strengths and weaknesses and external factors into threats and opportunities is difficult and it is not clear whether this is sensible and worth while. This is less important than the general detection of these internal and external factors combined with a profound assessment and analysis of their implications.

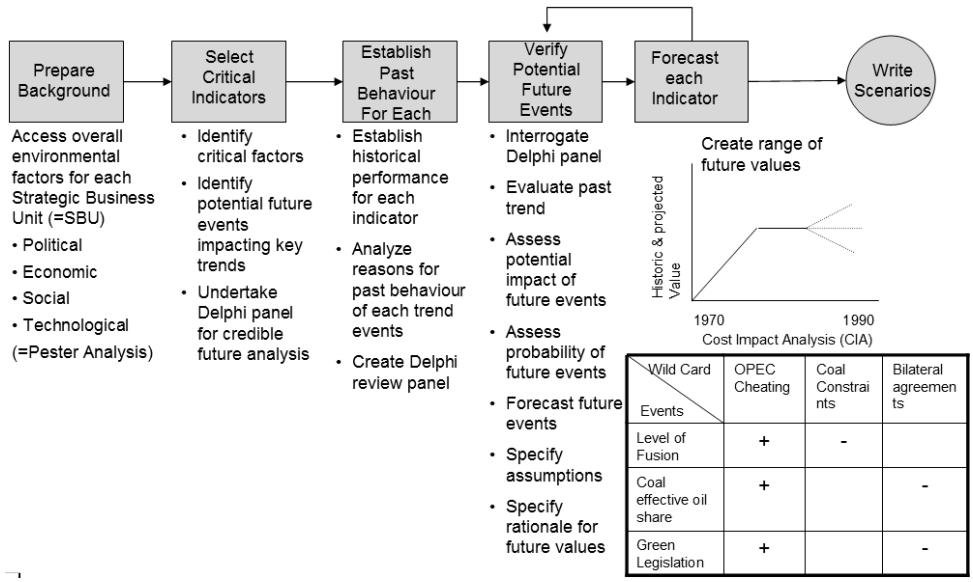


Figure 4 The process of scenario creation

D. Scenario and sensitivity analysis

A company's ability to adapt and react quickly to changes in its environment depends on its capacity to predict them. Even though a systematic analysis is a good preparation for the company's likely future, uncertainties remain. The wider environment and competitive actions can not be anticipated with certainty or the future might have been totally misinterpreted. When considering the future "tunnel vision" and "narrow-minded thinking" must be avoided. Adjustments to the initial strategy can be implemented more quickly if all possible outcomes have been considered and contingencies for further departures from the anticipated have been developed.

1) Description

In the case of scenario analysis, it is not a forecasting technique but a process to look at different hypothetical sequences of events. It focuses at different views on how the future may look. This process visualizes alternative futures in order to design flexible strategies to cope with identified visions and possible developments. The method has become well known ever since Royal Dutch Shell used it to successfully predict the first oil-price shock in 1973.

When we talk about sensitivity analysis, future cash flows are examined under different outcomes. Once a base case has been established it can be challenged in the sense that "What if demand or prices are 25% lower or higher". Or what happens to cost and earnings if predicted USD rate of 1.45 EUR/USD at year end will be +10 cents (/-10 cents) cents to 1.55 (1.35 EUR/USD) (e.g. for a company that produces in Euro zone but sells its products in USD)? Sensitivity analysis examines the robustness of the chosen strategy and reinforces the importance of planning with contingencies.

2) How to apply it

Scenario analysis begins in the present and makes assumptions about future environmental developments. The process of scenario creation is illustrated in Figure 4. Scenarios comprise subjective assessments of individuals or groups e.g. using Delphi analysis and consultations with experts and evince that decision makers have influence on future developments.

A series of 3 scenarios (optimistic scenario, best view scenario, worst case scenario) can be developed based on variant predictions.

3) When to use

Scenario analysis is especially useful when uncertainty is high. Environmental uncertainty can be currency appreciation or depreciation or the election of a different government. Industry uncertainties relate to for instance new legislation introduced or stalled. The aggressive expansion of a leading competitor can be characterized as an example for competitive uncertainties. Other conditions which favour the use of scenario analysis is when an industry is going through major changes and/or fundamentally different developments are conceivable. Or if new opportunities had not been detected previously and unforeseen trends have had costly aftermath for the company in the past.

E. War Gaming

1) Description

The purpose of driving a war game is providing the organizations with the possibility of testing their assumptions and beliefs and therefore better understanding the industry issues, emerging opportunities and threats.Specially when the market undergoes continues transformations is this technique of particular interest, as it helps management understanding how the competitors might react to the changes. The lessons learned from the War Game should be incorporated in the business strategy to improve the corporate planning processes

2) How to perform War Gaming?

There are several ways of carrying war gaming out. Usually several rounds are run, each one representing a different time period depending on the focus of the game (ranging from severtal months to one or two years). Shortening the length of the periods will help rather to take a tactical decision (instead of one strategic one).

Prior to the start of the game each team should be thoroughly briefed on the available knowledge on each organization. Typically, teams will then meet independently, in workshop sessions, and use the briefing information to plan what they would do during the first time period, playing the role of their chosen or allocated organization.

Following the completion of the round, players then announce their strategies and plans, leading to the second round. During the second round the teams take on board the different organizations' plans and modify their own for the following period. This process then continues for the agreed number of rounds. During each round, players need to anticipate the moves of other players, develop their own strategies, decide on what resources and funding are needed (and ensure that these exist and are allocated as necessary in their plans). Depending on the rules agreed prior to the start of the game, players may sometimes also communicate with other teams – for example to agree a joint-venture or merger. Following the actual game period, the participants then discuss the situation and the lessons learned.

3) Requirements for success

In order to produce the expected results it is indispensable fulfilling following preconditions:

- Sufficient Information on each of the organizations being examined
- Multidisciplinary teams with a good representaion of sales, marketing, general management, operations, finance, etc. This is important to widen the focus and to get people on the same page, so that jointly strategies can be defined

- A facilitator, especially if external, can be benefitial to run smoothly the workshops and act as an umpire to adjudicate in disputes between teams, and to suggest which teams' strategies are most likely to win out in the given situations. Sometime this role is taken up by a dedicated team.
- Adequate schedule and facilities.

4) When to use

War gaming can serve to multiple purposes depending on the way it is conducted:

- Gaining insights to the current situation, opportunities, threats and issues that may arise in the short-medium terms;
- Expressing recommendations and suggestions for future actions, that might have been tested during the game
- Corporate blind-spots identification (both in the organization and in the competitors), which allows for identifying vulnerabilities and thus, proactive strategies that can protect or take advantage of the weaknesses
- Identification of areas where intelligence on the market is missing.
- Corporate alignement between decision makers in different functional areas;
- Early warning on f how the market may change over the short-medium term

F. Others

Apart from the latter methods, there are further ones that might be also used, like financial analysis, win/loss analysis, etc. We intentionally leave them out of the scope of this work.

X. THE CI CYCLE

According to the classics process management discipline, a process is a sequence of value-adding stages designed to deliver a product and/or a service to external and internal customers. The methodology involves conceptualizing each step and substep where value is added through a series of transformation[14].

There have been many attempts of modeling the competitive intelligence process. Even before the competitive intelligence term was coined, plenty of information management processes were defined. If their scope was broader, they position themselves as the precursors of the CI models of our days.

The most prominent model has been re-adapted by Choo and highlights the classical six steps in information management, namely identification of information needs, information acquisition, information organization and storage, developing of information products and services, information distribution and information usage.

There are several models, as shown in following Figures, but all of them shared the same structure or at least the fundamental similarities:

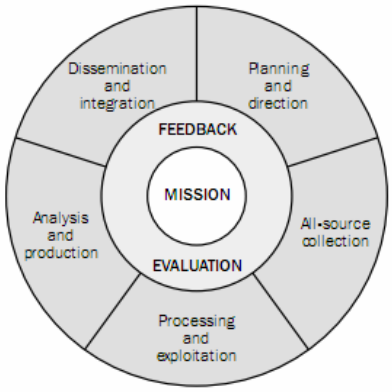


Figure 5 Intelligence process [15]

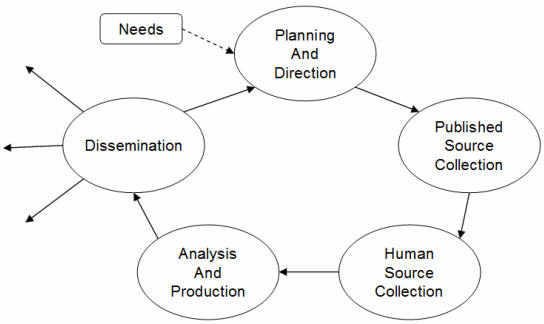


Figure 6 Know! proposed Intelligence Cycle

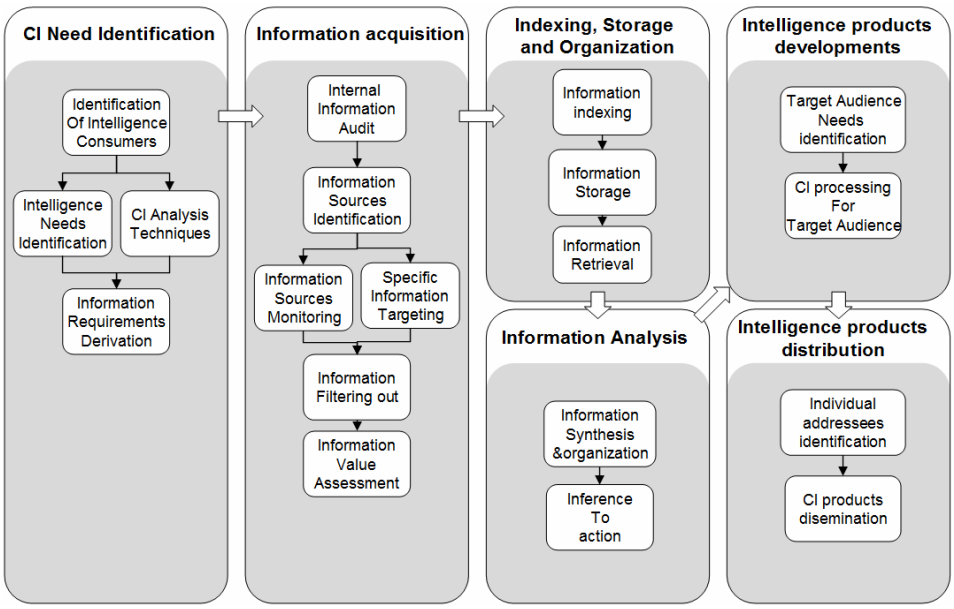


Figure 7 The CI cycle according Bouthillier

1) Identification of CI Needs, planning and direction

The intelligence process starts with a simple request of information that is required by somebody in your organization to help them making the right decisions. Defining intelligence needs is typically an iterative process performed according to a top-down approach. Starting with general "WH*" questions can help: Who? What? Where? When? Why? How?

The most complex yet crucial part is identifying where and how to focus the intelligence efforts (e.g.: What are the topics with which management should be concerned? How do we identify and define the critical decision challenges and information gaps?). In the CI lexicon this topics are known as Key Intelligence Topics (KITs)

A good start is a Five Forces Porter's analysis, but this is only the beginning because average manager is nowadays fully aware of the bargaining power of the customers and suppliers, the pressure of the intensifying competitive rivalry, the threats of substitute products and services, as well as new entrants in their markets. Defining intelligence needs requires thinking of priorities first: "What are those strategically relevant issues or topics 'where the outcome or resolution has a significant impact on the value of the firm'? [16] and "How can be define the legitimate intelligence needs of management?"

The KIT process involves translating the key decision-making needs of managers into topics and questions that can be made operative for collection and analysis by the competitive intelligence unit [15]. Executives express their concerns and challenges and the CI team reads out the intelligence requirements in terms of Key Intelligence Topics and Key Intelligence Questions they agree upon. KITs are basically statements regarding implications of a topic for the company and come directly from a systematic interaction between the intelligence consumer and the CI team.

There are 4 categories of KITs that are employed for the intelligence requirements definition:

a) Decision topics

Unlike other categories, decision topics are characterized by a deadline –after which the information will be of no use-, and can apply to any pending business decision (e.g.: development of strategic plans, capital expenditures likely to affect competitive positioning, acquisitions, joint ventures, etc).

b) Key player topics

Activities, capabilities, intentions and plans of industry competitors and others, for example alliance and joint-venture partners, emerging and existing competitors, major customers and suppliers, etc. These KITs are designed to provide a greater understanding of capabilities and intentions of targets and deeper insights unto their current and future actions.

c) Warning topics

Warning topics are those topics which are routinely monitored or tracked against pre-determined indicators, with no fixed end-date. Competitive Intelligence often reflect executives' suspicions or 'fears', if not a little paranoia, and start from the presumption that there are no 'happy surprises' in business. A warning topic, therefore, is one way of dealing with the question: 'What would we do if we were them?'

The three principal aims of an early warning topic are to [17]:

1. Identify current and future threats, including 'disruptive' changes in the industry, in government and in technology;
2. Avoid strategic surprise, especially competitor initiatives;
3. Spot new business opportunities.

d) Counterintelligence topics

Counterintelligence topics deal with at least four sets of important questions regarding the protection of a company's knowledge assets:

1. What must our firm protect?
2. What are our competitors (or foreign government agencies) trying to discover about us? And why?
3. How are they trying to do it?
4. What can we do, and what are we doing, to reduce their chances of getting it? What legitimate denial and deception tactics might we employ to safeguard our proprietary information?

As already stated, before starting any intelligence activity, company intelligence staff 'must be sure that [they] have the KIT issues right and the executive understands what will be done and can be expected as far as the final results'[19], which, in turn, involves ongoing dialogue between managers and intelligence.

Purpose

To identify your Competitive Intelligence (CI) needs

To understand how you would use intelligence

To obtain your ideas and suggestions regarding how the intelligence function, or system, cab best be developed by the company

Intelligence Needs

A. Decision-making (your area of responsibility)
- Planned/future
- Past examples
- Sources of external information
    - Written inputs
    - Experts
    - Personal networks
- Decision-making process:
    - Within business unit/division
    - For the company
- Suggestions to improve the quality of external information needed to make decisions

B. Early warning intelligence
- Examples of past surprises
- Concerns about the:
    - Company
    - Business
    - Industry
    - Others

- Subjects about which you believe the company needs to be well informed but at present is not

C. Competitors
- Which competitors are you most concerned about?
- What types of information intelligence do you need?
- What uses do you make of competitor intelligence?

D. Awareness
- Topics that you must regularly follow to do your job well
- External issues that have an impact on your business strategies and operations (e.g.: country risk, terrorist threats, regulatory)

Intelligence Uses
- What uses do you expect to make of intelligence (e.g,; market research, product and/or technology development, strategic planning, sales)?
- Who in your organization do you expect to be regular users of intelligence?
- What types of intelligence "products" would you like to see (e.g.: field reports, intelligence briefings, assessments, long-range estimates, research reports, warning alerts)?

Intelligence Capabilities
- Experience/familiarity with intelligence
- What types of intelligence/information do you receive at present?
- What intelligence capabilities does your business unit/division presently possess?
- What intelligence capabilities does your business unit/division need?
- Will your business unit/division be able to conduct intelligence operations to help other units/divisions? Any barriers=
- In your view, how should your company's intelligence system be organized?

Comments, Ideas, suggestions
- Today
- Afterthoughts (anytime)

Table 7 Key Intelligence Topics Survey Form [20]

Identifying the management's real intelligence needs and priorities, and then fulfilling these requirements within the context of a focused, systematic intelligence gathering and analysis process encompasses complex analysis of the industry to define for example the defining attributes of a company, a manager, a product, etc. Disregarding certain attributes might lead to insufficient information, whereas considering too many attributes results into a more complex and time consuming analysis that might render

the produced intelligence irrelevant due to changes in the competitive environment. Choosing the analytical technique by which information will be transformed into intelligence prior to deriving the basic information requirements is highly advisable, because different analytical approaches have different information needs (see Section IX).

The needs analysis depends very much upon the requester. As per to Bouthillier and Shearer [2], the output is highly dependant on the audience "consuming" it. For example, CI products that can be used as instruments in a company strategy decision taking process might not be relevant to other decision makers in the organization. In other words, this is a goal driven approach relying on the fact that CI client communities sharing the same goals will work with the same CI product.

The resulting deliverable of this step should be a well document list of KIT's and their derived information requirements together with the closure on one analytic tool. Once it is ready, the acquisition of competitive information is due to start.

2) Acquisition of competitive information

"'Data Data Data!' he cried impatiently. 'I can't make bricks without clay.'
Sherlock Holmes speaking with Dr. Watson."
Sir Arthur Conan Doyle -

After having defined the information needs, the second step in the CI cycle can take place. The collection or intelligence gathering determines how the responsible team for the intelligence will acquire the raw data and information that is required.

From the organizational point of view, the team has to decide from which sources the information will be obtained, when the collection should take place, who owns each particular collection activity, which is the most suitable format the information will be transformed in, what the integration method should be, etc.

It is important that people in charge of the analysis and those who perform the information sit together to develop gathering strategies that dictate which personnel will be tasked, which sources will be exploited, which gathering requirements have higher priority or required more granular gathering, how the information will be purged, formatted and consolidated to serve as input for the next step.

One recommended action item within the information gathering process is conducting an information audit. Usually the required information is implicitly available because have been already collected for other endeavor or simply exists as part of the organization documentation –reports, databases, publications, etc- and there's no need to collect it again resorting to external sources.

Apart from the gathering itself it is important to ideate a monitoring –scanning- strategy to fetch information that has been newly updated on the sources that have already been identified.

Once the information has been retrieved from a source, it is important to check whether the information obtained fulfills the information requirements identified in the previous step. It usually requires filtering out information that is superfluous or not relevant.

Depending on the type of source, the information gathering and monitoring methods may vary (see VIICI Sources), as well as the need for information validation, as already explained in the section "Choosing the right sources")

3) Organization, storage and retrieval

No matter how difficult the information gathering process was, it is imperative to implement an information organization and indexing strategy to ensure the efficient information retrieval.

Firstly the indexing criteria should be defined, such competitor name, product, customer, supplier, etc. The mere fact of introducing indexes ensures that pieces of information are linked to each other on related subjects, and can therefore be retrieved together.

Improved indexing quality results into more accurate information retrieval. Indexing should rely on a set of descriptors that represent every subject intelligence is required on. The major issue related to indexing is defining the appropriate categories so to ensure the consistency with the way the intelligence will be queried and retrieved.

4) Analysis

As discussed in the section X, there are several ways of analyzing the gathered and stored information. As mentioned above, it is highly advisable fix the analysis method on beforehand, as it can determine the information requirements and the way the information gathering takes place.

Analysis is the most important process of the CI, because out of information it produces intelligence, and consists basically of applying a set of techniques for sorting and comparing data and information, for deriving some interpretations and for developing and testing hypotheses based on various assumptions.

The best ingredients for a good analysis are subjectivity, intention, judgment, because grasping the big picture out of the information is not trivial. Actually, social sciences and humanities tools should be integrated, as this analytical step is rather of qualitative nature and requires the practitioners to develop a “sixth sense” to foresee the market situation.

The complexity inherent to the analysis and the lack of techniques as recipes that can be applied in all cases make the contribution of IT tools being just for information organizing purposes (it this sense, CI is similar to statistics, you can make use of tools to make data inferences but the interpretation of the results is up to the expert).

There are a couple of best practices related to this analysis [2]:

- Employ more than one analytical technique to extract the meaning of the information (e.g.: scenario development and war gaming are complementary): it allows for getting closer to the problem, reducing noise and getting more precision
- Get to the right level of information processing: it will ensure complete coverage of a technique and that you have considered all dimensions of the technique. Depending on the chosen technique, a tool can support the check for completeness (e.g.: in scenario developing, the generation of all potential scenarios)
- Get information synthesized to speed up the analysis
- Formulate recommendations as an outcome of the analysis: the intelligence has been well analyzed only if it leads to decision making and actions

$Data + Context = Information$
$Information + Meaning = Intelligence$
$Intelligence + Experience = Knowledge$
$Knowledge + Decision = Action$

Table 8: From Data to Action

There are some challenges associated to the analysis process [41]:

- Engaging in the policymaking. The role of the CI analyst is a supporting role… Use "if, then" to avoid crossing the line.
- Focus on low-probability, high-impact dangers and objectives… target for long-shot threats and opportunities. Intelligence analysis should provide expert, tough-minded assessments that concentrate, not on whether an event is likely, but how and why it might come about …'
- Choosing instead of just pointing: CI must identify and clarify the vulnerabilities of adversaries and the sources of the company's leverage' over partners and third parties as well as competitors.
- Timeline unawareness: delivering 'the best we've got' on time is better than disseminating a 'perfect' product too late.

5) Development of CI products

This step can be seen as a sub-step of the previous one. The following table shows typical CI products according to the intelligence type:

Intelligence Type	**Description**	**Deliverables and Examples**
Current Intelligence	- Provides the management with timely indicators about new developments that might have an impact on the company strategy. -- Tailored to meet particular management requirements with a given deadline. - A typical table of content of a current intelligence reports usually includes following sections: Key judgments, Scope, Introduction, Evidence/findings and Implications.	*CEO's daily brief (CDB).* The CDB addresses intelligence issues of the highest significance necessary for the chief executive officer to perform his or her duties to advance the strategic goals and objectives of the organization as well as safeguard its security.
		*Senior executive intelligence briefing (SEIB).*The SEIB is a compilation of current KITs. It is tailored to the needs those of executive and senior vice president rank, and is distributed several times a week.
		Competitive intelligence assessment This is an in-depth analysis of a strategically relevant development, event, issue or situation, providing the decision-maker with evaluation and judgments
Estimative Intelligence	- Includes long-range forecasts of key trends and their future implications for the organization. - Compiled at the direction of senior management for the purpose of helping executives envision the future (threats, opportunities, etc) the company is likely to face. - Purpose: to minimize the risk of major policy failure by reducing	*Estimative Intelligence report*: where following information can be found: - Outlook section: will assess the likely course and impact of important industry, market, scientific and technological developments, and identify the dynamics that will have the greatest impact on subsequent developments - Compilation of all relevant data and information – from human, open and other sources – that the CI unit possesses on the question; - Examination of the data and information: by an intelligence

	decision-makers' uncertainties about the external environment	team, who then make corresponding estimative judgments; - Description of the principal forces at work in the given question under examination
Research Intelligence	*Basic Research intelligence:* - Analysis on key regional, market, competitor and political events or topics. - Used by decision-makers to support new initiatives or is sometimes stored in anticipation of future 'crises'.	*Monographs and memorandums:* - Involves the preparation of what are commonly referred to as 'competitor profiles' by validating the key data and assumptions held about a competitor, combined with a 'so what' analysis of what the information means. - Is confidential and highly focused
	Operative Research Intelligence: - Intended to produce intelligence to satisfy the needs of managers with operational responsibilities	For example: Intelligence is due to answer following inquiry "What changes in the competitor's sales force composition?"
Scientific &Technical Intelligence	- Provide insights on rivals' process technologies (e.g.: where improvements in a competitor's efficiency could substantially improve their cost advantages) - Not only the "whats" and "hows" should be answered but also the "whys"	*Technology intelligence report:* responding critical questions like "why are they doing that?", "what do they know, or what assumptions do they hold that we do not, and what are the implications"?
Warning Intelligence	- The means by which companies anticipate, detect and where possible prevent, or at least mitigate, strategic surprise. - Intended to provide insights on future competitor intent or potential competitive threats posed by new industry and market entrants.	*Warning Watchlist:* regular report that tracks and assigns probabilities to potential threats to the company's strategic interests or security that may develop within a fixed time frame, e.g. three to six months.
		Warning alert: report that usually includes: statement of facts, analysis and outlook, implications for the firm, intelligence gaps and possible actions.

Table 9 Intelligence products classification (based on [15])

6) Distribution of Intelligence products

The dissemination step is crucial. Even if intelligence is collected and analyzed correctly, it will be of no value if the product is not conveyed to the end user in sufficient time for him to act upon it. A famous example in the Roman context was the episode in which a list of conspirators was thrust into Julius Caesar's hand shortly before he was assassinated. Caesar's intelligence network

had done its job. Had the dictator read the message and acted upon it, he might have survived. Taking advantage of the intelligence product–the decision to act–is not a function of the intelligence apparatus. If the commander or statesman has all the information yet makes a bad decision, it is not an intelligence failure but incompetence or poor judgment on the part of the intelligence consumer [30]

Depending on the intelligence consumer, the intelligence products should be presented in one or another way (e.g.: according with the management's decision taking, etc)

If we classified the intelligence in terms of damage the company will suffer if the information were to become public or end up in the competitor's hands, we will get a much better view of its value.

Traditionally, there are four ways of delivery:

1. Oral delivery. This is how intelligence analysts can be most certain of bringing to light what the user really needs to know. The results from the dialogue and feedback that takes place between intelligence staff and their customers when meeting face to face

2. Inclusion of intelligence reports from the field. Local assessments of intelligence issues or problems add considerable credibility to analysis completed by a central intelligence department. It helps answer the question: 'How do you know?'

3. Laying out the evidence. Decision-makers value seeing evidence that supports analysis and conclusions. The intelligence consumer is, in practice, the ultimate analyst, and in most cases will anyway exercise that privilege.

4. Inclusion of optional actions and implications. What actions might decision-makers wish to consider taking? What are the implications?

PART II: TECHNOLOGIES ASSOCIATED TO CI

XI. INTRODUCTION

This section will present the technologies that are typically associated to any competitive intelligence activity. After introducing them, a mapping of these technologies to the CI cycle sub-processes is also presented

XII. A TYPOLOGY OF CI TECHNOLOGIES

A. Technologies involved

1) Text mining technologies:

Based on linguistic patterns, it is about conducting higher-level text analysis to derive high-quality information. It usually makes use of language recognition technologies that rely on dictionaries where the concepts are semantically interlinked.

Unlike the current way web searching is understood, where the user is typically looking for something that is already known and has been written by someone else, the text mining tries to push aside all the material that currently isn't relevant to the user's needs in order to find the relevant information. Moreover, the goal is to discover or infer unknown information from the available one.

The difference between regular data mining and text mining relies on the fact that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. Databases are designed for programs to process automatically; text is written for people to read. We do not have programs that can "read" text and will not have such for the foreseeable future. Many researchers think it will require a full simulation of how the mind works before we can write programs that read the way people do.

a) Text discovering tool

These tools automatically extract concepts out of documents and map out the relationship between them. It can be of great help for the CI professional to automatically index and store the documents.

2) Natural Language Processing

The NLP has been research subject for many years and is now making a lot of progress in performing small subtask in combination with text mining. NLP is mainly concerned with the interactions between computers and human languages and therefore can be divided into natural language generation or how to put the database information in a human readable format, and the natural language understanding which takes up the other direction, namely converting human language into a more formal representation that can be manipulated by a program.

The NLP deals with particular sub problems, like *speach segmentation* or how to convert the analogical speech signal to a sequence of discrete characters, *text segmentation* or word boundaries identification (especially for Asian languages like Chinese, Japanese and Thai), *part-of-speech tagging* or marking the words in a text as corresponding to a lexical category, *word sense*

disambiguation or selecting the right meaning in a polysemy case depending on the overall context, *syntactic ambiguity* (see Table 10), etc

The usage of NLP in competitive intelligence is commonly done "behind the scenes", that is assisting other techniques like web mining, information extraction, information retrieval, etc to deal with the complexity of the natural language

We call a garden path sentence to the one for which the responder's most intuitive interpretation is an incorrect one, ultimately luring them into an improper parse. The NLP understanding systems will irremediably fall into this syntactical ambiguity trap unless a semantic disambiguation is applied. Following examples can illustrate it: - "The Company entered the Indian market because it required more off-shoring capabilities" (we know "it" refers to "The Company" because we now the properties of the entity "Company" and we now that Companies may require off-shoring capabilities) - "The Company entered the Indian market because it is expanding" (who is expanding? The Indian market or the company? No way to disambiguate it with the knowledge about the involved entities "the Indian market" or "The Company". It requires an explicit semantic disambiguation coming from the writer) In other situations where the meaning of a sentence depends upon the position where the spoken emphases is put, semantic disambiguation is also required. E.g.: "The CEO never stated it was our fault" "The CEO never stated it was our fault" – maybe someone else stated it? "The CEO never stated it was our fault" – maybe he simply didn't ever stated it "The CEO never stated it was our fault" – maybe he never explicitly stated it, but implied it in some way "The CEO never stated it was our fault" – maybe it didn't ever get stated, but something else "The CEO never stated it was our fault" – maybe it was someone else's fault

Table 10 NLP and syntactical ambiguity

3) Automation Text summarizing

Summarizing is basically creating a shorter version of a large text (usually called source) which makes sense and contains the most important points of the original document. This summary can be *generic* or targeted to answer certain questions known as *query relevant summary*

The most popular approaches are extraction based summarizing (which consists of merely copying the information deemed most important by the system to the summary –key clauses, full sentences or paragraphs-) and abstraction based summarizing (that goes beyond and paraphrase certain sections of the text to achieve more condensation). Developing text summarizing algorithms for the second variant usually requires the use of natural language generation

In some cases, the summary has multiple sources (i.e.: summarizing of news about the same topic).

Assessing the quality of a good summary is a very subjective task also for humans, but there are 2 factors that can be quantified like coherence and coverage (see [56])

Due to the huge volume of information that a CI practitioner has to deal with, these tools are becoming more and more important.

4) Information extraction and information retrieval

a) Information retrieval

Information retrieval is the term conventionally, though somewhat inaccurately […]. An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request[57]

Characteristic	**Data retrieval**	**Information retrieval**
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive
Matching	Exact match	Partial match, best match

Table 11 Information Retrieval vs Data Retrieval

Information retrieval is therefore a core technology for the intelligence information gathering. The web is at present queried by search engines that can be seen as key-words driven information retrieval systems.

b) Information extraction

Consists of the extraction of single facts from text fragments and their presentation in i.e. one schema (e.g.: facts about a competitor expanding in the Indian market, etc). The categories the facts belong to are known on beforehand –the user knows the categories about the facts she doesn't know yet). It differs from traditional information retrieval techniques in that it does not recover from a collection a subset of documents with are hopefully relevant to a query, based on key-word searching (perhaps augmented by a thesaurus). The goal of IE is to extract from the documents (which may be in a variety of languages) salient facts about pre-specified types of events, entities or relationships. Once these facts have been extracted, they get stored into typically a database to be analysed, exploited somehow or just made available for other applications.

Figure 8 Information retrieval vs Information Extraction

If we consider that over 95 percent of the digital universe is unstructured data (as per Jonathan Martin from HP Research Labs), we clearly see the importance of information extraction.

Let's think for a moment of internet as the universe of data. In the real universe, most of the stuff in clusters of galaxies is invisible and, since these are the largest structures in the Universe held together by gravity, scientists then conclude that most of the matter in the entire Universe is invisible. This invisible stuff is called 'dark matter' (see Figure 10)

The electronic dark matter exists: most of the stuff on the web is invisible and, since these unstructured documents are the largest data type in the web Universe held together by links, scientists then conclude that most of the data in the entire web is invisible.

A CI practitioner to be effective needs to explore and exploit the electronic dark matter. Hence, the importance of the information retrieval for the CI

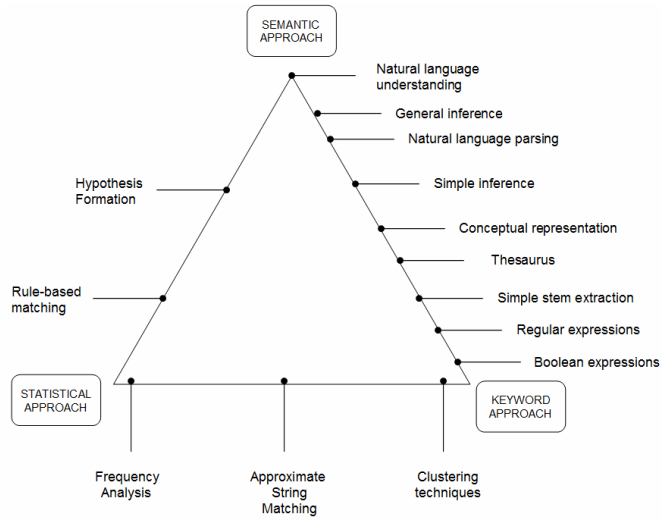


Figure 9 Representation of technologies around information extraction and retrieval

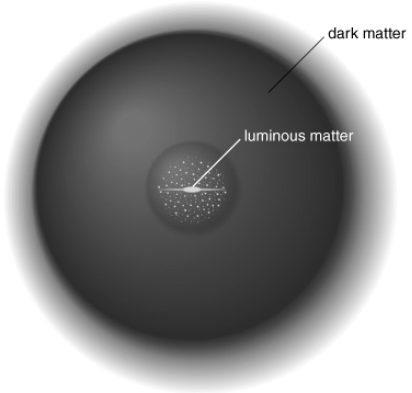


Figure 10 The dark matter

5) Analysis and reporting tools

a) Statistical packages

A statistical package is a suite of computer programs that are specialized for statistical analysis. It enables people to obtain the results of standard statistical procedures and statistical significance tests, without requiring low-level numerical programming. Most statistical packages also provide facilities for data management.

There are plenty of statistical software tools available on the market. Having such a variety of choices, it's essential to get clear on the specific requirements the software shall meet. A part from the cost factor, the use-ease, the availably of in-built self-explained tutorials, etc. you could compare them from the pure functional points of view:

- Descriptive Statistics: base stats, normality tests, etc
- ANOVA: one way, two ways, Manova, GLM, post-hoc test, Latinsqrs, etc
- Regression: linear regression, polynomial regression, other types of regression
- Non-parametric statistics: Cotingent Tables Analysis, etc
- Quality control:
- Time series analysis: Base Series Processing, f.ex. differing and smoothing, Analysis, i.e.. moving average etc.
- Survival analysis: cluster analysis, discriminal analysis
- Data processing: Base Data Processing (i.e. sorting) or extended processing (e.g.:.data sampling, transformation)
- Charting: for standalone packages

The CI practicioner, depending on the analysis method selected to transform the intelligence in knowledge, will require a lot of data statistical processing, which makes meaningful the usage of one of these statistical packages, or just a simplified set of operations that can be solved without such a tool overhead (i.e.: using standard tools like Excel, etc)

b) Analyzing and Reporting suites

These tools extract data, perform pattern detection, slice, dice and drill-down upon the data to allow end-to-end analysis and derive meanings and conclusions and offer several customizable reporting options

6) Intelligent agent technology

a) Active filtering tools

You can configure them to monitor websites, documents, emails, etc to filter information corresponding to certain input parameters that are used in the query.

Some of them are equipped with machinery learning algorithms, so that user preferences can be learned. Collaborating with other technologies, can automatically summarized findings, delete out-of-date records or forward information as part of an alerting mechanism

These agents are constantly used in the CI, especially in early warning systems

b) Media monitoring services

They are intended to provide documentation, analysis, or copies of media content of interest to the clients. These services are becoming more and more specialized by media or content (e.g.: services monitor for news and public affairs content, advertising,

sports sponsorships, product placement, video or audio news releases, use of copyrighted video or audio, infomercials, "watermarked" video/audio, etc)

Media monitoring services are probably the oldest method for automating competitive intelligence. Media monitoring is a relatively inexpensive and timesaving form of gathering information, though not as fast or comprehensive as some other methods.

7) Information searching, indexing and retrieving

According to whoever provides the actual search service, free search tools can be categorized into remote site search service and the server-side search engine. In the former, the indexer and query engine run on a remote server that stores the index file. When it comes to the time of search, a form on a user's local Web page sends a message to the remote search engine, which then sends the query results back to the user. A server-side search engine is what we usually think of as a search engine. It runs on the user's server, and takes that server's CPU time and disk space. In this paper, the term search engine refers only to server-side search engines.

According to what is indexed, search engines are classified as file system search engines and website search engines. File system search engines index only files in the server's local file system. Website search engines can index remote servers by feeding URLs to web crawlers. Most search engines combine the two functions, and can index both local file systems and remote servers. The nine search engine software packages compared here are all website search engines, some of which can index local file systems.

A fully functional website search engine software package should have the following four blocks:

a) Web Crawler

Also called Spider, basically follows HTML links in Web pages to gather documents. There are countless websites using this technology to provide up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam).

A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies.

Crawling the web is extremely difficult because of the fast changing of pages, the large volume of documents and the dynamic page generation. To deal with these difficulties, it is required to prioritize all downloads (as only a certain fraction of the web documents can be downloaded given a certain time). The dynamic nature of the web explains that after finishing the crawling of a site, new pages can have been added or deleted. Dealing with server-side generated pages introduces the problem of crawling duplicate content –massive http get request with all kind of query string parameter combinations might return the same content-as they must sort through endless combinations of relatively minor scripted changes in order to retrieve unique content.

To handle the explained issues, there are a set of policies whose combination determines the behavior of the crawler:

A *selection policy* that states which pages to download: a crawler always downloads just a fraction of the Web pages due to the unprocessable volume, that's why it is so important that the selection contains the most relevant pages and not just a random sample of the Web. To ensure that, it is essential to implement a prioritization mechanism that computes the importance of a page depending on its intrinsic quality, popularity (links and visits) and its URL. This mechanism should work with partial information, as the complete set of Web pages is not known during crawling.

There are 3 strategies that are commonly combined to get better results, namely breadth-first, backlink-count and partial Pagerank calculations

A *re-visit policy* that states when to check for changes to the pages, due to the volume of web documents, the crawling might take so long that before finishing, many events could have happened (creations, updates, deletions, etc). Therefore and to avoid (and measure) the cost of not detecting an event, functions like freshness and age of resources are defined.

The objective of the crawler is to keep the average freshness of pages in its collection as high as possible, or to keep the average age of pages as low as possible. These objectives are not equivalent: in the first case, the crawler is just concerned with how many pages are out-dated, while in the second case, the crawler is concerned with how old the local copies of pages are.

A *politeness policy* that states how to avoid overloading Web sites: Crawlers can retrieve data much quicker and in greater depth than human searchers, so they can have a crippling impact on the performance of a site. Needless to say, if a single crawler is performing multiple requests per second and/or downloading large files, a server would have a hard time keeping up with requests from multiple crawlers.

To keep it under control, the robot exclusion protocol (robot.txt) has been created to specify which parts of the web servers should not be accessed by crawlers.

A parallelization policy that states how to coordinate distributed Web crawlers: to maximize the download rate, the crawlers usually run multiple processes in parallel. A policy to assign new URLs discovered during the crawling process to threats is required to avoid crawling the same URL twice.

b) Indexer

The indexer module collects, parses, and stores data to facilitate fast and accurate information retrieval.

The most extended engines implement the full-text indexing of online, natural language documents. Media types such as video and audio and graphics are also searchable.

Unlike the cache-based search engines, the meta search engines reuse the indices of other services instead of storing them along with the corpus.

We can distinguish between full-text indices and partial-text services, where the depth indexed is restricted in order to reduce the index size.

Indexing can take place in real time, which is the case of agent based search engines, or within indexing intervals, for larger services due to processing cost.

The principal benefit of search engines is to speed up the information retrieval over a corpus of documents, that is, retrieving very fast, those that are relevant for the query. The lack of indexing would make necessary the full scanning of documents for each and every query, which will require a lot of computing resources and time.

The key points to be taken into account:

- Storage techniques: or how to store the index data
- Index size or how much store is required to support the index
- Lookup speed or how quickly a word can be found in the inverted index
- Fault tolerance or how reliable is the service
- Merge factors or how data enters the index in a multithread indexing

c) Query Engine

This module basically performs the actual search and returns ranked results. After the user enters the searching query (usually by means of key words), the engine examines its index and provides a list of best-matching web document (usually with a short summary containing the document's title and sometimes parts of the text).

To allow for the refining of the query, many modern search engines support the use of boolean operators AND, OR and NOT

Another modality offered by advance search engine consists of allowing the users to enter the distance (or number of words) between the key words to enable the so called proximity search. Some commercial search engines implement a modification of the proximity search that establishes the occurrence order between the key words.

The quality of the query engine can be measured by the relevance of the returned results, which is achieved by providing methods to rank the results.

The way the search engines determine the best matches and in which order depends upon each particular implementation and evolves over the time, as Internet itself evolves. Nonetheless, there are two main streams: human-driven predefined and hierarchical keywords –programmed by humans extensively-, or inverted index systems that takes place whenever a document is found –which relies much heavily on computer processing-

d) Interface

It can be defined as the aggregate of means that allows users to interact with the query engine. It basically provides an input, or a means to express the searching query, and an output, or a way the search engine presents the retrieved results.

The interface usually offers filtering and search options to help narrow down the search. Common filters include:

- Top level categories: web pages, web sites, news, images, video, blogs, etc
- File formats options: HTML, PDF, PPT, etc
- Date restrictions : updated last week, last month, last year, etc
- Location restrictions: within a site, within a country, etc
- Language restrictions: only pages in a certain languages, etc

It is advisable to present these options in a way that doesn't require much overhead and compromises that way the user experience.

Once the results have been returned, the results page presents the result ranking and items. This pages is usually divided in four sections [71]:

- Summary of search results (usually including the search phrase, the number of results displayed in the current page, total number of results, etc)
- Results item list (including the title of the page as a link, a small description, the URL, etc)
- Navigation to other result items and options for the next search
- Error handling (typos and spelling corrections, suggestions, etc)

e) Multi-media searching

The broadband is doubling over next 3-5 years and consequently the video enabled devices are emerging rapidly. We can talk about an emergence of mass internet audience and a shift of mainstream media into the web.

The depicted scenario ties with the so call "Democratization of Mass Media in the 21st Century", where media is

- Of the People (media is ultimately defined by users)
- By the people (production, tagging)
- For the people (consumption, devices, personalization

The media search relies on leveraging the media meta-content from the web (inferred meta-data), from communities (tagging and user submissions, or the usage of structured Meta-Data from different sources

More and more relevant information will be available in a non-textual format, and the CI professional should be concerned

f) Sentiment analysis

Generally speaking, it aims to determine the attitude of a speaker or a writer with respect to some topic. The attitude may be their judgment or evaluation (see appraisal theory), their affective state (that is to say, the emotional state of the author when writing) or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

We can see it as a mere translation of the vagaries of human emotion into hard data. For many businesses, online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace.

Yet many companies struggle to make sense of the caterwaul of complaints and compliments that now swirl around their products online. As sentiment analysis tools begin to take shape, they could not only help businesses improve their bottom lines, but also eventually transform the experience of searching for information online.

For an CI professional, this field is a gold mine and a for-free tool to find out what the customers thinks about the own products and services and the competitors' ones

8) Document and content management

a) Document management systems (DMS)

Document management systems can have a much focused scope, comprising just certain document collection within a give department, or have a broad focus, like enterprise content management systems. Nevertheless, there are certain commonalities these systems share:

They need a location where documents are stored and a means to ensure that documents will be kept secure and protected from unauthorized accesses. In case of a disaster, a disaster recovery plan is also required to recover the affected documents. Archiving is also a core functionality to enable audit trailing and future accesses.

Filing methods to organize and index the existing documents for further retrieval are also a key feature. The retrieval itself can be understood as searching capabilities together with browsing on the encountered collection of documents.

System administrators usually require a workflow mechanism to reproduce the document life cycle. Stati are defined and a set of valid transactions to move from one state to further ones. Transactions are bound to activities that need to be performed in order to enable the new state. It is also important to provide some traceability to understand the transactions history for a given document.

For a Competitive Intelligence practitioner, the interaction with document management systems is two folded: on one hand they must be able to retrieve corporate information to cross with information gathered from other sources; on the other hand, some deliverables can be incorporate to the CI section of a corporate DMS… even the CI cycle can be implemented as an information transformation process whose steps are configured in the DMS workflow

9) Information aggregators

Information aggregation is a service that gathers relevant information from multiple sources in order to provide value by analyzing the aggregated information for specific objectives using Internet technologies.

We call the providers of these service aggregators. In a broader sense, information intermediaries such as newspapers, magazines, professional journals, and more recently, increasing number of web portals are information aggregators since they all collect information from multiple sources and disseminate it for convenient consumption [80]

The aggregators collect, categorize, and regroup information from multiple sources. In addition, they perform analysis to the aggregated information.

We distinguish:

- Comparison Aggregation: retrieves information about a set of attributes for a product/service offered by many competing vendors and normalizes the information for meaningful side-by-side comparison.

- Relationship Aggregation: enables customers to manage multiple accounts with a single logon. A relationship aggregator can collect the information on a customer's behalf and generate various useful reports.

- Intra-organization and Inter-organization Aggregation can aggregate relevant information from disparate sources to promote knowledge sharing and perform firm level analysis.

10) Multipurpose portals

Are considered like an integration space (service hubs of mails, news, etc) to provide access to internal and external sources.

They are usually designed as a platform, with groupware capabilities, automatic information retrieval, classification, monitoring software, etc.

The aggregation and delivery of such a variety of sources makes these portals especially attractive for CI

11) Business Intelligence and e-Business applications

Business Intelligence suits are conglomerates of software applications aims at providing assistance to the corporate strategy. If we dissect a Business Intelligence tool, we will find following components:

Online analytical Processing or OLAP: intended to quickly answer multi-dimensional analytical queries. It is typically used in business reporting for sales, marketing, management reporting, business process management, budgeting, forecasting, financial reporting, etc. It relies on a multidimensional data model, allowing for complex analytical and ad-hoc queries with a rapid execution time (based on hierarchical and navigational databases that are faster than traditional relational databases
Digital dashboards: or executive information system's user interface designed to be easy to read, to save time by running multiple reports in parallel, to gain total visibility of all areas instantly, to align strategy and organizational goals, to measure efficiencies and inefficiencies, to identify and correct negative trends, etc… that is, to make more informed decisions based on the collected business intelligence

Reporting and query software: tools that extract, sort, summarize, and present selected data

Business performance management: consists of a set of processes that help organizations optimize their business performance. It provides a framework for organizing, automating and analyzing business methodologies, metrics, processes and systems that drive business performance

Process mining which allows for the analysis of business processes based on event logs. The basic idea is to extract knowledge from event logs recorded by an information system. Process mining aims at improving this by providing techniques and tools for discovering process, control, data, organizational, and social structures from event logs

Data mining capabilities to extracting patterns from data and so, transform these data into information.

Local information systems to support geo- based reporting

The CI professional can leverage each and every of these subsystems to ease the CI information gathering, analysis and the creation of intelligence products (i.e.: presenting the in a dashboard, etc)

XIII. CI TECHNOLOGIES MAPPED TO THE CI CYCLE

Depending on the particular phase within the CI Cycle presented before, the need for a particular technology may vary. Following table contains a summary on which technology is required when:

	Needs definition	**Information Gathering**	**Information Organization**	**Analysis**	**Creation of Intelligence Products**	**Dissemination**
Text mining technologies		*	*			
Text discovering tool		*	*			
Automation Text summarizing				*	*	
IE and IR						
Information retrieval		*	*			
Information extraction		*	*			
Analysis and reporting tools						
Statistical packages				*	*	
Analyzing and Reporting suites				*	*	*

Intelligent agent technology						
Active filtering tools		*	*			
Media monitoring services		*	*			
Information searching, indexing and retrieving						
Web Crawler		*	*			
Indexer		*	*			
Query Engine		*	*			
Interface		*	*			
Sentiment analysis		*	*			
Multimedia content		*	*			
Document and content management						
Document management systems (DMS)	*	*			*	
Information aggregators		*			*	*
Multipurpose portals		*	*			*
Business Intelligence and e-Business applications				*		

Table 12 CI Technologies along the CI Cycle

PART III: THE SEMANTIC WEB

XIV. INTRODUCTION

I invite the reader to have a look back to the computers history… Computers were born to help processing complex numerical calculations. After a few years corresponding with PC boom, the computers entered plenty of homes and enterprises and their usage extended to the information processing and entertainment area, being the typical applications database systems, text processing applications, games, etc. The internet started like a side product of a researcher's community, but its development's pace has been at least dramatic, as shown in the Appendix I.

Whatever topic one might think of can be found in this internet space… Tones of terabytes about this topic… but the discovery and retrieval have become highly complex because the information is too unstructured, too repetitive, too difficult to retrieve…

That's why there is one subject one should retake, namely making machines better understand the human beings in order to better support them handling this complexity… That's the semantic web *raison d'être*

XV. THE FUNDAMENTALS

A. *The subject, the verb, the object and RDF*

RDF stands for Resource Description Framework and is an approved recommendation for the Semantic Web at the World Wide Web Consortium (W3C). It is basically a standard specification for data and modeling used to encode metadata and digital information and it's of highest interest for us, because the semantic web vision revolves around it.

Let's recall for an instant how we learned the sentence structure in grammar school and apply it to a given sentence. For example:

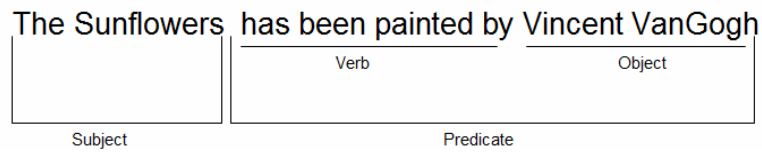


Figure 11 Sample of typical semantic analysis

We used to identify the subjects the predicates, and objects of this sentence:

- The subject as the noun or noun phrase that is the doer of the action. The subject of the sentence tells us what the sentence is about (in the example *"The Sunflowers"*)

- The predicate as the part of a sentence that modifies the subject and includes the verb phrase. In other words, the predicate tells us something about the subject (in the example *"has been painted by"*)

- The Object as a noun that is acted upon by the verb (in the example *"Vicent VanGogh"*)

In RDF is very similar, the subject is the resource that is being described by the ensuing predicate and object. Therefore, in RDF, we want a URI to stand for the unique concept “The Sunflowers” like “http://www.example.org/art#TheSunflowers” to denote that we mean the master piece and not the name of a sunflowers oil brand. An RDF resource stands for either electronic resources, like files, or concepts, like “person.” One way to think of an RDF resource is as “anything that has identity.” The predicate is a relation between the subject and the object. Thus, we would define a unique URI for the concept “has been painted by” or better “painter” like “http://www.example.org/art#painter”.

The object is either a resource referred to by the predicate or a literal value. In our example, we would define a unique literal value for “Vincent VanGogh”.

Note that there is a difference between resources and literals: whatever referred by an URI is a resource (e.g.: “http://www.example.org/art#TheSunflowers”). The other possibility is the literal (or sequence of characters) “Vincent VanGogh”. Thus, we distinguish between resource-valued predicates –that may vary over time- and literal-valued predicates –that can be seen as constants- (see Figure 12)

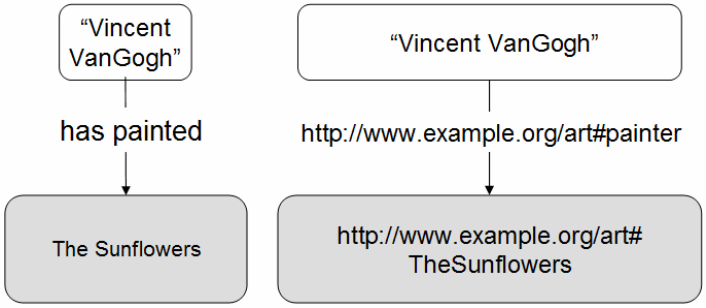


Figure 12 Triple and its URI based representation

As you might have noticed, in order to provide unique names in RDF, URIs are employed. When an URI is used as a qualifier for a given set of names is also known as namespace; however, not all URIs are namespaces. Because RDF is often used to describe federated data, the URIs are often addressable, but they don’t have to be.

Commonly used vocabulary namespaces in RDF:

- RDF: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- Dublin Core: http://purl.org/dc/elements/1.1/
- SKOS: http://www.w3.org/2004/02/skos/core#
- FOAF: http://xmlns.com/foaf/0.1/

There are many ways of expressing RDF description, but perhaps the most extended one is XML (see Table 13), but other formats like N3 (see Table 14) or Turtle are also quite extended

```
<?xml version=”1.0”?>
 <rdf:RDF xmlns:art=" http://www.example.org/art#"
     xml:base=" http://www.example.org/art"
     xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
```

```
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <rdf:Description rdf:about="http://www.example.org/art#TheSunflowers">
    <art:painter rdf:resource="http://www.example.org/art#Vincent_VanGogh"/>
   </rdf:Description>
 </rdf:RDF>
```

Table 13 RDF example

```
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix art:  <http://www.example.org/art#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl:  <http://www.w3.org/2002/07/owl#> .
art:VanGogh  art:painter art:SunFlowers;
```

Table 14 N3 example

B. The Web Ontology Language

The Web Ontology Language or OWL (and not WOL, intentionally given to leverage the association with the animal and the connection to the Greek symbol of Intelligence) is a W3C specification and formalism to enable the creation, publishing and distribution of ontologies (we will discuss about ontologies in detail later on). It's about the formal description between the terms in a domain and their meaning so that not only humans but also software agents understand this meaning.

OWL is based on RDF syntax and goes beyond the RDF Schema by extending it to allow the formulation of predicate-logic like expressions.

From the OWL perspective, an ontology is a set of "individuals" and a set of "property assertions" which relate these individuals to each other, or in other words, a set of axioms which place constraints on sets of individuals (called "classes") and the types of relationships permitted between them. These axioms provide semantics by allowing systems to infer additional information based on the data explicitly provided.

The competitive advantage of OWL in comparison with the previous attempts to build large ontologies relies on the explicit logical basis for the language based on description logics (which is a family of logics that are decidable fragments of the first order logic).

OWL is this respective both:

- a syntax for describing and exchanging ontologies
- has a formally defined semantics that gives them meaning.

The reliance on Description Logics grants the OWL with two basic yet indispensable properties:

- OWL implements the open world assumption: stating that whatever is not specified is not automatically taken as false
- OWL is monotomic, meaning that adding new statements (information) to our knowledgebase never falsifies a previous conclusion

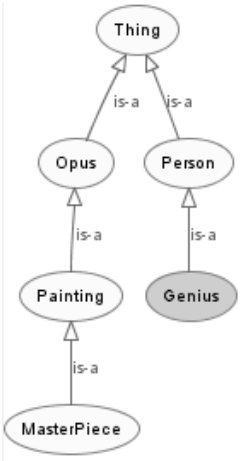


Figure 13 Asserted model of OWL Genius Ontology

1) OWL elements

The Web Ontology Language enables the creation of classes and its properties, instances or individuals and their operations:

Classes: User-defined classes which are subclasses of root class owl:Thing. A class may contain individuals, which are instances of the class, and other subclasses (see the Table 15 Genius Ontology OWL example lines 16 to 36)

Properties: A property is a binary relation that specifies class characteristics. There are 2 kinds of properties:

- Datatype properties help describe individuals — they are not typically used to describe classes and are certainly not dependent on classes. The set of allowable values for datatype properties are typed literals (or literal values with a specific datatype)
- Object properties: allow you to create associations or relationships between two individuals. That means the subject and the object the triple are both individuals. Let's see it in Table 15 Genius Ontology OWL example lines 11 to 15: we define the object property "hasPainted" and we use it in line 47 between the individuals "VanGogh" and "SunFlowers"

Instances: Instances are individuals that belong to the classes defined. A class may have any number of instances. Instances are used to define the relationship among different classes. In our example in Table 15, Vincent VanGogh, "The Sunflowers" and the "TheOldMill" are defined as instances.

Operations: OWL supports various operations on classes such as union, intersection and complement. It also allows class enumeration, cardinality, and disjointness.

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://www.semanticweb.org/ontologies/2009/8/6/GeniusSample.owl#"
     xml:base="http://www.semanticweb.org/ontologies/2009/8/6/GeniusSample.owl"
     xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:GeniusSample="http://www.semanticweb.org/ontologies/2009/8/6/GeniusSample.owl#"
     xmlns:owl="http://www.w3.org/2002/07/owl#">
     <owl:Ontology rdf:about=""/>

<!-- Object Properties -->
     <owl:ObjectProperty rdf:about="#hasCreated"/>
     <owl:ObjectProperty rdf:about="#hasPainted">
        <rdfs:subPropertyOf rdf:resource="#hasCreated"/>
     </owl:ObjectProperty>

<!—Classes -->
     <owl:Class rdf:about="#Genius">
        <owl:equivalentClass>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#hasCreated"/>
                <owl:someValuesFrom rdf:resource="#MasterPiece"/>
            </owl:Restriction>
        </owl:equivalentClass>
        <rdfs:subClassOf rdf:resource="#Person"/>
     </owl:Class>

     <owl:Class rdf:about="#MasterPiece">
        <rdfs:subClassOf rdf:resource="#Painting"/>
     </owl:Class>

     <owl:Class rdf:about="#Opus"/>

     <owl:Class rdf:about="#Painting">
        <rdfs:subClassOf rdf:resource="#Opus"/>
     </owl:Class>

     <owl:Class rdf:about="#Person"/>

<!--  Individuals -->
     <owl:Thing rdf:about="#SunFlowers">
        <rdf:type rdf:resource="#MasterPiece"/>
     </owl:Thing>

     <MasterPiece rdf:about="#TheOldMill">
        <rdf:type rdf:resource="&owl;Thing"/>
```

```
44      </MasterPiece>
45
46      <owl:Thing rdf:about="#VanGogh">
47         <rdf:type rdf:resource="#Person"/>
48         <hasPainted rdf:resource="#SunFlowers"/>
49         <hasPainted rdf:resource="#TheOldMill"/>
50      </owl:Thing>
</rdf:RDF>
```

Table 15 Genius Ontology OWL example

2) OWL sublanguages

The W3C-endorsed OWL specification includes the definition of three variants of OWL, with different levels of expressiveness:

- ***OWL Lite*** was originally intended to support those users primarily needing a classification hierarchy and simple constraints. For example, while it supports cardinality constraints, it only permits cardinality values of 0 or 1. It was hoped that it would be simpler to provide tool support for OWL Lite than its more expressive relatives, allowing quick migration path for systems utilizing thesauri and other taxonomies. In practice, however, most of the expressiveness constraints placed on OWL Lite amount to little more than syntactic inconveniences: most of the constructs available in OWL DL can be built using complex combinations of OWL Lite features. Development of OWL Lite tools has thus proven almost as difficult as development of tools for OWL DL, and OWL Lite is not widely used.

- ***OWL DL*** was designed to provide the maximum expressiveness possible while retaining computational completeness (either φ or ¬φ belong), decidability (there is an effective procedure to determine whether φ is derivable or not), and the availability of practical reasoning algorithms. OWL DL includes all OWL language constructs, but they can be used only under certain restrictions (for example, number restrictions may not be placed upon properties which are declared to be transitive). OWL DL is so named due to its correspondence with description logic, a field of research that has studied the logics that form the formal foundation of OWL.

- ***OWL Full*** is based on a different semantics from OWL Lite or OWL DL, and was designed to preserve some compatibility with RDF Schema. For example, in OWL Full a class can be treated simultaneously as a collection of individuals and as an individual in its own right; this is not permitted in OWL DL. OWL Full allows an ontology to augment the meaning of the pre-defined (RDF or OWL) vocabulary. It is unlikely that any reasoning software will be able to support complete reasoning for OWL Full.

The following Table 16 summarized the differences between these 3 sublanguages.

	Lite	**DL**	**Full**
Compatibility with RDF	Theoretically,no rdf document can be assumed to be compatible with OWL Lite	Theoretically,no rdf document can be assumed to be compatible with OWL DL	All valid rdf documents are OWL full

Restrictions on class definition	Requires separation of classes, instances, properties, and data values	Requires separation of classes, instances, properties, and data values	Classes can be instances or properties at the same time..
RDF Mixing	Restricts mixing of rdf and owl constructs	Restricts mixing of RDF and OWL constructs	Freely allows mixing of RDF and OWL constructs
Classes Descriptions	The only class description available in OWL lite is IntersectionOf	Classes can be described as UnionOf , ComplementOf, IntersectionOf, and enumeration Eg: class can be exhaustively defined by its instances. For example defining a class DaysOfWeek exhaustively by Sun, Mon, Tue, Wed, Thurs, Fri, Sat	Classes can be UnionOf, ComplementOf, IntersectionOf, and enumeration Eg: class can be exhaustively defined by its instances. For example defining a class DaysOfWeek exhaustively by Sun, Mon, Tue, Wed, Thurs, Fri, Sat
Cardinality Constraints	Cardinality: 0/1 MinCardinality: 0/1 MaxCardinality: 0/1	Cardinality>= 0 MaxCardinality >= 0 MinCardinality >= 0	Cardinality>= 0 MaxCardinality >= 0 MinCardinality >= 0
Value Constraints	owl:allValuesFrom Owl:someValuesFrom Object type for owl:valueFrom should be a class name or class identifier	Owl:allValuesFrom Owl:someValueFrom Owl:hasValue	Owl:allValuesFrom Owl:someValueFrom Owl:hasValue
Metamodeling	Does not allow metamodeling	Does not allow metamodeling	Allows metamodeling. Thus RDF and OWL constructs can be augmented or redefined
Class	OWL:class is subclass of RDFS:class	OWL:class is subclass of RDFS:class.	RDFS:class and OWL:class are equivalent

Table 16 OWL Lite vs DL vs Full [55]

C. The reasoning capabilities

Reasoning is the act of making implicit knowledge explicit. For example, an OWL knowledge base containing descriptions of students and their parents could infer that two students exhibited the 'brother' relationship if there were both male and shared one or more parent. No explicit markup indicating the 'brotherhood' relationship need ever have been declared. A *Reasoning Engine*

is computational machinery that uses facts found in the knowledge base and rules known a priori to determine Subsumption, Classification, Equivalence, and so on. F-OWL, FaCT, and Racer are examples of such engines. OWL Full is so expressive that there are no computational guarantees that inferences can be made effectively and it is unlikely that any such engine will be able to support all its features soon. However, OWL Lite and subsets of OWL DL can be supported.

In general, one can talk about several types of inference which are demonstrable in Description Logics-based systems, including those systems captured in OWL:

- ***Consistency -*** determine if the model is consistent. For example, presents an OWL model containing the facts: (a) cows are vegetarian, (b) sheep are animals, and (c) a 'mad cow' is one that has eaten sheep brain. From these facts a computational reasoning engine can infer that 'mad cows' are inconsistent since any cow eating sheep violates (a).

- ***Subsumption –*** infer knowledge structure, mostly hierarchy; the notion of one artifact being more general than another. For example, presents a model incorporating the notions (a) 'drivers drive vehicles', (b) 'bus drivers drive buses', and (b) a bus is a vehicle, and subsumption reasoning allows the inference that 'bus drivers are drivers' (since 'vehicle' is more general than 'bus').

- ***Equivalence*** - determine if classes in the model denote the same set of instances

- ***Instantiation -*** determine if an individual is an instance of a given Class. This is also known as 'classification' – that is, determine the instance of a given Class (e.g.: in the example of Table 15, we can see that (a) "The Sunflowers" is a Master piece (b) "The Sunflowers" has been painted by "VanGogh" (c) VanGogh is a person (d) Genius is defined as a person that has at least painted one Master piece. Hence, it can be inferred that "VanGogh a genius is"

- ***Retrieval -*** determine the set of individuals that instantiate a given Class

XVI. DATA INTEGRATION

Nature laughs at the difficulties of integration.

– Pierre Simon de Laplace

A. The current situation

The biggest challenge and effort driver in the IT resources available in the industry is the integration of existing disparate data sources. The problem exists since the world's second computer was built.

The problem can be formulated as a single and (apparently) quite simple requirement: the data from one system should work effectively inside a completely different system. This requirement that has been formulated in one sentence needs to be fulfilled whenever a new system is built and needs to work within a given landscape with legacy systems.

The same investment is done over and over by several projects because of emerging requirements that makes the integration with legacy systems highly difficult

The integration and recycling of information sources lead to the creation of several data bases, which keeps the problem ongoing and getting a larger dimension

From the developer point of view this situation obviously presents drawbacks:

- High redundancy among the resources implies wasting of development capacity
- Parallel scanning of resources is almost impossible. Plenty of queries to be executed in order to retrieve the available information
- Relevant dependencies or contradictions between data remain hidden, because the information is spread out among different data bases
- Users to get used to different data models and different interfaces, which requires ramp up time
- Need for custom development to integrate each and every data source
- Suboptimal data exploiting: small data bases stay unexploited. Adding of new data sources requires manual intervention for discovery and biding, which means the model does not scale.

There have been approaches to overcome such problems but their results have been in the best case a good short-term solution, but not sustainable. The fact of adding a new data source implied remapping with other databases schema or with a central data schema and eventually building a new connector, which lead to a cost explosion and to a poor designed home-grown system

The problem starts right after a new data source project goes live: the developed data source becomes difficult to access from outside because of the lack of a semantic basis and application context

B. Understanding the problem

If we have a look at the way application are built, we will notice, that data integration is an irrefutable need: in the first place business analysts start interviewing people or observing the behaviour of a process or activity, and out of that produce a process modelling document or artefact. The people in charge of data modelling take as input the process modelling document and create a data model, that highly depends on the skills and modelling style of the modeller. The software architect comes then into picture to produce the so called model-driven architecture, including class diagrams, object modelling, etc. There are a lot of

methodologies and architectural styles (RUP, GoF, etc) and therefore the resulting architecture can heavily vary. Once the implementation partners get involved three tendencies motivates the heterogeneity of these projects: they usually know better and try new technologies, they try to leverage experience from other projects, and Commercial off-the-shelf (COTS) products.

The immediate result of this process is a distancing from the reality on each step, but it doesn't change the fact that the software aims at solving: the users and the funding exist in the real world.

The same process model in different hands can produce different data models; the same data model, different architectures; and different implementation teams even given the same data model and same architectural specification will follow different ways of implementing the system. The interoperability problems start with this indeterminism: hardly ever two applications handle the same information in the same way

C. Making systems interoperable

The only way to ensure that 2 applications intended to address similar problems can interoperate is by designing them together. There are several ways to get interoperable systems:

- Use one system only (à la ERP): it requires adapting the business processes to the software or in other words IT ends up driving the business. The problem is that there's no "out-of-the-box" reality and there's no system able to scale to the whole enterprise and trespass the enterprise borders. So at the end of the day you will have to make a lot of local adaptations that will compromise the interoperability.
- Create a corporate data model: having them as reference can be a good idea, but as they are inevitably someone's view of the world, they end up failing. Beside that, there are specialist terminologies used by specialist communities that are reluctant to employ a foreign vocabulary (nobody is willing to work according to someone else's view of world)
- Apply the Enterprise Application Integration: despite of a lot of total loose coupling required, they might work in an IT environment that is always supervised and controlled. Some semantic integration issues are present
- Move to a Service Oriented Architecture: it requires splitting the business and the IT into components and creating then an interoperable services-based enterprise, but again IT ends up driving the business. To be successful it has to be top-down designed and it shifts the interoperability problem to lower, encapsulated levels.
- Design for interoperability: illusory, utopic… no deterministic technique to ensure that the same data structures will be the result of 2 implementation of the same problem
- Keep databases in perfect synchronization by building a replicator: the complexity might render the attempt unuseful, The maintenance of such a replicator can be huge. It has to be reprogrammed each time a new data source needs to be integrated.
- Create your custom ETL integrator: a lot of effort and money for a short term solution that is not sustainable. Moreover, ETL and replication solutions are typically part of a larger solution that may include business intelligence, analytic applications, or other data integration solutions — but incompatible system metadata results in lost productivity.

D. The root cause of the problem

The existing data and the way they are structured needs to be taken into consideration, from the very beginning on. You can get the best information about how the business works by analyzing and examining the data the systems in place (also the legacy ones) rely on.

E. The solution: ontology based modeling and a building methodology

Leveraging the OWL/RDF as a common metadata framework for enterprise infrastructure will be one step towards sustainable and flexible data integration. OWL/RDF is powerful and expressive and enable new benefits like higher reuse, better developer productivity, and end-to-end impact analysis features that prevent unforeseen technical outages.
Even the developers that used to build a custom-made integrator, will profit from a data model view layer so that, ensuring highly expressiveness and portability

At the end of the day, it is about creating a commonly accepted model that all the system intended to be interoperable should comply with, and a deterministic way to create the model. This model should be a way to express formally a shared understanding of information [50] or an ontology. Ontologies have become the cornerstone to structure the complex knowledge domains and establish standards.

1) Ontologies between Philosophy and Computer Science

As indicated in [51] the fact that ontologies is a plural raises the major difference between the philosophical and computer science approach to the term.

A philosophical ontology would encompass the whole of the universe, but computer scientists allow the existence of multiple, overlapping ontologies, each focused on a particular domain.

Indeed an understanding of the ontology of a particular domain may be crucial to any understanding of the domain.
The combination of ontologies, and communication between them, is therefore, a major issue within computer science, although such issues are problematic with the philosophical use of the term. At the limit, an ontology that perfectly expresses one persons understanding of the world is useless for anyone else with a different view of the world. Communication between ontologies is necessary to avoid this type of solipsism.

2) Ontologies usage

a) Reference for naming things

The motivation behind it is establishing a set of controlled terms for labelling entities in databases and data sets.

It will ensure the consensus between people on the name to be given to certain entity, and the consensus between people and machines to identify and name things. The immediate consequence of this consensus is the fact that computers can help researches to make sense of massive data available to perform analysis on. The challenging side is the variety of synonymous terms and polysemy or lexical ambiguity, defined as the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings[52]

The biggest effort driver is the unification of disparate data that are labelled differently in different data sources. Thus, where the ontology adds value is in “fixing” the terminology so that people can label medical entities in a consistent way.
Additionally, synonymy, acronyms and abbreviations can augment the ontologies.

The most generic ontologies provide their entities with the *is-a* and *part-of* relations to other entities, which constitutes the basis for the knowledge representation. These relations support the creation of computer reasoning applications, able to infer subsumption (is-a relations) or composition (part-of relations) between entities.

Making use of a central ontology as a disambiguation mechanism it is possible to query different data sources at a time.

Apart from term fixing, these ontologies can be used for term extraction and better information retrieval on web documents

The description of audio visual information is also address by the usage of ontologies to provide i.e. names for ontology entities present in images

b) Representation of encyclopaedic knowledge

The second natural step to capture and represent knowledge is by means of rich relationships between the entities of a domain. The textual description of complex knowledge gives the humans the possibility to access this knowledge, but not the machines. Using well-defined, univocal, standardized relationships to structure and make explicit the knowledge enable the access to machines and humans.

These ontologies should be the result of the subject matter experts and knowledge engineers collaboration and shouldn't be created with the aim of a particular application, but trying to provide a holistic representation of the encyclopaedic knowledge belonging to a domain.

c) Information model specification

Specifying information and data models using ontologies instead of a conventional modelling language, like UML, provides several advantages: explicit specification of the terms used to express information in the particular domain, augmented capabilities like explicit relationship making among data types and automatic reasoning –subsumption and composition-, visualization capabilities of complex structures (i.e.: the ones offered by tools like Protégé), publishing of the information model in the semantic web (standards have been adhered –like OWL-), etc

d) Specification of data exchange format

The emerging of multiple data based containing information related to a particular domain requires a mechanism to specify the standard exchange format. The ontological capabilities for structuring information are being more and more used.

e) Semantic based Information Integration

The integration scenario of heterogeneous yet related data sources requires manual ad-hoc processing currently based in syntactic-based methods (e.g.: linking object with the same name facing polysemy, acronyms, abbreviations and synonymy related issues). Specifying the semantics of data in a variety of databases can enable researchers to integrate heterogeneous data across different databases. Linking entities in different data sources based on shared characteristics supported by an ontology that provides a common declarative foundation to describe domain specific content has proven to be a better approach. The additional ontological reasoning capabilities can support the linking process and resolve ambiguity and at the end of the day facilitate the integration and validation of disparate information.

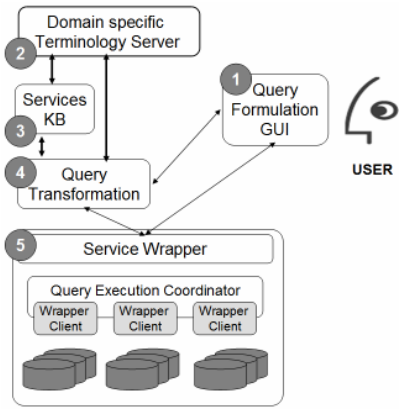


Figure 14 Ontology driven integration middleware

The Figure 14 shows the architecture of a generic information retrieval system based on an ontology driven middleware (inspired by [53])

1. The user interacts with Query Formulation Dialogues, expressing queries in terms of the domain model. The dialogues are driven by the content of the model, guiding the user towards sensible queries. The query is then passed to the transformation process, which may require further user input to refine and instantiate the query.
2. The Terminology Server provides services for reasoning about concept models, answering questions like: What can I say about concept Y? Or what are the parents of concept X? It communicates with other modules through a well-defined interface
3. The Services Knowledge Base links the domain ontology with the sources and their schemas. This information is used by the transformation process to determine which source should be used.
4. The Query Transformation module takes the conceptual source-independent queries and rewrites to produce executable query plans. To do this it requires knowledge about the information sources and the services they offer information about particular user preferences - say favourite databases or analysis methods - may also be incorporated by the query planner. The query plans are then passed to the wrappers.
5. The Wrapper Service coordinates the execution of the query and sends each component to the appropriate source. Results are collected and returned to the user.

f) Computer reasoning with data

The competitive advantage of representing the knowledge by means of ontologies is the possibility to exploit knowledge by means of computer reasoning or the capability of making inferences based in the knowledge contained in the ontology, the contextual information and the asserted facts. For a scientist the panorama looks like a huge amount of well-structured information and a set of tools to analyze this information and allow for drawing meaningful inferences.

This steps means shifting from the mere information retrieval to the meaning of information mindsets. Typically, when a researcher is formulating hypothesis, it's extremely difficult to verify that the data available support this hypothesis and if no, to figure out where the inconsistencies are. The need for tools capable of querying and interpreting the information at hand is becoming more and more incipient.

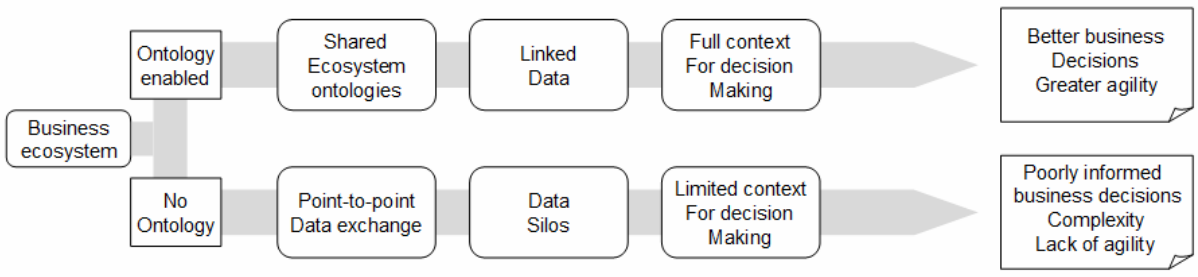


Figure 15 How corporate strategy can benefit from ontologies

3) The BORO Method

It is a systematic way of making the ontology building process deterministic.

It's recommendable to re-engineer legacy data or in general every kind of data intended to be integrated into a new model. One very beneficial characteristic is the repeatability in results.

Additionally, the process is defensible because it is based on anyone's view of the world (this is an ontology in the sense) and traces back to things in the real world.

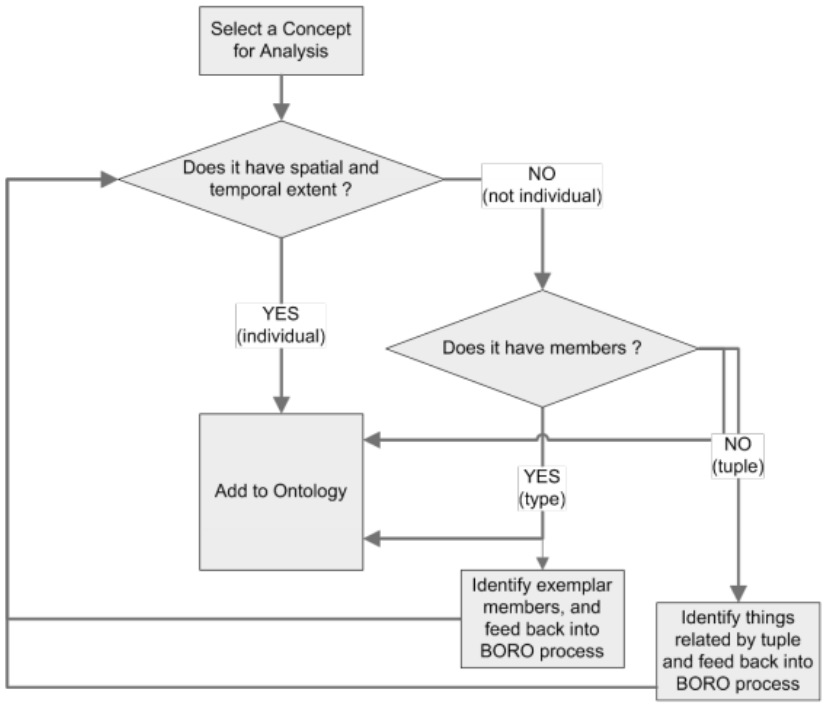


Image Crown Copyright, BORO Process Owned by BORO Centre

4) A wishful vision

The Semantic Web as integration tool is not intended to extend or improve existing integration tools, but to displace them

Bringing the data into a semantic web compatible format eliminates or reduces substantially the need for physical integration of different formats, syntaxes, structures, etc. That's exactly the wishful vision for a near future where most of business software applications will make their data available in RDF/OWL format, so that data will be accessible, linkable, combinable, re-usable, etc. and all of that for granted… No need any longer for manually connecting data in a non-standard infrastructure

XVII. BRIDGING THE GAP TO THE SEMANTIC WEB

The internet has become the most important shop window for companies to present their products and services. The advent of the Web 2.0 and the user generated content possibilities have transformed the web from a facts space to an opinion space. Reviewing products, services has become a common practice

The information processing techniques have not moved at the same pace and still work with facts that are usually assumed to be true (e.g.: web crawlers, etc).

Facts can be expressed by means of key or topic words but not opinions. This reality renders the entire document retrieval system and document ranking strategy implemented by the most popular search engines simply inappropriate.

1) The importance of Natural Language Processing

Before the semantic web becomes a widely accepted reality in the World Wide Web and the mostly of the sites guarantee the compliance, the web stays as an space where everything is for human consumption (for people to view) rather than for computer systems to process.

With the more and more significant exception of the multi-media content, we can say that all the information is expressed in natural language.

In the field of information extraction, when the problem comes down to find instances of a particular expression, it clearly won't do to just look for word-by-word matches; to be at all successful, matching must occur at a structural level. So the crucial problem here, at the heart of many NLP applications, is the accurate identification of the structure of sentences and entire discourses.

2) Syntactically structured information extraction

For design and usability reasons, the internet content is confined in regularly structured data objects. The inherent nested nature of HTML and the design pattern repetition and the HTML nesting capabilities enable the existence of these structures. Even free text fields where users can add their own content, like for example blog entries or product reviews, etc are displayed on the target page in a structured way after been submitted by the creator.

To automatically retrieve for example the competitors' product palette, to benchmark products, to monitor price variations, etc it is important to know how they are displayed on their site.

We can distinguish basically two types of pages with structured data: List pages, and detail pages

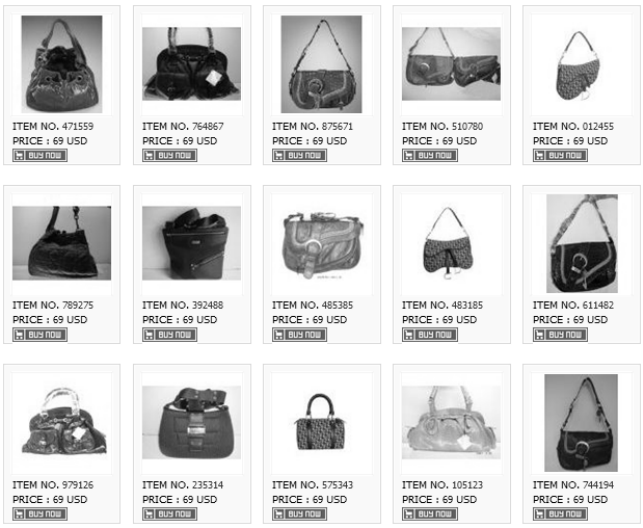


Figure 16 List of products example

Figure 17 Product detail example

The HTML documents can be exploited as tag trees having <HTML> as root. The advantages of using the DOM-like tree consists of applying navigation methods

a) Wrapper induction

Wrappers are commonly used as such translators. A wrapper is a procedure, specific to a single information resource (i.e.: a list page or a product detail page) that translates a query response to relational form (e.g.: in the Figure 17 “What it the price of Item No. 423833?”). Wrappers are typically hand-coded; unfortunately, hand-coding is tedious and error-prone. A good alternative is applying machinery learning methods to learn extraction rules and patterns on a given set of manually labeled pages).

Wrappers can be used for information integration purposes including a mediation middleware capable of breaking down the user’s query as wrapper-specific sub-queries and once the results are retrieved, able to combine them.

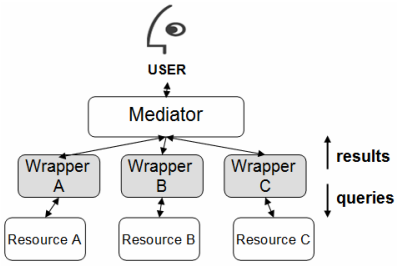


Figure 18 Wrapper-based information extraction architecture

Wrapper induction requires the manual labeling of the examples, which is an intensive and time consuming labor, especially if one wants to extract data from a huge number of sites. This supervised learning technique learns data extraction rules from the set of manually labeled positive and negative examples. The learned rules are then applied to extract target data from other pages using the same template.

There are a lot of wrapper induction systems depending on the learning algorithm:

Wrapper	**Description**	**Pros**	**Cons**
WIEN [59]	There are several approaches: - *Left-Right (LR):* LR consists of a set of k delimited pairs (left and right) that have to be matched in the text. - *Head-Tail-Left-Right (HLRT):* same as LR, but filters out firstly delimiting regions by learning their delimiters - *Open-Close-Left-Right (OCLR)*: using open and close delimiters to indicate the beginning and end of each tuple - *Head-Tail-Open-Close-Left-Right (HOCLRT)*: combines HTLR and OCLR - *N-LR and N-HLRT*: modifications of LT and HLRT to handle nested structured	- Very simple yet powerful - Can implement multiple databases - Handles delimiting regions - Covers almost 70% of all existing web pages	- Does not allow optional attributes - No handling of nested structures. - Assumes that the items are always fixed and their order is known (Cannot handle permutations and missing items) - Must label entire page - Requires large number of examples
STALKER [61]	Treats a web page as a tree-like structure Handles IE hierarchically Use disjunctions to deal with variations.	- More efficient than the WIEN -Higher expressiveness - Powerful extraction language (eg, embedded list) - One hard-to-extract item does not affect	- Generate imperfect rules - Higher complexity - Still no answer on how to generate label pages automatically for the learning

		others - Can handle almost 90% of the web sites	
Softmealy [60]	First learns a finite-state transducer (FST) that encodes all possible sequences of attributes where each state represents a fact to be extracted. Unlike the previous ones, it uses separators ("invisible" borders) instead of delimiters, that are learned by defining their left and right context with contextual rules (state transitions)	- Simple enough to be learnable from a small number of examples of extractions - Complex enough to handle irregular attribute permutations (missing attributes, multiple attribute values, variant attribute ordering) - Uses wildcards (eg, Number, AllCaps, etc)	- Must "see" all possible permutations
RAPIER	Based in a logic framework (ILP) Integrates some NLP (part-of-speech tags) Bottom-up learning with lgg: select two examples and compute the minimal generalization that covers both		
SRV	Uses a large variety of features both for structured and unstructured text Implements a top-down rule learning (Ripper-like)		
WHISK	Rules represented as perl-like regular expressions Can handle (semi-)structured and unstructured text Implements a top-down rule learning with seed instance Can be enhanced with user-specified semantic classes to handle polygraphy		Slower

The major shortcomings of wrapper induction based Information Extraction can be summarized in following points:

- They are not generic but dependant on each page.
- A page structure change might render the learned method unusable. It opens a research direction on how to automatically re-label a wrapper that stopped working based on two questions: a) if the site changes, does the wrapper know the problem? –wrapper verification problem-, and b) if the change is correctly detected, how can the wrapper be automatically repaired? –wrapper repair problem-.
- Preparing a set of instances as example can be labor intensive and time consuming. Indeed, choosing the set of learning examples for the user to label is deciding and not trivial. To avoid unnecessary labelling, active learning is proposed as an approach to help identify informative unlabeled examples.
- The extracted information can't be exploited as it is, but needs to be stored in a relational DB or any other medium that allows for querying it.

b) Automated data extraction

Due to the problems related to the supervised wrapper induction, automatic extraction has been studied by researchers in recent years. Automatic extraction is possible because data records in a Web site are usually encoded using a limited number of fixed templates. It is possible to find these templates by mining repeated patterns in multiple data records.

We will focus now on the problem of generating extraction patterns for a given page with multiple data records without supervision.

We can address the problem by following approaches:

- identify data regions and/or data records based on pattern detection, like string matching algorithms (handling HTML as an string) or tree matching algorithms (exploiting the tree-like HTML structure)
- using multiple alignment algorithms

Let's talk first about the particular problem of generating extraction patterns for a given page. The first task to be performed consists of identifying the page regions that might contain the target data to be extracted. A valid assumption to make states that a group of data records that contains descriptions of a set of similar objects are typically rendered in a contiguous region of a page and are formatted using similar HTML tags.

To identify the so called data record regions, a tree matching approach that compares different sub-trees to find similar ones can be applied. This approach is computationally prohibitive because every single start and end tag has to be analyzed.

Another valid assumption relies on the fact that a set of similar data records are formed by some child sub-trees of the same parent node. This observation makes it possible to design a very efficient algorithm based on tree mapping to identify data records because it limits the tags at which a data record may start and end [58]

The tree matching algorithms start building a DOM tree (usually relying not only on the HTML string structure but also in the visual information or on how the rendering engine would interpret ill-formatted tags.

Once the tree has been built, the mining for data regions is performed intended to find a list of similar data records. This mining usually relies on some string or tree distance definitions between two sub-trees… further information like HTML tag labels, visual information and textual context might also be used to determine the matching of two nodes.

Finally, the data records are extracted from the identified regions by matching the corresponding data items from all data records. Usually alignment techniques are pursued that align multiple DOM trees by progressively growing a seed tree.

The question now is how to generalize the extraction pattern we have employed. For that, it is possible to generate a grammar for data extraction (grammar induction), which is not a trivial tasks. The result is a set of tree-based regular patterns for later extraction.

If instead of one page, you are given a set of positive pages where data records exist, to generate extraction patterns, the problem should be approach in another way.

The generated wrapper should be an union-free regular expression (i.e., no disjunction). The method works progressively, taking a sample page to start that is considered the first wrapper. This algorithm is called RoadRunner and described in [62]

This wrapper is then refined by solving mismatches between the wrapper and each sample page, which generalizes the wrapper. That is, it progressively infers a common grammar.

Matching is the core of this inference. The HTML sources are treated as lists of tokens, each token being either an HTML tag or a string, and works on two objects at a time: (i) a sample, i.e., a list of tokens corresponding to one of the sample pages, and (ii) a wrapper, i.e., a regular expression. The idea is to parse the sample with the wrapper: a mismatch occurs when some token in the sample does not match the grammar of the wrapper. The matching algorithm tries then to generalize it. This is done by applying suitable generalization operators. The algorithm succeeds if a common wrapper can be generated by solving all mismatches encountered during the parsing. To start, one of the sample pages is taken as the initial version of the wrapper.

Another approach proposed in [63] starts by defining a set of tokens or equivalence classes that have the same frequency of occurrence in every page. Then, the sets are expanded by differentiating the roles of the token using context related information, so that the same token in different contexts are treated in a different way. The page template is built by using the equivalence classes based on what is between two consecutive tokens, empty, data or list.

c) Natural Language processing to assist the structured data extraction

The structured information extraction can be supported by NLP techniques in two particular scenarios: Item matching based on text and integration of data from multiple sites.

3) (Web) Data integration

The information integration problem has been intensively researched since the early 80s. Firstly the problem was identified in the context of data base information. The section IXVI deals thoroughly with the problem, but now we will focus in the particular problem of integration data from different web pages

The key issue is the schema matching or the creation of a mapping between the attributes of the schemas that need to be integrated so that the resulting one corresponds semantically to the integrating ones. In other words, it is about merging the schemas into a single, global one.

For example, the Figure 17 attributes could have also be described naming the attributes in another way or just splitting “Price” into “Price” and “Currency”.

Represent the mapping with a similarity relation, $\cong$, over the power sets of S1 and S2, where each pair in $\cong$ represents one element of the mapping. E.g.,

Item No ≅ Reference Article

Price ≅ {Price, Currency}

Model Name ≅ Model

Available sizes ≅ Sized

Before starting any integration task, the information needs to be brought to the adequate state by means of pre-processing techniques like tokenization (breaking items into atomic words using a dictionary), expansion of abbreviations, removing of stop-words, stemming, standardization of irregular words, etc

There are several levels of matching, depending on the information considered:

a) Schema-level

Only schema information is considered. It relies on information such as name, description, data type, relationship type (e.g., part-of, is-a, etc), constraints, etc. The matching can be more difficult depending on the cardinality of the schema attributes –i.e.: 1:m is less complicated than n:m- work on it.

The automatic detection of matching candidates can be solved by means of linguistic techniques applied to names (equality, synonyms, equality of hypernyms, common sub-strings, cosine similarity[3], user-provided name match like a domain dependent match dictionary, etc), descriptions (after preprocessing can cosine similarity can be also applied) or any other information about schema attributes.

b) Domain and instance-level only matching

Where some instance data (data records) and possibly the domain of each attribute are used. This case is quite common on the Web. Then, the value characteristics are used in matching.

The domain (also known as type) can be simple –having only a single component whose value cannot be decomposed- or composite, where each value contains more than one component.

For simple domain learning techniques to extract data type pattern by means of regular expressions can be applied (e.g.: zip code, phone numbers, real, integers, emails, money, etc). When the data type is numeric, statistical magnitudes can be employed. In case we compare text, the cosine similarity can be an adequate method…

For composite domain matching methods based on the delimiters can be applied.

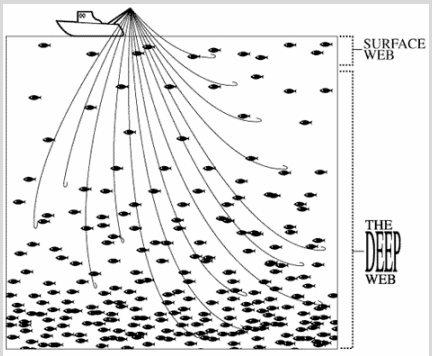
c) Integrated matching of schema, domain and instance data

Both schema and instance data (possibly domain information) are available. Similarities from many match indicators can be combined to find the most accurate candidates.

[3] Cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining. Given two vectors of attributes, A and B, the cosine similarity, θ, is represented using a dot product and magnitude as

What have been mentioned so far is valid for any kind of data base integration. In the particular case of web information integration, there are additional tasks related to the particularities of the web documents.

If we make the abstraction of web usage as a *search&retrieve* system, where every single site is like a small isolated database with a particular querying interface (i.e.: a search mask, etc), the integration problem extends from the data itself to the way this data is queried. Moreover, certain information is only available after a query have been launched and the results have been retrieved and the resulting web document is dynamically created, which makes impossible the information extraction just by browsing (see Table 17 about the deep web)



Approximately only 20% of the total Internet information belongs to the surface web, that is can be indexed by the standard search engines. The deep web cannot be "seen" by these search engines – these pages simply do not exist until they are created dynamically as the result of a specific search. The type of content that is under the surface can be classified in:

Dynamic content: dynamic pages which are returned in response to a submitted query or accessed only through a form, especially if open-domain input elements (such as text fields) are used; such fields are hard to navigate without domain knowledge.

Unlinked content: pages which are not linked to by other pages, which may prevent Web crawling programs from accessing the content. This content is referred to as pages without backlinks (or inlinks).

Private Web: sites that require registration and login (password-protected resources).

Contextual Web: pages with content varying for different access contexts (e.g., ranges of client IP addresses or previous navigation sequence).

Limited access content: sites that limit access to their pages in a technical way (e.g., using the Robots Exclusion Standard, CAPTCHAs, or no-cache Pragma HTTP headers which prohibit search engines from browsing them and creating cached copies).

Scripted content: pages that are only accessible through links produced by JavaScript as well as content dynamically downloaded from Web servers via Flash or AJAX solutions.

Non-HTML/text content: textual content encoded in multimedia (image or video) files or specific file formats not handled by search engines.

Table 17 The deep web

The automated extraction of data behind form interfaces is a precondition to have automated agents search for desired information, when we wish to wrap a site for higher level queries, and when we wish to extract and integrate information from different sites.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

The question here is firstly how this information can be automatically and systematically accessed (users should be supported in finding online databases useful for their queries) and then how to make the deep web uniformly usable (users shall be assisted in querying on-line databases). The result is thinking of the deep Web as a large collection of queryable databases whose information should be integrated (also with the one available in the shallow web).

Since sources are proliferating and evolving on the Web, they cannot be statically configured for integration, so this integration is intrinsically dynamic and the dynamic discovery of new sources is a requirement. On the other hand, queries are formulated by users for different purposes, which require an ad-hoc integration disabling any possibility of configured per-source knowledge.

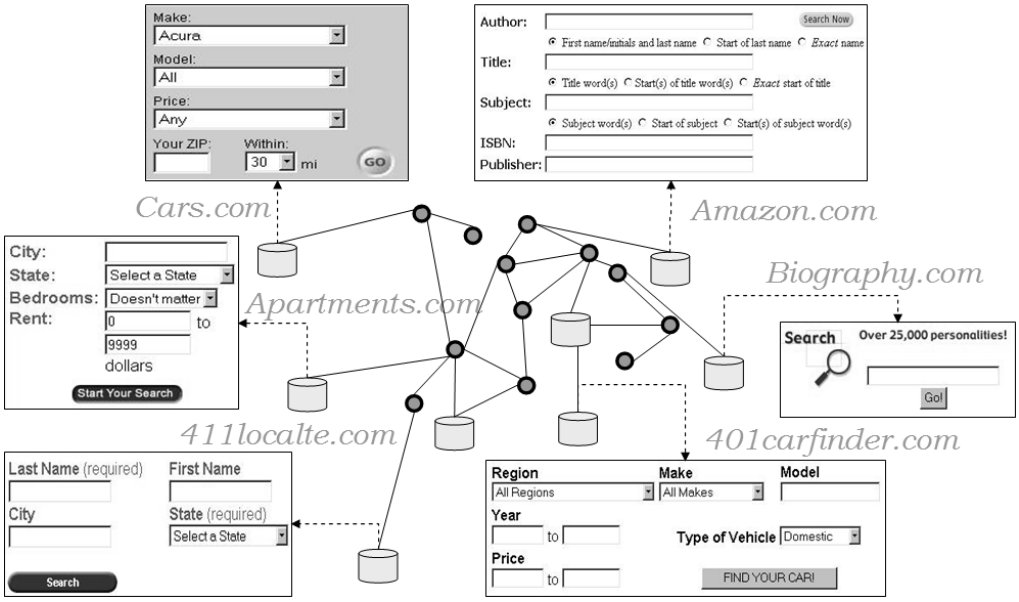


Figure 19 The conceptual view of the deep web [64]

Many researchers have tried to define a method to create a global web query interface that shall fulfill following requirements [64]: *conciseness* or the capability of semantically combining similar fields over source interfaces, *completeness* or retaining source-specific fields, and *user-friendliness* or placing highly related fields together.

If we considered the input fields as schema attributes, we could see the web querying problem as a schema matching scenario. As we already explained, the identification of synonym attributes and other relations between terms in an application domain is usually required to enable the matching.

XVIII. SEMANTIC WEB ENABLING TECHNOLOGIES

A. Semantic web technology stack

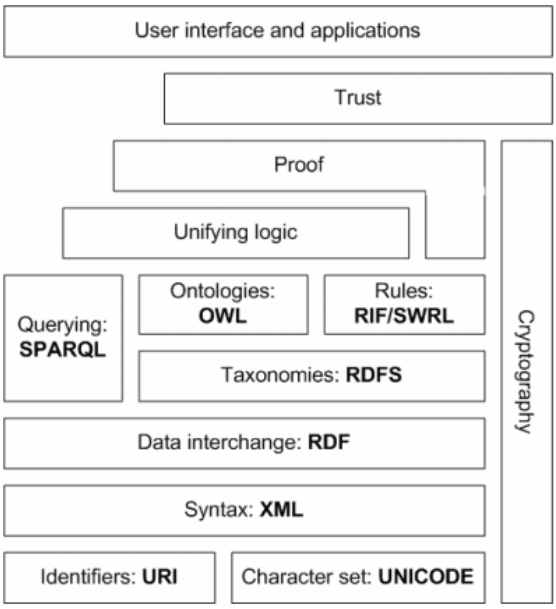


Figure 20 The Semantic Web Technology Stack

1) Unicode

Unicode provides a common representation and technical encoding for text in any language (which is very important, because people communicate using different kinds of alphabets around the world and all of them need to be represented). It ensures compatibility between the text and all types of software. UTF-8 (multi-byte) and UTF-16 are very common Unicode formats [42].

2) Uniform Resource Identifier

The URI provides the address for how to find any kind of Web resource and is therefore is the foundation of the World Wide Web. A URI may consist of a name and/or a locator. URIs are the basis for finding Web pages inside browsers and linking RDF data objects across the vast expanse of the Internet [44].

3) XML

The eXtensible Markup Language (XML) is a language for marking documents and messages with tags that can make it simpler for machines to parse data from files.

The first versions of the RDF and OWL specifications were encoded exclusively in XML, but new alternative formats appeared like N3, Turtle, and N-Triples. Thus, XML is crucial as semantic technologies building block [43].

4) Resource Description Framework and RDF Schema

Are mature data formats that truly serve as the central defining feature of the Semantic Web. RDF is the core model semantics for an open and extensible graph data model of interconnected data items linked by URIs. Whereas the RDF schema provides the core model semantics for describing simple class taxonomies (concepts) that group the RDF data into more complex sets that can be organized and queried via different query languages [45].

5) Ontology Web Language

OWL brings an advanced, computationally stable way of defining highly complex and interdependent data models in the Semantic Web. OWL adds data modelling semantics that are more powerful than conventional databases, but maintains their essential reliability and correctness guarantees that make them so valuable for software applications.

OWL is what gives the Semantic Web an element of grounding and stability for defining the meaning of data in an unambiguous yet powerful data model that rests upon a strong mathematical foundation [46].

6) Simple Protocol and RDF Query Language

Even if the critics believe that W3C should have leveraged the work already put into XQuery or SQL, SPARQL is the standard RDF query language and is being developed to enable the fully compatibility with OWL [47].

7) Rule Interchange Format

The RIF set out to define a standard format for the exchange of business rules between various kinds of software engines. The RIF Working Group has since decided to develop a family of languages aimed at solving specific kinds of problems because the complexity of defining a single technical language for all types of business rules became undesirable.

RIF describes a number of dialects, initially including a Basic Logic Dialect (BLD) and Production Rule Dialect (PRD) [48].

8) Unifying Logic Layer

The aim is to provide a single interface to the semantic web data and rules, encapsulating individual complexities and enabling the applications writing against this single façade.

There is a theoretical aim behind this unifying layer, which is the creation of a formal mathematical logic that reconciles all the different model semantics of the parts (RDF, RDFS, OWL, SPARQL, and RIF) into a consistent and holistic model theory.

Due to the lack of well-defined implementation requirements and documented details around the technical implementation of this layer in the practical senses, there are a lot of frameworks that have implemented their own unifying layer. The result is portability in terms of RDF and OWL, but lack of portability at application level.

9) Proof, Trust and cryptography

The intent of the "proof" element is intended to supply a mathematically correct way of explaining which inferences and which business rules have led to a particular conclusion or recommendation, or in other words, a means for humans to validate what the software has inferred.

The "trust" element enables the rating in terms of trustworthiness (e.g.: data that is likely to be good can be distinguished from data that is more likely to be bad)

Cryptography is built upon the encryption techniques defined for lower layers of the stack like Unicode and XML.

B. Technologies to enrich existing documents with semantics

One of the key challenges towards the adoption of the semantic web is the coding of semantic information with the existing mark-up languages. There have been several approaches:

1) Microformats

They are an approach to semantic markup that seeks to re-use existing XHTML and HTML tags to convey metadata and other attributes. This approach allows information intended for end-users (such as contact information, geographic coordinates, calendar events, and the like) to also be automatically processed by software [54]

Microformats emerged as part of a grassroots movement to make recognizable data items (such as events, contact details or geographical locations) capable of automated processing by software, as well as directly readable by end-users.

XHTML and HTML standards allow for semantics to be embedded and encoded within the attributes of markup tags. Microformats take advantage of these standards by indicating the presence of metadata using the following attributes: class, rel and rev. For example:

```
<div class="vcard">
   <div class="fn">Juan Bernabé Moreno</div>
   <div class="org">University of Granada</div>
   <div class="tel">+34 958 50 00 00</div>
   <a class="url"
href="http://tomywebpage.com/">http://tomywebpage.com/</a>
 </div>
```

Table 18 Example of microformat vcard

Here, the formatted name (fn), organization (org), telephone number (tel) and web address (url) have been identified using specific class names and the whole thing is wrapped in class="vcard", which indicates that the other classes form an hCard (short for "HTML vCard") and are not merely coincidentally named. Other, optional, hCard classes also exist. It is now possible for software, such as browser plug-ins, to extract the information, and transfer it to other applications, such as an address book.

RDF even being a framework and not a format specification is often compared with microformats (see Table 19), as both have similar purpose. Microformats are in reality a bottom-up approach that might not be generic enough, but for now allow people to encode metadata about the data into Web pages, which is good for the semantic awareness.

	RDF	**Microformats**
Designed for	Machines first, humans second	Humans first, machines second
Scope	General representation of metadata	Specific problems
Nature	Is a language	Is not a new language
Extensibility	Infinitely and open-ended	Restricted
Resource representation	URIs allowing for metadata remote access	HTML structures
Semantic web stack	OWL builds upon	Just a hook
Adoption ease	Requires a lot of effort. Might require a behavioural change a tools rewriting	Highly correlated with semantic XHTML
Reliance on pre-defined formats	No limitation	Limited by the types of data that can be encoded

Table 19 RDF vs Microformats

2) RDFa

The RDF for attributes is a set of XHTML extensions that allow for including metadata in any XML document, primarily a web page written in XHTML.

The set of defined attributes for RDFa can be found in the Table 20. Unlike microformats, the RDFa uses namespaces and is therefore highly extendable.

Attribute	**Usage**
About	URI or CURIE specifying the resource the metadata is about; in its absence it defaults to the current document
rel and rev	specifying a relationship or reverse-relationship with another resource
href, src and resource	specifying the partner resource
Property	specifying a property for the content of an element
Content	attribute that overrides the content of the element when using the property attribute
Datatype	optional attribute that specifies the datatype of text specified for use with the property attribute
Typeof	optional attribute that specifies the RDF type(s) of the subject (the resource that the metadata is about)

Table 20 RDFa attributes usage

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
    "http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    version="XHTML+RDFa 1.0" xml:lang="en">
  <head>
    <title> Juan's Home Page </title>
    <base href="http://myexample.net/juan-b/" />
    <meta property="dc:creator" content="Juan Bernabe" />
  </head>
<body>
    <h1>Juan's Home Page</h1>
    <p>My name is <span property="foaf:nick">Juan Bernabé</span> and I
```

```
like
      <a href="http://www.thesimspns.com/" rel="foaf:interest"></a>.
    </p>
    <p>
      My <span rel="foaf:interest"
resource="urn:ISBN:0752820907">favorite
      book</span> is <span about="urn:ISBN: 0062502182"><cite
      property="dc:title">The Alchemist</cite> by
      <span property="dc:creator">Paulo Coelho</span></span>
    </p>
  </body>
</html>
```

Table 21 Example of RDFa

3) eRDF

The embeddable RDF is similar to RDFa and mostly interchangeable, although eRDF is meant only to be used in XHTML documents, whereas RDFa cab be used in all XML-compliant documents. Basically, it is a syntax for writing HTML in such a way that the information in can be extracted by using a parser or an XML Stylesheet) into RDF.

4) GRDDL

GRDDL is a markup format for Gleaning Resource Descriptions from Dialects of Languages. It is a W3C Recommendation, and enables users to obtain RDF triples out of XML documents, including XHTML. The GRDDL specification shows examples using XSLT, however it was intended to be abstract enough to allow for other implementations as well. It became a Recommendation on September 11, 2007

5) RDF Extraction

The variety of formats and specifications mentioned above pursuit the same purpose, namely enabling applications to find data within the unstructured text in a web document.

The most popular and generic way of extracting RDF form i.e. GRDDL is writing a XSLT but it can be tedious. Additionally, you can find a variety of parsers available on the Web to also extract from RDFa and eRDF documents.

The next natural step following the RDF extracting is typically storing it into a RDF database so that more advanced queries and information analysis can be performed.

The fact that dumb web documents hosted in web servers begin containing structured atomic data items brings the data mashups to the next level… Data is there, can be found and extracted and can be combined in any kind of mashups.

XIX. ALL ABOUT (META)DATA

Metadata is data about data. A very rough, yet useful and clear definition of metadata points it as a simply way to enrich data so that software systems can interact with information. Metadata enables that data preserves its meaning outside of its original context.

The fact that the semantic web allows for much more metadata expressiveness on the structure of data makes the semantic web database knowledge bases, unlike relational databases. And that enables enhanced reasoning capabilities and inference on data structures

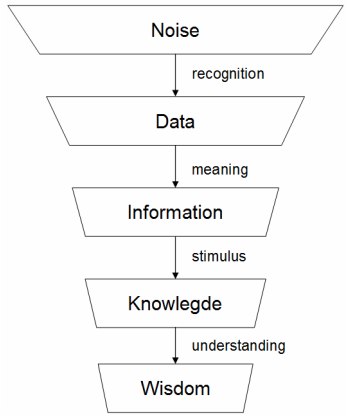


Figure 21 The knowledge funnel

The Table 22 shows the different levels of metadata sorted by increasing complexity.

Metadata Level	**Explanation**
Instant data and records	- Has no data type associated - Exists outside any particular software - Represents an innate irreducible fact associated with its value
Syntactic data	- Defines "how to say things so that the software compiler can understand" Enables the writing of software programs, as it abstracts the human programmer from the machine code - It is the way in which program variables and literals are defined - In the case of markup languages, it is usually called tagging
Structural metadata	- Defines "what can be said" Puts in place a governing schema to enforce a prescribed structure of the document so that it can be validated and checked for consistency against (e.g.: XML documents and XSD, or a relational DB and the ANSI SQL query standards) - Is generally speaking "data about data" - For data models, the governing schema can be one of these: ▪ XML: data is organized in hierarchies, like tree's limbs ▪ Relational: data is organized in tables, like a spreadsheet

	▪ Object: data is organized in a software main memory space ▪ Graph: data is organized in a network where any item can link to another The structure determines which algorithms can navigate it and the limitations associated (e.g.: graphs cannot be indexed and the query answering takes much longer than relational data)
Referent metadata	- A referent is basically an object, action, state, relationship or attribute that defines a relationship to anything real or imaginary (in other words, the links between objects and instances) - Referred to a particular model, the referent metadata is the set of allowable relationships that may exist between objects and instances (e.g..: typical relationship types are Unions, Intersections, Disjointedness, Equivalence, Aggregation, Composition, Hierarchy, Inheritance, etc) - In the particular case of the Semantic Web languages (OWL, RDF, etc), the reference properties are first-class objects in the model, that can be inherited and modeled just like other objects - Referent metadata also describes how objects and instances are related across different schemas and domains to serve to integration purposes (e.g.: point2point mapping, etc)
Domain metadata	- Puts the structural and metadata in context Plays a decisive role in the cross-system data exchange as helps understanding how foreign data fits into local data models. - The most commonly used models are undergoing an standardization process

Table 22 Levels of metadata

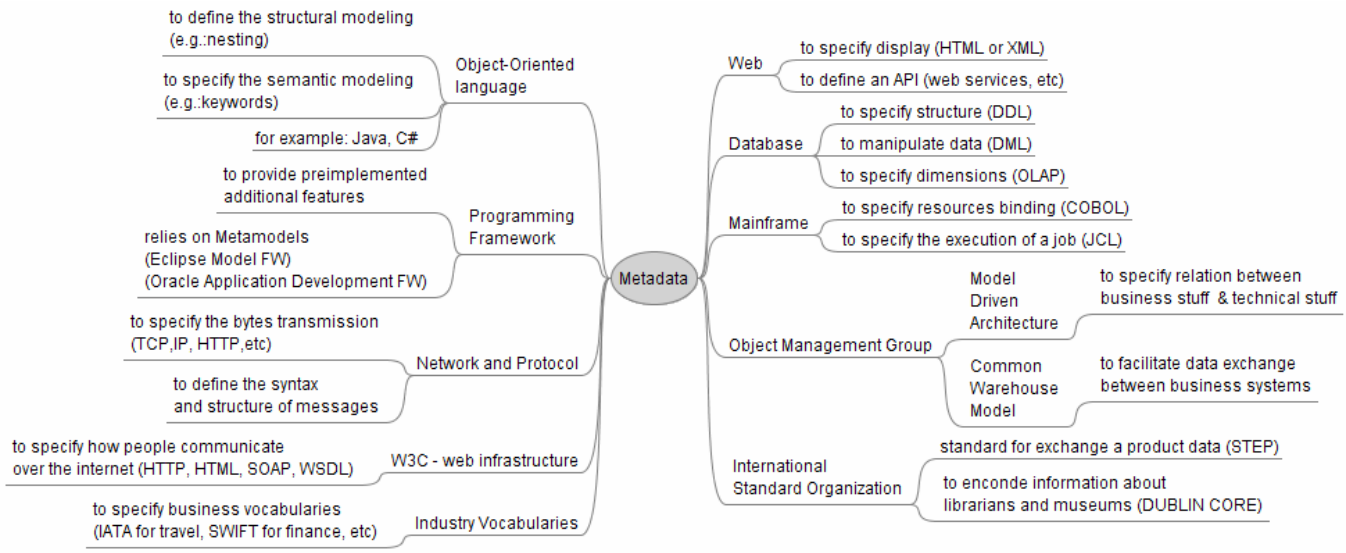


Figure 22 Different usage of metadata

A. The semantic web as superset of metadata

The core issue around metadata is not its purpose, but its variety and its quantity… There is simple too much and the reusability is too little.

The semantic web is just yet another data modeling specification, but with a set of attributes that makes it solid candidate to become the panacea among the metadata and modeling formats:

- Graph modeling with first-class relationship properties
- Open-world assumption (assuming that all answers are possible and then tries to find data that supports or refutes the query)
- High expressivity
- Inference logic
- Decidability (or ability to say for certain whether some bit of data belongs to a set)

But what is a superset metadata language good for? Having a look at Figure 22 it's clear that metadata is present all over the software applications and the metadata formats are incompatible. Wouldn't be great just having a superset of metadata applications can be built against, without having to deal with integration and migration issues? The Semantic Web as a superset metadata format for all data modeling may just be on the horizon for us all: efforts underway since the early 2000s have aimed at focused areas to map existing metadata structures into the Semantic Web's RDF and OWL formats. Numerous standards are migrating toward RDF and OWL support. There are tools and utilities to generate object-oriented metadata, even program code such as Java, from the Semantic Web's RDF and OWL formats. There are numerous utilities to generate RDF and OWL from relational databases, and vice versa. Semantic Web vocabularies are already commonplace in some industries such as life sciences, healthcare, and defense. So, while many people are still debating whether it's even possible, others are going out and doing it [49]

XX. THE EVOLUTION: WEB 1.0, WEB 2.0 AND FINALLY WEB 3.0

In the beginning the world was full of genetic code, which after centuries formed into vessels of information. The basic unit of biological information was known as the gene. It could store and transmit data. It could also duplicate, what gave place to interpretation and mutation. Mutation led to specialization. Genes that were better at performing certain tasks started working together, collaborating. The first organisms were the results of these beneficial partnerships.

The world can be seen as a compendium of loose ideas. Ideas about technology, art, catch phrases, beliefs, fads, earworms, etc. Ideas that propagate from one mind to another… The basic cultural information unit can be called the meme[4]. Memeplexes, or aggregations of memes, form the basis of beliefs, trends, social eras, etc, that spread by means of syndication, feedback, social groups, word of mouth, etc… The internet is the greatest big meme machine.

If we go beyond with this analogy, we can also speak of a natural selection of species à la Darwin, who stated that nature only selects those organisms that are best-suited to the particular environmental condition. Memes show similar behavior, some of them are merely a flash in the pan, others never even get so far as a flash, whereas a few succeed in changing the fabric of our internet culture.

[4] Richard Dawkins coined the term "meme" in his *The Selfish Gene* in 1989: "We need a name for the new replicator, a noun that conveys the idea of a unit of cultural transmission, or a unit of imitation. 'Mimeme' comes from a suitable Greek root, but I want a monosyllable that sounds a bit like 'gene'. I hope my classicist friends will forgive me if I abbreviate mimeme to meme. If it is any consolation, it could alternatively be thought of as being related to 'memory', or to the French word meme. It should be pronounced to rhyme with 'cream'."

A. The web 1.0 or the web as information portal

At the very beginning, the web was a compendium of documents, mostly written in a markup language that could be accessed by URL, by clicking on links on already found documents, or retrieving them with a search engine. The information was exclusive… Only the webmaster owned it as content owner. The WWW was divided into usable directories where everyone had their personal own little corner in the cyberspace.

The main drawbacks were lack of context, reduced to inexistent interaction and low scalability

B. The web 2.0 or the web as a platform

The community got the focus… they are in foreground to create and validate content. The organization shift took place from directories to a freer form called "tags". Integration capabilities were added by means of hooks, like the definition of RSS Feeds and content syndication and APIs

In this scenario, personalization was still missing, the interoperability quite limited and based on time consuming workarounds and the true portability was lacking.

The web 2.0 is not a new form of the existing web… actually it is not about the web itself, but about the way people use web 1.0. The shift consists of moving from a static fixed context serving modus to an interactive place people congregate and do things together.

Collective intelligence is leverage by massive collaboration and interaction. Even the vocabularies are evolving in a much more natural way based on tagging projects that establish hierarchies of tags building folksonomies.

Content re-usage via mashups is unprecendently easy and makes the internet a space where collaboration-based creativity has emerged like never before.

From the technological point of view, there has been only a very little change –new programming languages have appeared or been consolidated, like Ruby on Rails or Flash, hooks from the past have been standardized, like AJAX or JASON, and XML is used for almost everything-, but the web remains driven by document and pages. Therefore, Web 2.0 is rather an evolution of the way people use the web.

The peak of web 2.0 has sharpened the limitations of the original document based web, because data cannot be reused and is isolated in information silos.

C. The web 3.0 or the web as a meeting place for humans and machines

The web 3.0 relies on following memes:

- Semantic enabled web, aiming at changing the web by introducing a language that can be read and processed by software rather than humans.
- Personalization, or contextualization of the web based on the way people are using it.
- Artificial Intelligence or how to extract meaning from the way people interact with the web.
- Mobility or how everything can be accessed from everywhere at any time

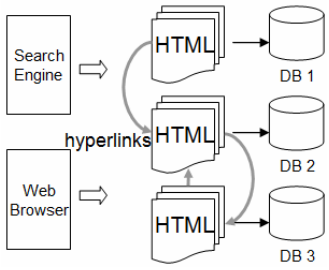
	Web 1.0	**Web 2.0**	**Web 3.0**
Slogan	"the mostly read-only web"	"the wildly read-write web"	"the portable personal web"
Focus	On companies	On communities	On the individual

Basic online unit	Homepage	Blogs	Livestream
Users main activity	Owning content	Sharing content	Consolidating dynamic content
Information compendium	Britannica Online	Wikipedia	DBpedia
Core technologies	HTML, portals	XML, RSS	Semantic web technologies
Web building blocks	Web forms	Web applications	Widgets, drag&drop mashups
Information organization	Directories (taxonomy)	Tagging (folksonomy)	User behavior ("me-onomy")
Reference application	Netscape	Google	iGoogle, NetVibes
Ecommerce units	Pages views	Cost per click	User engagement
Advertising form	Advertising	Word of mouth, viral, rich media	Advertainment

Table 23 'Wevolution', inspired by Marta Strickland

D. Towards the web of data

The semantic web is about getting the data out of the information silos and enabling the qualified interlinking of them. There have been attempts in the context of Web 2.0 to enable web applications consuming data or functionality from heterogeneous data sources, what we know as a mashup. As we pointed out before, the Web 2.0 has been a social evolution rather than a technical transformation. Thus, the ambitious approach of seeing Internet a big data base of open data can be only achieved with some technological advances brought by the semantic web technologies

	Pros	**Cons**
Document based web 	- Status quo - No complexity added	- Content not well structured - Difficult to retrieve combined information and to express combined queries - Application cannot process the content
API-based mashups	- APIs expose structured data that	- Proprietary interfaces

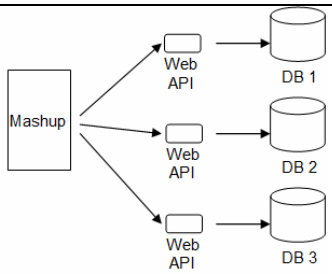
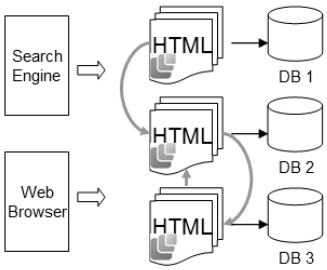
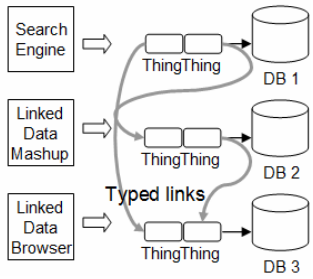
	can be easily consumed - APIs enable combining the structured information to create new applications	- Mush-ups based on fixed datasets - You cannot set links between the objects (neither within a dataset nor between objects located in 2 different data sets)
Semantic Microformats 	- Embed structured data into existing HTML - hCard, hCalendar, xFN, etc - Easy to adopt	- It doesn't allow for interconnecting data items - Only a small and fixed set of microformats exist
Opened linked data 	- Builds on standards in contrast to Web APIs - Enables applications that work against an unbound set of data sources - Enables incorporating new data sources as they become available on the web - Interlinking enables new content discovery - Information retrieval based on SPARQL	- Not all data are available for the big audience - Need for compliance with standards - SPARQL still far away from SQL

Table 24 Web of Data steps of maturity

XXI. SEMANTIC WEB SERVICES

Today, the web services present several problems that semantic web services strive for solving.

The interfaces of the web services nowadays are designed to enable the communication with a single program or with the same program.

Commonly it is about web browsers that don't process the information they get, but just display it enclosed within HTML tags.

The web browser knows nothing about the information between <H1></H1>, but only that this information should be displayed with a font-size higher.

Many services in the web have the problem that they can't be combined. A service can invoke another service only after big programming efforts (e.g.: parsing the output of the other service for a given token and applying some scraping techniques). The

fact that almost all web services in the web have different output and this can vary along the time makes the integration a short of Sisyphus work.

On one hand we have the different standards that are developed by different companies to serve different purposes, which avoid the existence of uniform interfaces. On the other hand, a web service can't describe itself to other programs and therefore, can't be easily found (discovered) without the intervention of a human user.

All together should work in the imminent future of the web services with the support of semantics.

Using web services with standardized interfaces should save a lot of time and money in the development of web-based applications. Semantic web services provides Internet with additional automatisms, which enables that one application running on the home computer autonomously searches, selects and invokes web services and even more, combines them to achieve a higher level functionality.

This is the dream of software agents come true.

A. Laying the fundamentals

According to W3C a software system designed to support interoperable machine-to-machine interaction over a network. Following conditions are the task for each entity:

- Providers must describe the capabilities and constraints on offered services
- Requestors must create abstract characterizations of required services to facilitate matching with published capabilities.
- Requestors must locate and interact with peers or matchmakers that can respond to queries for advertised service descriptions.
- Matchmakers must compare descriptions of queries and capabilities.
- Requestors must decide if they can satisfy the preconditions specified in a prospective service's self-description in order to use it.

B. What are Semantic Web Services good for?

The goal of the semantic web services is the increase of automation in following web services processes:

- Automatic discovery of web services: the search and discovery of web services providing a given functionality in a Service Registry should be performed automatically, which is only possible by means of semantics

- Automatic invocation of web services: here are web services meant, that consists of several method calls (non-atomic) (e.g.: buying a CD in internet implies searching for it, selecting it, adding it to the shopping cart and paying for it). Without the support of semantics, this couldn't be possible

- Automatic composition of web services: should a service not be able of fulfilling the user requirements, then a composition of other web services is automatically created and performed to fulfill these requirements.

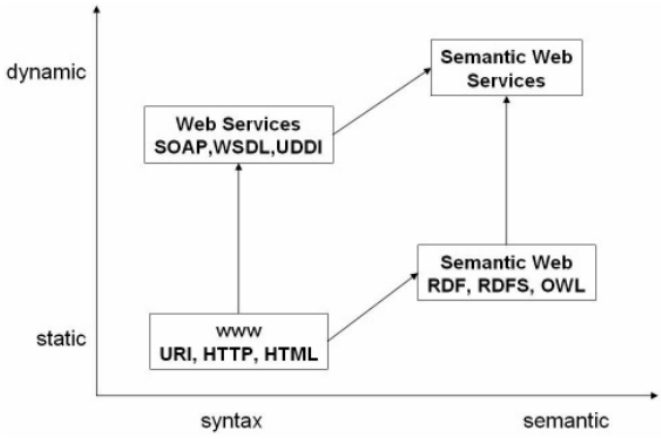


Figure 23 Evolution to Semantic Web Services

C. Bringing semantics to the Web Services

Depending on their starting point of the approach, we distinguish top-down approaches, that model the web services and their semantics independently on existing web services technologies. This modeling is accomplished according to an ontology language, which aims at providing the best means to describe the web service.

By means of the so called "grounding", where the mapping of semantic description elements to WSDL elements is specified, the relationship to the WSDL is established. There are two main top-down approaches: the OWL-S and the WSMO.
On the other hand we have the so called bottom-up approach, which pursuit the semantic enrichment of existent technologies (like especially WSDL and BPEL).

The initiative the W3C organization is fostering the most represents the evolutionary and less ambitious approach, just because it relies on extending already existing components with semantic capabilities to overcome their limitations.

The Web Services Definition Language (WSDL) is an extensible, platform independent XML language for "describing" services. It provides mainly functional information about the service: IDL description, access protocol and deployment details, etc… in general, all of the functional information needed to programmatically access a service, contained within a machine-readable format.

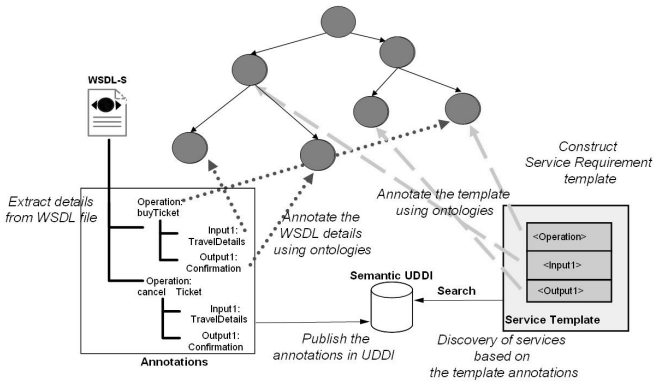
WSDL does not include QoS, Taxonomies or Business information.

1) WSDL-S

Its key success factors are:

- Ease in adoption: as this approach is simple, light-weight and upwardly compatible with the existing WSDL standard
- Semantic representation language independency, which allows for the re-usage of domain models, the flexibility of modeling language choice and the annotation with multiple ontologies
- Ease in tool upgrades (e.g. wsif / axis invocation)

Even other more revolutionary approaches to the semantic web services are pursued leveraging the existing WSDL and XML schemas for business documents and the set of tools to exploit them is and will be critical. WSDL-S already positioned itself as the best candidate to bridge the gap between those revolutionary approaches and the existing WSDL



2) WSMO

Web Service Modeling Ontology is a conceptual model for the web services description, in other words, one semantic web services core elements ontology

Similarly to other initiatives, the ultimate goal of WSMO is enabling the automatic service discovery and their execution, as well as paving the way to a holistic yet simple integration solution. This will allow for the automatic cooperation of non-dependant services to achieve a common functionality at a higher level.

The four pillars of WSMO are:

- Ontologies, that define the language space for all WSMO elements
- Web Services, that provide access to certain functionalities
- Goals set by the clients that invoke the web services
- Mediators that are responsible for the interoperability between the WSMO elements

3) OWL-S

The Web Ontology Language for Services (OWL-S) is created upon the DAML-S, which is based on DAML+OIL (currently in Version 1.1).

The goal of the developers of OWL-S is the connection of Web Services and the semantic web to end up providing the Semantic web services.

To achieve this goal, OWL-S should provide following functionality:

- Automatic Web Service Discovery
- Automatic Web Service Invocation by a client or software agent. The execution is seen as a sequence of functional invocations and requires that the Software-Agent recognizes the interface semantic of the to-be-called WS
- Automatic Web Service composition and interoperation: given a request, the selection, composition and cooperation of web services to fulfill this request.

To provide the mentioned functionality, OWL-S is based on technologies already in place for Web Services (like SOAP and WSDL), adding types and classes to them, with the purpose of describing the web services functionality in a machine understable way (covering not only the control and data flow, but also preconditions and effects).

Semantic Web Services (SWS) in OWL-S are described by four ontologies: “Service“, ”ServiceProfile“, ”ServiceModel“ and ”ServiceGrounding“

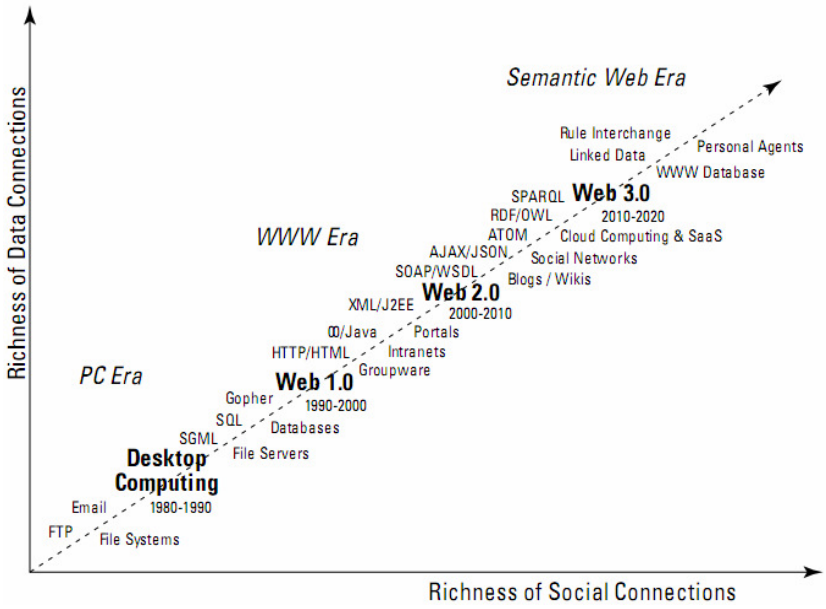


Figure 24 Nova Spivak's vision of the semantic development

PART IV: THE SEMANTIC WEB MEETS CI

XXII. INTRODUCTION

As already mentioned in this work, knowledge and intelligence are tightly connected… The semantic Web aims at creating the ability to capture knowledge in machine understandable form, to publish that knowledge online, to develop agents that can integrate that knowledge and reason about it, and to communicate the results both to people and to other agents, will do nothing short of revolutionize the way people disseminate and utilize information [3]

We will approach the assessment of the advantages that the semantic web brings to the Competitive Intelligence word according to a model that takes into account all dimensions of interests (as shown in the Figure 25)

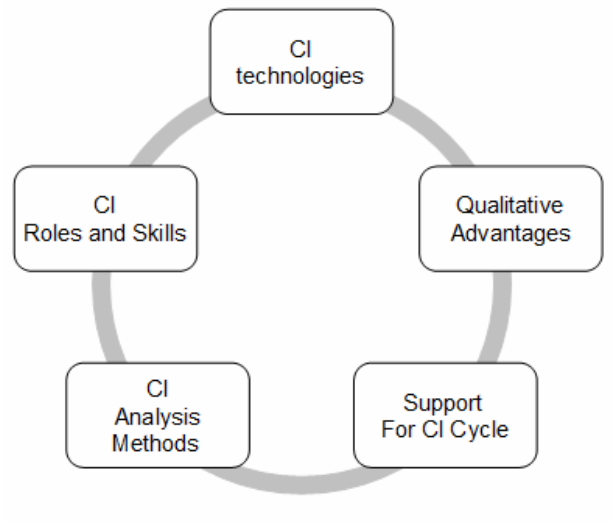


Figure 25 Competitive Intelligence dimensions of interest

XXIII. ROLES AND SKILLS IN THE COMPETITIVE INTELLIGENCE TEAM

In this section we will explain the typical composition of a Competitive Intelligence Unit, with a rough description on the activities and the skills required by each role.

A. Roles required to perform CI Activities

1) Intelligence requestor

The intelligence needs required to be expressed. The intelligence requestor is the trigger of the entire intelligence process. Behind a formulated intelligence need, there is usually a strategic decision to be taken, or a question about the company in comparison to their competitors, etc.

The role of requestor is commonly played by executives, heads of departments, C-level, etc. It is all about gaining intelligence about a particular business and the inquiry comes therefore from the people in charge of driving it.

2) Archivist

The archivist is in charge of assessing, organizing, preserving, maintaining control over, and providing access to information determined to have long-term value.

They are required to have research skills (especially when cataloguing record or when assisting users), be able to preserve both paper and electronic records, have logical and analytical thinking to organize the information to be archived in a way that is easily accessible, etc

3) Librarian (or cataloguers)

They are specialized in organizing information of different complexity. Specialization in a particular industry or domain can be beneficial, but is not a must. The modern librarian is expected to develop search strategies that can be applied to several search engines, online catalogues, etc.

The core competencies of a librarian consist of locating, enriching, organizing and disseminating data. They work closely with taxonomist and sometimes they may assume the role of the taxonomist itself by creating taxonomies, but in principle, they are not intended to. Actually, they work with predefines tags/taxonomies to manually classify information for further enrichment and distribution.

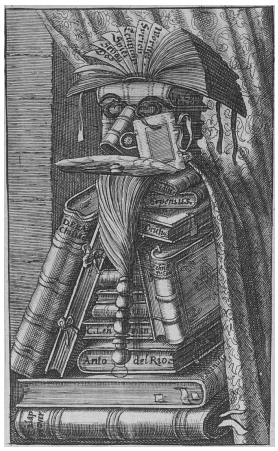


Figure 26 The librarian

4) Taxonomists

The taxonomist is the person responsible for defining the category system and tags, which requires a much more technical background an average librarian would posses. The category and tagging systems are usually part of a bigger systems landscape where taxonomies are consumed by automated software programs and may be maintained in technical formats like XML documents and indexed master files. It's a taxonomist' responsibility to specify and maintain complex taxonomies with IT dependencies that require a deeper technical understanding of code syntax and programming skills in order to produce technically valid IT inputs.

The close cooperation with the business analyst is crucial to get the information needs and requirements expressed in the appropriate IT requirements for the many uses of that corporate information and reference data.

The working artefacts of a taxonomist are model hierarchies, tag sets, file lists, master files, property files, some relational data models or indices, etc. In order to organize content that varies from fully unstructured to highly structured, the taxonomist produces classification systems that are applied to the available data.

5) Information architects

They are experts in the underlying technologies and system that supply the lifeblood of data throughout a large business. This role doesn't require industry or domain specific knowledge and it is located within the IT to create new data formats that comply with the existing technology stack. Sometimes this role is also known as software architect or system architect, as they are experts in the software systems that feed and are fed by the information management applications.

This role to be successful requires close cooperation with the business analysts, taxonomists, and DBAs working on implementation design and construction.

6) Database architects

They are highly focused on non-database models and more focus on performance optimizations for relational databases and work directly with the business analysts and taxonomists to understand the system requirement. They produce a database data #model that can match those business needs with other non-functional requirements.

A key success factor for database architecture is their ability to see the big picture from a business perspective and understand the database technology is merely an enabler.

7) Information scouts

The information scout has the required skills to perform the information gathering discovering new sources and selecting the ones out of the known repository that better fulfils the information requirements.

They can be working with librarians to access currently available information, but also need to specialize in finding information sources that are not known at the searching time. Thus, this role requires a very high technical affinity even to retrieve information from the deep web.

Once the documents containing the requested information have been retrieved, checking the quality of information or integrating despair sources are typical task of the information scout.

8) Concurrence Monitors

This role relates to those that are constantly monitoring the sources that address general or specific aspects about the competitors. They are usually familiar with internet monitoring systems and retrieve information according to the corporate guidelines (i.e.: using the terminology defined by the taxonomist, etc)

9) Data governors

This role guarantees the quality of the data by ensuring its compliance with corporate models, taxonomies, etc, by eliminating inconsistencies, redundancies, etc. Unlike the other roles commented so far, the data governors operate directly on the data and are held responsible for ensuring good data.

10) Business analysts

The business analyst is an expert in a particular domain or industry, but without a horizontal skills set that makes him transferable to different domains.

The business analyst advises the IT team on behalf of the business and sets objectives for the management and dissemination of high-quality and reliable business information. Usually takes care of translating the intelligence needs into information requirements, so that the entire CI process can start.

The analytical method is usually determined by the business analyst depending on the intelligence needs

B. How the CI required roles benefits from semantic technologies

1) An emerging role: from Taxonomist to Ontologist

This is a new emerging discipline of information modelling, structured data definitions and description logics

Unlike taxonomists, that are usually skilled in working according to tree structures in a thesaurus-style, the ontologist is skilled in a definitional logic that is much more expressive.

Unlike the information architect, that works with UML and ER models, the ontologist produces a higher-level of modelling experience using formats like OWL, KIF (Knowledge Interchange Format), or SCL (Simple Common Logic).

2) Information scouts and the semantic-enabled search optimization

Competitive Intelligence requires finding the right information at the right time. Unlike a search performed via a search engine like Google, the competitors intelligence depends on very rich and sophisticated taxonomies that guide the Information Scouts towards the right content.

Once the information sources have been gathered and their suitability has been checked, it's determining categorizing the content according to term lists, keywords, data models, etc, so that it can be found with the maximal efficiency. This task that traditionally has performed manually or semi-automatically is a tedious and error-prone. Traditional thesaurus-based techniques with nested taxonomies have been employed to introduce some level of automatism in this process.

With the advances of the semantic techniques, it is possible to leverage Natural Language Processing for documents classification by enhancing semantic web based ontologies seeding the NLP models with more dynamic taxonomies. What traditionally was stored in rigid rational database can now populate RDF database with graph data granting more flexibility and allowing for enhanced navigation. Not only database, but also master files shall profit from semantic techniques… the master file structure that is currently being specified by means of flat word list in text documents can be encoded as proper ontologies with all the additional richness and power of complete business logic for linking word descriptions.

Obviously these possibilities don't seem so revolutionary, but they are gaining in attention and many large corporations are adopting them to make one of their principal assets, the information, faster retrievable and bring the searching experience to its next level.

3) Data governor 2.0 or how to get to higher data quality standards

The gathered data needs to be proved for consistency and accuracy. Bad data can have incommensurable impact on the valuation of many public companies itself. More often, the public doesn't hear about the cases where bad customer data or bad product data cost a company millions.

So far there have been rule-based systems that locate bad data and try to fix it. Even some statistical method can be leverage to cluster the data and report outliers that point to bad data.

The semantic-concept based approach tries first to organize the data according to the concepts that the data is prone to related with, and then by means of consistency rules normalizing that same data. This approach has proven to outperform the others in terms of data quality and cleansing operations when the data domain is complex.

Applying the semantic approach to the data quality problem will make the data governor's life much easier

4) The "Fall of the Wall" between the Information Silos

The information scout sometimes acts like a information broker. Moreover, to answer an intelligence inquiry, part of the required data might already be available within the organization. If we assume that this data is correct, we can start thinking of the next problem, which consists of not being able to get the data that you know is there somewhere. Separated content management systems for intranet and extranet, document management systems –like Documentum, etc- that are available for certain departments, but not globally, a variety of department specific databases, etc… A lot of sources with information that somehow is all related but cannot be retrieved holistically because the information is trapped in silos.

Each silo contains a fragment of the big pictures, might use its own taxonomy, its on searching and retrieval procedures, etc

The Semantic Web technologies can be leveraged to build a kind of enterprise ontology to unify the taxonomies, schemas, etc of different information silos. The richness and flexibility of an ontology language are key success factors in such endeavour.

5) The business intelligence-driven business

Intelligence needs are usually part of an overall strategic decision process. Companies usually rely on business intelligence systems that aid executives with scenario planning, forecasting, visibility into operational systems, analysis of market conditions, etc.

These systems have been typically grown around a relational database and improved to get the most of the rational data over years. That's why many companies are so reluctant to completely shifting to a semantic web based system. Nevertheless, those that require flexible and dynamic integration of big volume of data coming from a large variety of heterogeneous sources have immediately adopted the semantic web data structures to avoid the integration overhead inherent to ER systems. An early adopter has been the pharmaceutical and the bio-sciences industry, where experimental data are published and consumed in real time by researchers all over the world relying on semantic web structures that allow for linkage of data entities across disparate sources. Another good example is the defence industry that uses semantic web data as a place to consume and analyze open-source intelligence gleaned from public sources[5]

For the rest of industries, the semantic web systems can be embraced due to their graph-base data representation to enable the extension of classical ER models to uptake unstructured documents, and thus, improving the classical business intelligence systems.

In general, business intelligence systems that deals with data you know very little or nothing a priori about (not even the source, or not even that they exist at the present moment) but require integration with existing data would enormously profit from the intrinsic flexibility, agility and extensibility of the semantic web structures.

The semantic web can also significantly improve the traditional business intelligence by leveraging all variety of underutilized data assets, as shown in Figure 27

[5] As indicated in the historical note at the beginning of this work, the Competitive Intelligence has its origins in military intelligence. If the Defence of the most larger nations relies on semantic web technologies, wouldn't it be appropriate to strive for such a shift also for Competitive Intelligence?

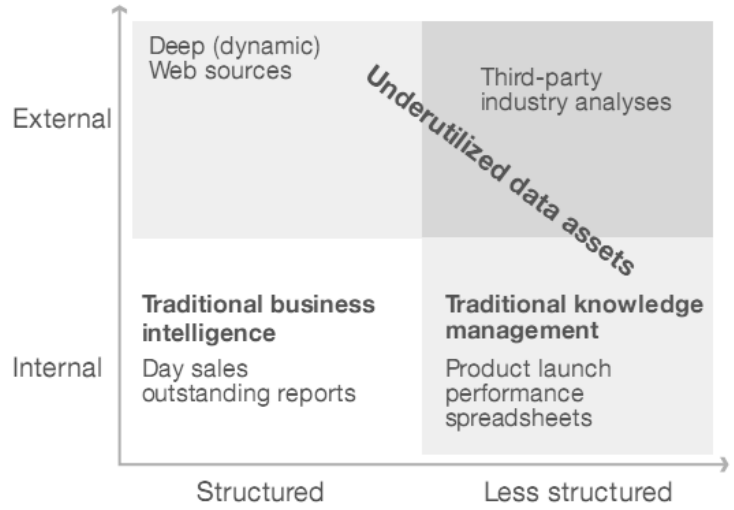


Figure 27 Leveraging Business Intelligence (source PriceWaterhouseCoopers'09)

XXIV. How the Semantic Web can support the CI cycle

If we examine the CI cycle phases, we will discover that the semantic web supports to some extent each and every phase. Actually this enablement happens at technological level, which means that tracking the technologies being used in each CI cycle step, we will get to a benefit.

We revisit the Table 12 CI Technologies along the CI Cycle putting a semantic "scoring" (from S to SSS, being SSS the maximum) according to the semantic potential to leverage a given technology (rows) within a particular sub-process in the CI cycle (columns).

	Needs definition	**Information Gathering**	**Information Organization**	**Analysis**	**Creation of Intelligence Products**	**Dissemination**
Text mining technologies		SSS	SSS			
Text discovering tool		S	S			
Automation Text summarizing				S	S	
IE and IR						
Information retrieval		SSS	SSS			
Information extraction		SSS	SSS			
Analysis and reporting tools						
Statistical packages				S	S	
Analyzing and Reporting suites					S	S
Intelligent agent technology						
Active filtering tools		S	S			
Media monitoring		SSS	SSS			

services						
Information searching, indexing and retrieving						
Web Crawler		SSS	SSS			
Indexer		SSS	SSS			
Query Engine		SSS	SSS			
Interface		SSS	SSS			
Sentiment analysis		SSS	SSS			
Multimedia content		SSS	SSS			
Document and content management						
Document management systems (DMS)	SS	SS			S	
Information aggregators		SSS			SSS	SS
Multipurpose portals		S	S			S
Business Intelligence and e-Business applications				SS		

Table 25 Semantic web support to the CI cycle

XXV. QUALITATIVE ADVANTAGES

Now we suggest leaving for a moment the Competitive Intelligence domain aside and we focus on the net benefits that the semantic web brings to all information professionals:

1) *Lower the number of clicks to get what you want*

Decrease the number of clicks to find the data you need to get what you are searching for (this is already a reality with Yahoo! SearchMonkey)

2) *Avoid repetitive tasks*

Avoid entering over and over the same information (the Blue Glue toolbar leverages your metadata for the same)

Exchange with others in a more effective way: topics are univocally identified and there's no room for ambiguity (ever tried Twine?)

3) *Improve the way you search*

Move away from key word searching... Start requesting information from the web scanning for ideas and concepts (the Calais Web Service from Thomson Reuters enables it for news)

4) *Optimize the time to information*

Lower substantially the time-to-information. Find not only documents, but also people that you need (NASA is already locating people based on their skills with semantic technology)

5) Real time analysis

Get real-time answers to the your analysis (the semantic web is already helping Renault in deciphering complex data for root-cause analysis to cross-reference data and technical problems in real time)

6) Seamless data integration

Leverage a flexible and extensible integration means for any kind of sources and perform queries to jointly retrieve information from them and get answers supported by all sources (like Oracle, the British Telecom and other are already doing)

7) Dig into the deep web

Find easier and link faster the available web services which usually give access to deep web information (IBM is actively researching on the semantic web services area)

8) Leverage what is already there

Perform classical data mining techniques in the web to derive "hidden" intelligence

XXVI. How the semantic web can improve or leverage existing CI technologies

1) Semantic Text summarization

Current search engines and information retrieval systems only perform shallow string processing due to the lack of deep understanding of natural languages and human intelligence, and users usually have to go through pages before they find something useful or give up.

Depending on the domain in which the information request is launched, this latency to get the information might be critical and the time to get the essentials of a document becomes too precious.

That's why the effective text summarization is essential for CI professionals.

Unlike the pre-semantic surface features-based approach that selects summary material from the source based on position information, specific terms or cue phrases, the knowledge-based approach builds a semantic representation for the summarization task, such as a set of logical forms, using ontology knowledge, or a template describing some key concept, etc.

Let's have a detailed look at the approach proposed in [65], which is totally term-based: only terms defined in the general purpose or specific domain ontology are recognized and process, whereas all other terms are ignored. Thus, the original document is represented with a semantically connected concept network.

With a purely term based approach, a subset of sentences form the original document is picked as the summary.

The summarization procedure relies on following steps:

1. The query is refined using general purpose and domain specific ontologies by adding relevant keywords or removing redundant ones. The user is typically ask to provide feedback to the re-fined query
2. The distance from each document's sentence to the re-fined query is computed. If the distance is smaller than a threshold, the sentence is a candidate to be include in the summary
3. The pair-wise distance among the candidate sentences is calculated which allow for clustering the candidate sentences into groups according to the threshold value. The sentences are picked for the summary according to their ranking and to the degree to which the text shall be comprised.

The advantage of this technique relies firstly on the semantic based query enhancement, and secondly on the creation of the semantically connected conceptual network. Although there is room for improvement (like semantic-based redundancy detection, better syntax analysis to improve the quality of the documents, better query handling procedure, combination with NLP techniques, etc), the ontology knowledge enables better results than a mere keyword-based approach.

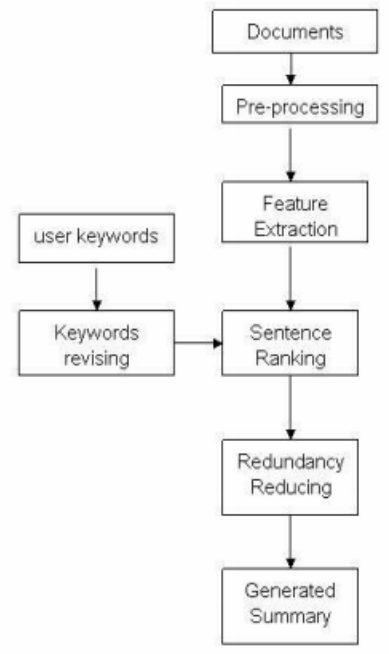


Figure 28 Semantic enabled text summarizing

2) Semantic Information Retrieval

The heterogeneous and distributed nature of information results into two issues that impede the effectively query answering:

- Information semantic heterogeneity and lack of semantic annotations. In such conditions, a search engine can hardly map a query to documents where the requested information is available but not explicitly annotated
- The result of a query may be distributed across several documents that may or may not be explicitly connected with a hyperlink. How can a search engine identify such a distributed result?

The semantic annotations and the ontological support address the first issue, but the internet status quo demonstrates that hardly any web document is explicitly semantic enriched. The second issue can be solved using graph based query models as long as the links are available.

In [66] a Light-Weight semantic web is presented to overcome such issues by employing well-established Information Extraction tools enhanced with additional external knowledge, as discussed before in this work, to find information of certain classes (like person, location, etc). These classes are automatically annotated in the documents found by the search engine by means of type-specific wrappers. This annotation enables the identification of relevant answers to queries as well as the creation of virtual links between documents when an instance of a class is identified to be the same in both documents.

Other approaches pursuit semantic knowledge from collaborative knowledge bases (i.e.: Wikipedia) to improve the effectiveness of information retrieval. In [68] an approach is presented that performs the same in a multilingual way. Basically Wikipedia is leveraged as corpus to derive the conceptual space where the world similarities are calculated (applying different similarity measure depending on the co-occurrence of terms in the same article). The retrieved word clusters are then applied to compute sentence-based document similarity [69]. It can be also modified to represent each term contained in Wikipedia as the concept space as a vector, derived from the term's occurrence in the respective Wikipedia articles. The similarity of two documents is then computed using a centroid-based classifier. The concept vector of each term in the document is weighted with the term's value. From these weighted concept vectors an average vector is calculated which represents the respective document in the concept space. The similarity score of two documents is then computed using the cosine metric.

In [67] both statistical and semantic models are combined to increase the retrieval effectiveness (as both use different types of information represented in queries, documents, and possibly external knowledge a combination of both might deliver better results)

To handle the richness of natural languages used by humans a fuzzy-ontology based approach has been proposed [70]. Expressions of interest like "good Asiatic restaurant, a very sunny resort, relevant patents, etc". The dynamic definition of the dynamic knowledge of a domain, adapting itself to the context needs to be addressed. As a matter of fact, there is a gap between the correct definition of an object (given by the ontology structure) and the actual meaning assigned to the artefact by users (i.e. the experience-based context assumed by every person according to his specific knowledge). By leveraging the semantic correlations among the entities that are searched in a query or when a document has been inserted into a data base, this gap can be bridged. Correlations can be defined by a semantic evaluation of objects stored in a data base, query representations and ontology structure (with fuzzy weights).

Introducing fuzzy ontologies in Information Retrieval enables the exploitation of additional knowledge hidden in the entities-documents relationships and therewith the enrichment of the system semantics.

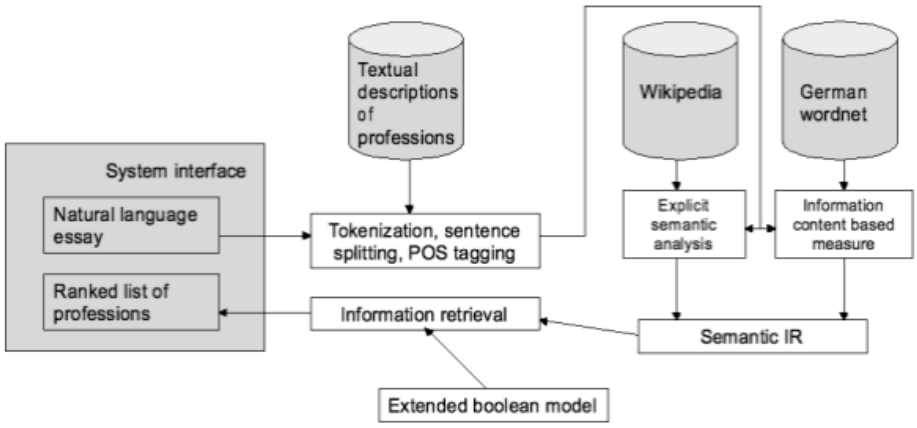


Figure 29 SIR system architecture

3) The ideal Search Engine

As already mentioned in this research, search engines are the be-all and end-all out of the tools for the modern CI practitioner and in general for each and every worker that has to do with information.

The following points will describe the features of an ideal search engine and subsequently will show how far the semantic web can contribute to improve each feature

a) Querying

Spelling correction: in the Search interface, unrecognized words are noted and the user is given a list of alternative spellings o select from. The user may also leave the word as it is. The user should have the choice to specify that the Semantic NLP module search on words if they are not recognized

Linguistic Boolean Search: the searches can be formed using fully recursive Boolean expressions with AND, OR, WITH and AND NOT operators. The expressions connected with the Boolean operators are interpreted for meaning.

Proximity: supports the use of some form of word proximity searching, exclusive of the ability to search on phrases.

Restrictors (search by field):

- domain, language, date, page level, file format, occurrences
- Supports the following prefixes: intitle:,allintitle:,site:,inurl:,filetype:
- putting ~ before a word searches for synonyms
- Search for numeric range by placing two periods between numbers

Truncation: enabled word stemming either automatically or on request

Syntactical Exclusion: terms that are not intended to appear in the results

Semantic Exclusion: a certain contexts related to a query can be excluded (e.g.: searching for “apple” and excluding all electronic devices to retrieve only documents about fruits)

Limits:

- Date (User Specified or Specific Choices)
- Domain
- Containing a media type
- Document Directory Depth
- Page Depth

b) Recall and Precision

Relevance ranking: the search engine returns a list of retrievals containing documents in which the query concepts were found together in a sentence. Those with exact word matches to the query terms come first. The next group is documents in which there were exact matches to some query terms in the body of the document, but other query terms only matched conceptually. Which terms match exactly is indicated. The last group is documents in which there were conceptual matches to the query terms. Within each group, documents are listed according to the number of sentences in which all query terms were found.

Fuzzy Search: the search engine searches can be formed using wildcards and fuzzy operators (e.g., specify matches names that sound like something, or matches words that start with some string and end with another one, etc) such that proper names and other words can be matched approximately.

c) Results presentation

Find Like: the engine provides a quick link on every return to allow finding similar sites.

Specific retrievals highlighted: when the user clicks on the “Highlighted Text” link corresponding to a retrieved document, the search engine should highlight the relevant section.

Specific words highlighted: when the user clicks on the “Highlighted Text” link corresponding to a retrieved document, the search engine colour-codes the specific words which matched the query within the relevant highlight section. As the user hovers the cursor over a matched word, the corresponding query term is indicated. Sometimes a given word may correspond to more than one query term. In this case the word is highlighted according to the first query term matched, but the hove over text indicates all matched terms.

d) Extras

Review and annotation tool: the search engine offers a review tool for reviewing and categorizing document sets. Features include the ability to create project categories and classify documents into those categories, to limit searches within categories, to browse document sets without querying and to export archives of categorized documents. When combined with a relational database, the features provided allow the user to create and modify database tables, display document-related data from tables as part of query results and restrict searches based on table column values. Also included are scripts that allow the user to take notes and have the notes indexed and available to be searched in tandem with the related document set.

e) Crawling and Indexing

Formats: the search engine indexes documents in HTML, XML, OCR'd text and plain ASCII text. Documents in Word, Power Point RTF, or WordPerfect are converted to HTML before being indexed. Some engineering may be required for XML. Documents in PDF are converted to plain text before being indexed. The user may view retrievals in documents converted to HTML or plain text with the specific retrieved sections highlighted, but may also choose to view the original file without highlighting.

Deep Crawl: all crawlers will find pages to add to their web page indexes, even if those pages have never been submitted to them. However, some crawlers are better than others. This section of the chart shows which search engines are likely to do a "deep crawl" and gather many pages from your web site, even if these pages were never submitted. In general, the larger a search engine's index is, the more likely it will list many pages per site. See the Search Engine Sizes page for the latest index sizes at the major search engines.

Customer and meta tags: the search engine can search in customer-defines tags, if desired, with some engineering assistance from Cognition Technologies.

Indexing local directories: the search engine interface permits the user to select individual files or whole directories for indexing.

Spider configuration: the indexing GUI allows specifying a list of seeds, the desired depth to crawl, and parameters to include or exclude particular URLs.

Authentication: the spider can be directed to use passwords or cookies to enter sites that require authentication, so that the user can index these sites.

Languages: the search engine should be able to handle any semantic NLP search coded in any language handled by Unicode.

Search and retrieval pages: customizable search and retrieval pages as provided for the user. The user can make the search and retrieval look any way desired, as well as enabling and disabling filters and options

Partial updating: the admin may select any number of files to re-index, rather than having to re-index an entire document base when individual documents are added or changed.

Automatic updating: a console-level indexing command is provided so that system administrators can automatically update new files or changed files on a regular basis, initiated by the computer clock.

Load balancing: the indexing interface automatically distributes document indexing across as many servers as the administrator selects.

Brokering: administrative tools enable the system administrator to manually control indexing and searching load, and membership access to document bases. The tools send queries to servers in response to load and allocate databases to specified servers.

Support of customer-specific criteria that may involve user parameters such as subscription membership.

Categorization: the interface queries users for categories and saves whole retrieval lists or individual files into the categories. New categories can be created on the fly. Subsequent searches can be restricted to categories.

User defined ontology: Users can add ontological classes by editing a standard file. In this way users can define search into classes unknown to the search engine semantics, such as company widget names or phrases. For example, Sony could add a category video-recorder with specific video-record names as category members. Using this new category, and user could ask "video−recorder" and retrieve to specific video recorder names mentioned in indexed documents.

User control of names: users can force preference for name or non-name interpretations of words like Bush and Stone, which can either, be names or common words.

4) How the semantic web contributes to get to the ideal search engine

The following table will explain at a feature level, the potential for improvement that the semantic web can supply.

Category	Feature	Semantic Web potential	Comments
Querying	Spelling correction		
	Linguistic Boolean Search	*	Ontology might already have the boolean relationships between terms in-built
	Proximity		
	Restrictors	*	By leveraging semantic relationships (i.e.: from Wordnet)
	Truncation		
	Syntactical Exclusion		
	Semantic Exclusion	*	Only possible with semantic structures
	Limits		
Recall and Precision	Relevance ranking	*	Semantic outperforms in terms of recall and precision syntactical search
	Fuzzy Search	*	Fuzzy ontologies might be the enablers
Results Presentation	Find Like	*	Semantic structures as result enables the graph-based fetching of related results with wider semantic scope
	Specific retrievals highlighted		
	Specific words highlighted		
Extras	Review and annotation	*	The advantages of semantic annotation are explained in [73]
Crawling And Indexing	Formats		
	Deep Crawl	*	The semantic web offers new ways of crawling the deep web, as explained in [72]
	Customer and meta tags		
	Indexing local directories		
	Spider configuration		

	Authentication		
	Languages	*	Semantic multi-language enablement
	Search and retrieval pages		
	Partial updating		
	Automatic updating		
	Load balancing		
	Brokering		
	Support of customer-specific criteria		
	Categorization		
	User defined ontology	*	
	User control of names	*	

5) Semantic Crawling and Indexing

The amount of information being made available on the Internet and the complexity of the new internet technologies have motivated the building of vertical search engines; each containing specialized indices and thus serving particular user needs.

A search engine with a specialized index has consequently a more structured content and offers higher precision than a generalized search engine, as it has already been intelligently extracted from the web.

This is known as focused crawling and requires special techniques to be performed. The term “focused” is given because of the selective seek out for pages that are only relevant to a set of pre-defined topics. If these topics are specified by using a domain knowledge means, like an ontology, instead of a plain set of key-words, the talk about semantic focused crawling.

The immediate advantages of a focused crawler are the saving in terms of hardware and network resources and the fact that the information is kept more up-to-date with much less effort

Designing a semantic focused crawler is difficult if you want to maintain a high harvest rate and yet stick to the topics of interest.

The right mix between keayword and a domain specific ontology allows for the best results in terms of the relevance between web pages and crawling topics. The issue with ontology-based focused crawling relies on the fact that concept weights and structures are heuristically predefined before being applied to calculate the relevance scores of web pages, which makes very difficult to acquire the optimal concept weights and adaptive structures to maintain a stable harvest rate during the crawling process. Introducing a learning mechanism can overcome the problem (i.e. the ontology self-evolving mechanism can dynamically adapt to the particular environment of the relevant topic, as proposed in [75])

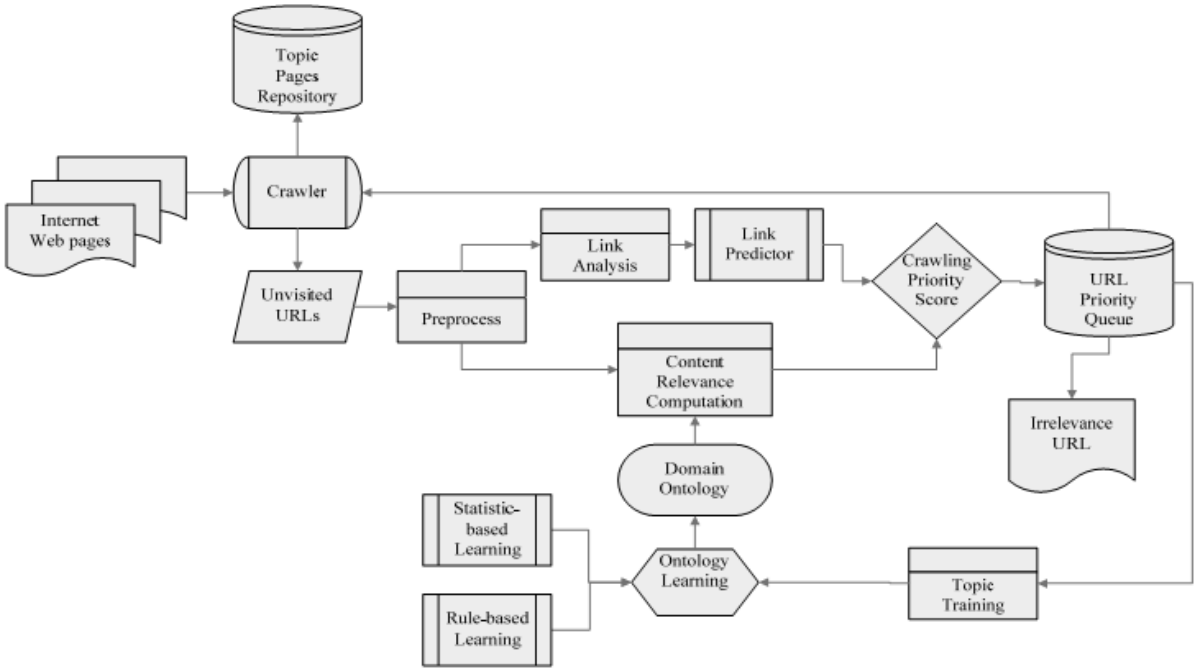


Figure 30 Semantic Focused Crawler Architecture

In [76] a pilot for an ontology driven crawler to gather information about the competitors and market events has been developed. This prototype combines advantages of existing Internet search engines with modern text analysis functionalities and an intelligent ontology based storage system for documents and knowledge items

6) Semantic Agents

The fact that agents acts on behalf of another person, an entity or a process to perform some tasks in an unsupervised way requires a perfect understanding between the agent and the medium the task should be performed in.

The usefulness of intelligent agents is demonstrated in various scenarios, like application automation and Internet commerce. Enhancing the conventional agents with ontology-based intelligence, as proposed in [77], can obviously enhance application integration and thus improve Internet commerce.

Focused information harvesting agents have been developed based on one intelligent online services called ITalks [78] that enables the communication between users and agents for locating IT Talks.

It's expected that this area will be extensively researched because the creation of intelligent agents is one of the most promising use cases in the entire semantic web initiative.

7) Knowledge Management (KM)

The usage of ontologies allows for establishing meaningful relations among data, unlike the linked Web.

The semantic web is relevant to knowledge management because it has the potential to dramatically accelerate the speed with which information can be synthesized, by automating its aggregation and analysis. Information on the Web now is typically presented in HTML format, and while very beneficial in some respects, the format offers neither structure nor metadata that is useful for effective management. Without structure, elements of content cannot be related to each other, and without metadata,

the nature of the elements themselves cannot be known. The discovery process is therefore human-centric and time-consuming[79]

8) Composition of complex systems

As we have extensively explained in the section about the Competitive Intelligence technologies, there are a enormous variety of tools in the spectrum of a practitioner.

The dream of all-in-one can come closer with the semantic enablement: as a matter of fact, it's possible to compose numerous Web services and Web contents to produce one more complex system.

Let's assume that the functionalities of each system are exposed as consumable web services. The complexity relies then on the discovery, the substitution, the composition and finally the management of services. In addition, there are also non-functional requirements such as availability, security and trust.

Leveraging the semantic web services will enable the (partially on the fly) discovery and composition of all complex systems.

9) Multimedia collection

Especially with the advent of the new media within the context of the Web 2.0, the volume of non-textual information duplicates in even shorter period of times. A lot of CI relevant information is in non-textual format, and needs to be gathered

The semantic web via ontology offers a way to enable semantic annotations that could be easily organized and found

10) Information filtering

Semantic driven filtering should be more effective than keyword based filtering. Information filtering occurs in the processes of information receiving, sending and storing by filtering unwelcome data and Web services invocation, sending selectively to the right clients and storing in the right place for the value-added consumption. An enterprise always has to process a huge amount of e-commerce data.

11) Machine dialogue across the domains

Ontology provides formal semantics, thereby making not only humans but also machines understandable. Additionally ontology mappings or translation could improve the domains alignment as well as enabling the dialogue across multiple domains. To really establish bridges, translation is required.

Internet commerce requires automatic negotiation and contracting for all searched results. This feature could significantly help machines process a large amount of business partner information that humans cannot handle, and thus save time and money.

12) Virtual community

The current globalization trend requires the building of virtual teams that work in a shared space. Ontology can be used to define relationships in the community from forming, organizing, communicating to demising automatically.

XXVII. How the Semantic Web improves The Analytical Methods

If we for example take War Gaming and have a look at the steps to perform it, we can easily recognize where the semantic web techniques can have a major impact and bring a higher benefit. The industry assessment can be performed with various frameworks, but the results should more or less contain and assessment of the key customers, the key suppliers, the intensity of the competition, as well as a description of the level of threat of alternative products and services (actually, nothing but the Porter five forces).

The competitive analysis consists of assessing how each team can respond to each particular issue. For the formulation of strategies, it is required that each team presents them to the group, where their validity is discussed and a re-assessment might be required.

A semantic focused crawler can be set up to perform the information retrieval of industry relevant only documents. Information extraction techniques would do nicely, especially if enriched with semantic annotation and indexing of deep web content (which can also contribute to certain automatism to extend the domain ontology as new knowledge is discovered)

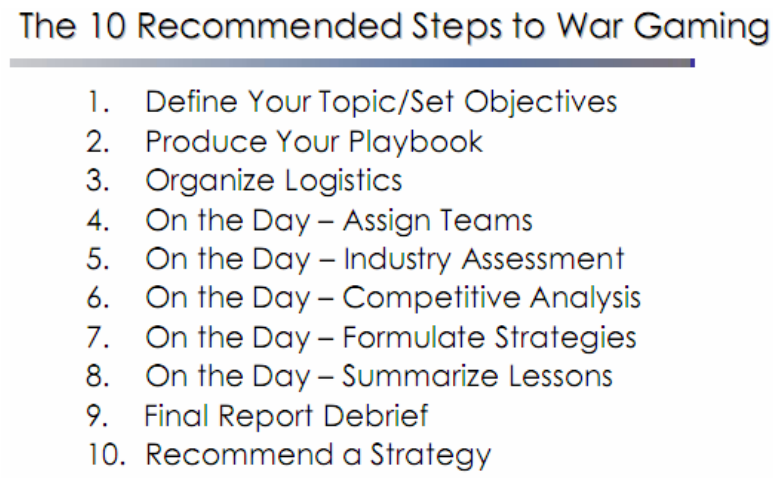


Figure 31 Ten Steps to run War gaming

A semantic focused crawler can be set up to perform the information retrieval of industry relevant only documents. Information extraction techniques would do nicely, especially if enriched with semantic annotation and indexing of deep web content (which can also contribute to certain automatism to extend the domain ontology as new knowledge is discovered)

No matter which analytical technique we use to infer intelligence, the semantic web might even allow for building up an application where each one is driven by the so called “analytical ontology”. Governing them, there can be an “umbrella ontology” enabling the cross-check of the findings.

The use of semantics is essential to establish connections among the inputs information and findings of the employed techniques towards a more automatic analytical process

XXVIII. Conclusion

The advent of the semantic era is a matter of fact. So it is the need for faster, more precise and easier to gather Competitive Intelligence in the global economy setup.

Never before there has been so much and so up-to-date information available on the internet and never before the online medium has had so much potential for Competitive Intelligence activities. Unfortunately, there is an inconvenient "other side of the coin" for the CI practitioner, which is the unmanageable complexity of such an information overload.

The technologies and tools that are currently available for a CI practitioner can hardly deal with the augmented complexity and volume. The only way forward is to shift tasks that have traditionally been by humans to the software terrain, which requires a machinery understanding of the web documents. This is exactly the credo of the semantic web: enabling automatic processing of web documents by machines according to a semantic basis.

In this research, we have broken down the CI universe to its dimensions of interest and we have demonstrated to which extend the semantic web technologies can contribute to the success of CI practitioners:

From the staffing perspective, we have demonstrated how each and every role within a typical CI unit can be streamlined by decreasing the amount of repetitive manual work and enabling the access to resources that weren't reachable so far, the efficiency increases and the error-proneness decays.

From the technological point of view, we fully explained how the current technologies can "go semantics" to exploit their potential. Especially in the fields of information retrieval and focused crawling, information extraction and information integration the use of semantic web technologies enables unprecedented results.

From the support to the CI Cycle perspective, we fully discussed how each particular step could significantly benefit, along with the semantic leverage of technologies available at present and the introduction of innovative pure semantic techniques.

If we talk about the analytical methods that are most commonly employed to gain competitive intelligence out of the available information, we have seen the tremendous potential that this step towards semantics would bring.

There are cross-the-board benefits that are not only exclusivity of CI professionals but very attractive for those who embrace this new semantic-based web: get more for less (clicks), do things only once (avoiding repetitive tasks), get your data easily integrated, get faster to the things you seek out, get the information that was hidden to the conventional search engines and last but not least, leverage whatever you have in place with the semantic flavor.

If we were asked to point out the semantic web core differentiator, we would say that it brings the rigor of science and logic to the management of data, models, and business rules for the first time in the history of information management. Large-scale data processing, that was only possible with people in the process to help out, can for the first time be rigorously driven by machines.

Provided that the Competitive Intelligence plays already a determining role in the strategy of each modern corporation, it is essential that the entire infrastructure of business strives for supplying good information about the competition and the market, so that the decision makers move out of their gut instinct and take well-founded decisions.

Another aspect of the semantic web value proposition is the attention shift of information technology (IT) moving away from technology and emphasizing the information aspect. The data that uses to be second class citizen within the organizations' landscape has become a precious and necessary asset and competition enablement and the arsenal of semantic web techniques can make the most of this data.

May our vision statement be the colophon of this research: "*With the progressive introduction of the semantic web the competitive intelligence will be brought to a next level"*

XXIX. Outlook or next steps

Having demonstrated the way the semantic web can bring the Competitive Intelligence to the next level, we now present a set of research directions and initiatives that can be followed to materialize all benefits we have presented:

- Defining a system architecture to enable the CI professionals to take advantage of the explained benefits of introducing semantic web techniques along the CI cycle.
- Ontology driven modeling of the competitive market for a given industry
- Ontology mapping research to leverage in-house models combined with industry standard models
- Elaborating agent-driven benchmarking products (i.e.: by learning the attributes, extracting their values after information retrieval and mapping them back to the learned ontology)
- Enable sentiment collection in forums and other collaborative platforms to exploit
- Analyze the social media networks related to your competitors to identify alpha users and potential churn candidates that might be targeted for your own acquisition
- Semantic agent-based monitoring of the competition
- Ontology driven trustworthiness of the resources (bringing the “Trust” layer of the semantic web stack into picture)

APPENDIX I: INTERNET DEVELOPMENT TIMELINE

Year	Invention(s)/Event(s)	Inventor(s)/Author(s)	Description
1945	Memex	Vannevar Bush	Memex is photo-electrical-mechanical device which could make and follow links between documents on microfiche
1963	Term Hypertext coined	Ted Nelson	First use of the term hypertext
1965	Term Hypermedia coined	Ted Nelson	First use of the term hypermedia
1967	HES	Andries van Dam, IBM	Hypertext Editing System developed
1968	Hypertext on NLS	Doug Englebart	Doug Englebart publicly demonstrates Hypertext on the NLS on December 9, 1968.
1979	Usenet first started		Usenet will grow to allow millions of users to access millions of articles on various subjects
1980	Enquire	Tim Berners-Lee	Enquire allows links to be made between arbitrary nodes
1987	Perl 1.0	Larry Wall	Larry Wall introduces Perl 1.0
1990	Archie, the first search engine, released	Alan Emtage, Bill Heelan, and Mike Parker at McGill University	The first search engine Archie, written by Alan Emtage, Bill Heelan, and Mike Parker at McGill University in Montreal Canada is released on September 10, 1990
1990	Gopher, a menu driven search and retrieval tool	University of Minnesota	Gopher is a menu-driven search-and-retrieval tool that helps Internet users locate information online.
1990	Hypertext system proposed	Tim Berners-Lee and Robert Cailliau at CERN	In 1990, Tim Berners-Lee, working with Robert Cailliau at CERN, proposes a 'hypertext' system, which is the first start of the Internet as we know it today.
1990	WWW introduced to the public on August 6, 1991.	Tim Berners-Lee	The World Wide Web is launched to the public August 6, 1991. Tim Berners-Lee, a scientist at the European Partial Physics Laboratory (CERN) in Geneva, Switzerland develops the Web as a research tool.
1992	MIME		MIME standard is defined, providing browsers and mail programs with a standard method for processing non-textual media files and streams
1993	Mosaic	NCSA	The NCSA releases the Mosaic browser, the first

			widely available browser, April 22, 1993.
1993	HTML		HTML programming language released
1994	Netscape founded	Marc Andreesen and James H. Clark	Netscape is found by Marc Andreesen and James H. Clark.
1994	W3C	Tim Berners-Lee	The World Wide Web Consortium is founded by Tim Berners-Lee.
1994	PHP	Rasmus Lerdorf	Rasmus Lerdorf creates PHP, a widely used scripting language for web page development
1994	YAHOO	David Filo and Jerry Yang	YAHOO, the first web portal and search engine, is created in April, 1994.
1995	Java	Sun Microsystems	Java programming language is introduced.
1995	Wiki	Ward Cunningham	The first Wiki is created.
1995	Amazon.com	Jeff Bezos	Amazon.com, one of the largest and well known e-commerce sites today opens its website for the first time on July of 1995.
1995	eBay	Pierre Omidyar and Jeff Skoll	EBay is founded and goes on to become the largest online auction website
1996	Macromedia Flash	Macromedia	Macromedia purchases FutureWave and later releases Macromedia Flash 1.0
1996	Altavista	DEC	DEC introduces Altavista, the first search engine that stored every word of every web page in a fast searchable database
1997	Altavista Babel Fish	DEC	Altavista introduces its free online translator Babel Fish, the web's first machine translation service capable of translating words and sentences to and from English, Spanish, French, German, Portuguese, Italian, and Russian
1998	MySQL	Michael Widenius	MySQL is introduced; MySQL is a relational database management system which is publicly available for use under the GNU General Public License; became very popular for website backend databases
1998	Google	Sergey Brin and Larry Page	Google is founded by Sergey Brin and Larry Page September 7, 1998.
1999	MySpace	Tom Anderson, and Chris Dewolfe	MySpace starts
2004	Firefox	Mozilla	Firefox 1.0 is first introduced on November 9,

			2004.
2004	iTunes	Apple	As broadband becomes more popular, media companies start selling music and video online. Apple's download store for its trendy iPod portable music players
2004	Paid mp3 download store	Napster	Napster relaunches as a paid music download store.
2004	Facebook	Mark Zuckerberg	Launch at Harvard University. Within three years, the social networking site has 30 million members. By 2009, Facebook boasts of over 200 million active users (those who have logged in the last 30 days).
2004	Photo sharing (Flickr)	Ludicorp	Coinciding with the rise in digital photography. (Kodak discontinues reloadable film cameras in Western Europe and North America in this year.)
2005	Video over internet (Youtube)	Steve Chen, Chad Hurley and Jawed Karim	Platform to enable people to easily publish videos online. Google acquired that within one year for $1.65 billion. Youtube users were at the time uploading 65K new films and watching 100 million clips a day.
2005	Voice over internet (Skype)	Ahti Heinla, Priit Kasesalu and Jaan Tallinn	Enables 2 million calls at any moment and has an user base of 53 million. Got acquired by eBay
2006	Twitter	Jack Dorsey	Microblogging emerges In stark contrast to the proliferation of lengthy blog posts online, Twitter messages are limited to 140 characters
2008	Scoring concept in auction platforms	Jack Sheng	Jack Sheng becomes the first person to earn an eBay feedback score of one million. From startup capital of $500, Sheng has built a $40 million business selling gadgets. eBay creates the shooting silver star to designate users with a feedback score of over a million.
2008	Google’s 10th birthday	Google	The company that began with a search engine now also dominates online advertising and has a leading presence in online mapping, webmail and online document collaboration. Google's search engine indexes 1 trillion unique URLs and there are several billion new webpages published every

			day. Google encroaches on Microsoft's territory with the launch of the Google Chrome browser.
2008	Mobile web advertising		The mobile web reaches critical mass for advertising, according to Nielsen Mobile. In the US, there are 95 million mobile internet subscribers and 40 million active users. US mobile penetration is 15.6%, compared to 12.9% in the UK. Mobile internet generated $1.7 billion in revenue in the first quarter of 2008.
2009	Twitter’s one million subscriptions	Ashton Kutcher	Actor Ashton Kutcher becomes the first person on Twitter to have a million followers subscribing to his 'tweets'
2009	iPlayer goes high definition	BBC	The BBC announces its iPlayer will go high definition. It was first launched Christmas 2007 and is used to stream programmes over the internet for up to a week after their television broadcast. Two thirds of Britons have broadband access at home, and there were 1.5 million new broadband subscribers in 2008.

REFERENCES

[1] Metcalf Carr, M. *Super Searchers on Competitive Intelligence: the online and offline secrets of CI researches*. Information Today, Inc. Medford, New Jersey 2003

[2] Bouthillier, F. and Shearer, K. *Assessing Competitive Intelligence Software. A guide to evaluating CI Technology. Information Today*, Inc. Medford, New Jersey 2003.

[3] Davies, J., Studer R. and Warren, P. *Semantic Web Technologies. Trends and Research in Ontology-based Systems*. John Wiley & Sons Ltd, West Sussex, England. 2006. Foreword

[4] SCIP: Society of Competitive Intelligence Professionals (accessed 2008), available at http://www.scip.org/

[5] Industry Canada. *SME Direct: Competitive Intelligence*. Available at http://smedirectcanada.com/ci [Accessed June 2008]

[6] Fleisher, C.S and Benkhorn, D.L. *Managing Frontiers in Competitive Intelligence.* Westport, CT: Quorum Books, 2001

[7] Day, G.S., Reibstein, D.J. and Gunter, R.E. *Wharton on Competitive Strategy*. John Wiley & Sons Inc. Toronto, ON, 1997

[8] Gruber, T. *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Journal Human-Computer Studies 43, 907–928 (1993)

[9] Fleisher, C. and Bensoussan B.: *Business and Competitive Analysis: Effective Application of New and Classic Methods.* FT Press, 2007.

[10] PowerHomeBiz: *Building a Competitor Profile*, available at http://www.powerhomebiz.com/vol4/competia.htm [Accessed October 2008]

[11] Kahaner, L. *Competitive Intelligence: How to gather, analyze, and use information to move your business to the top*. New York: Touchstone 1998

[12] Bernhardt, D. *Competitive Intelligence How to acquire and use corporate intelligence and counter-intelligence*. Prentice Hall, NY 2003

[13] Grzanka, L. Competitive Intelligence. Knowledge Management Magazine. Available at http://www.destinationkm.com/articles/default.asp?ArticleID=726 [Accessed Summer 2008]

[14] Electronic College of Process Innovation. Framework for Managing Process Improvement. 2001

[15] Bernhardt, Douglas. *Competitive Intelligence*: Acquiring and using corporate *intelligence* and counter-*intelligence*., 2003, *FT Prentice Hall*

[16] Hussey, D. and Jenster, P. *Competitor Intelligence: Turning Analysis into Success.* Chichester: John Wiley & Sons. 1999

[17] Hoge Jr, J.F. and Rose, G. *How Did This Happen: Terrorism and the New War.* Oxford: PublicAffairs, 2001

[18] Sheth, A. *Semantic Meta Data for Enterprise Information Integration.* Information Management Magazine, July 2003 [Accessed October 2009]

[19] Herring, J.P. *Key Intelligence Topics: A Process to Identify and Define Intelligence Needs*. Hartford, CT: Jan P. Herring & Associates. 2002

[20] iMentor Management Consulting. Geneva 2002

[21] Cook, M. & Cook, C. *Competitive Intelligence*. London: Kogan Page 2002

[22] Kassler, H.S. Mining the Internet for competitive intelligence. How to track and sift for golden nuggets. Online: The Magazine for Online Information Systems, 12(5), 34-45. 1997

[23] Vriens, D., Achterbergh, J. (2003), "The source map: a means to support collection activities", in Vriens, D. (Eds),*Information and Communications Technology for Competitive Intelligence*, Idea Group Publishing, Hershey, PA, pp.181-93.

[24] Hofstede, G.H.. *Culture and organizations—Software of the mind.* London: Harper Collins. 1991

[25] Likert, R. *A Technique for the Measurement of Attitudes*. Archives of Psychology 140: 1–55. 1932

[26] O'Brien, J. *Introduction to Information Systems (2nd edition).* New York: McGraw-Hill, 1998

[27] Kirk, E. Evaluating Information Found on the Internet. Available at http://www.library.jhu.edu/researchhelp/general/evaluating/ [Accessed October 2009]

[28] *Know! Competitive Intelligence:* Published Source Collection. A division of Knowledge inForm, Inc.Know!Books Press, 2005

[29] Q. Tao and J. E. Prescott. *"China: Competitive intelligence practices in an emerging market environment."* Competitive Intelligence Review, Vol. 11, 2000, pp. 65-78.

[30] HistoryNet Staff, *Espionage in Ancient Rome*, Available at http://www.historynet.com/espionage-in-ancient-rome.htm [Accessed October 2009]

[31] Chanakya, *Arthasastra*, Available at http://en.wikipedia.org/wiki/Arthasastra [Accessed October 2009]

[32] Holmes, George. *Oxford History of Medieval Europe*. New York: Oxford University Press, 1988.

[33] Sable, M.H. *Industrial espionage and trade secrets: an international bibliography.* New York: Harworth Press 1985

[34] Porter, Michael E., *Competitive Strategy. Techniques for Analyzing Industries and Competitors*, New York: Free Press 1980

[35] Prescott, J.E. *The Evolution of Competitive Intelligence*. Proposal Management APMP 1999

[36] Blog. Wikipedia. Available at http://en.wikipedia.org/wiki/Blog [Accessed October 2009]

[37] Microblogging. Wikipedia. Available at http://en.wikipedia.org/wiki/Microblogging [Accessed October 2009]

[38] Church, Sally. Using Delicious to mine for pharmaceutical marketing and competitive intelligence information. Available at http://www.pharmastrategyblog.com/2009/01 [Accessed October 2009]

[39] SciTech Library Question. Available at http://stlq.info/ [Accessed October 2009]

[40] Fisch, K., McLeod, S. Bronman J. *Did You Know? 3.0 (Official Video) -2009 Edition*, Available at http://www.youtube.com/watch?v=PHmwZ96_Gos&feature=fvst [Accessed October 2009]

[41] DI Analytic Toolkit. Available at http://www.cia.gov/cia/di/toolkit/index.html [Accessed October 2009]

[42] Unicode Org. *Unicode* Available at http://unicode.org/ [Accessed November 2009]

[43] W3C, *XML* Available at http://www.w3.org/XML/ [Accessed November 2009]

[44] W3C, *URI* Available at http://www.ltg.ed.ac.uk/~ht/WhatAreURIs/ [Accessed November 2009]

[45] W3C, *RDF* Available at http://www.w3.org/RDF/ [Accessed November 2009]

[46] W3C, *OWL* Available at http://www.w3.org/TR/owl-features/ [Accessed November 2009]

[47] W3C, *SPARQL* Available at http://www.w3.org/TR/rdf-sparql-query/ [Accessed November 2009]

[48] W3C, *RIF* Available at http://www.w3.org/2005/rules/wiki/RIF_Working_Group [Accessed November 2009]

[49] Pollock, J. Semantic Web for Dummies. Wiley Publishing Inc. Indianapolis, 2009

[50] Noy, N., et al., *Creating Semantic Web contents with Protege!-2000*, IEEE Intelligent Systems 16 (2001 March/April).

[51] Parry, D. ACM International Conference Proceeding Series; Vol. 54 *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation* - Volume 32, New Zealand 2004

[52] Slomin D. and Tengi R., *WordNet*, Princeton University Cognitive Science Lab., 1-003

[53] Goble, C., Stevens, R., Ng, G., Bechhofer, S., Paton, N., Baker, P., Peim, M., and Brass, A. *Transparent access to multiple bioinformatics information sources.* IBM Systems Journal, Issue on Deep computing for the life science 2001

[54] *Microformats*, Available at http://en.wikipedia.org/wiki/Microformats [Accessed November 2009]]

[55] Agrawal, R. *Difference between OWL Lite, DL, and Full*. Available at http://ragrawal.wordpress.com/2007/02/20/difference-between-owl-lite-dl-and-full/ [Accessed November 2009]

[56] Mani, I. 2001. *Automatic Summarization*. Amsterdam: John Benjamins. 2001

[57] Lancaster, F.W., *Information Retrieval Systems: Characteristics, Testing and Evaluation,* Wiley, New York (1968).

[58] Zhai, Y. and Liu B.Automatic *Wrapper Generation Using Tree Matching and Partial Tree Alignment*. Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, Boston, 2006

[59] Kushmerick N. *Wrapper induction: efficiency and expressiveness.* Artificial Intelligence, 118:15–68. 2000

[60] Hsu C.; and Dung M. *Generating finite-state transducers for semi-structured data extraction from the web*. Inf. Syst., 23(9):521–538. 1998

[61] Muslea I.; Minton S.; and Knoblock C. *A hierarchical approach to wrapper induction*. In AGENTS'99. 1999

[62] Crescenzi V., Mecca G., Merialdo P., *Handling irregularities in Road Runner.* In AAAI Workshop on Automatic Text Extraction and Mining (ATEM2004) 2004

[63] Arasu, A. and Garcia-Molina, H., *Extracting structured data from Web pages*. SIGMOD-03, pp. 337-348, 2003.

[64] Chang, K. and He, *Accessing the deep web.* Communications of the ACM. May 2007/Vol. 50, No. 5

[65] Rakesh Verma, Ping Chen, Wei Lu, *A Semantic Free-text Summarization System Using Ontology Knowledge*, IEEE Transactions on Information Technology in Biomedicine, Vol. 5(4), pp. 261-270, 2007

[66] Graupmann, J. and Schenkel, R. *The Light-Weight Semantic Web: Integrating Information Extraction and Information Retrieval for Heterogeneous Environments*. SIGIR 2005 Workshop on Heterogeneous and Distributed Information Retrieval, Brazil 2005

[67] Müller, C. and Gurevych, I. *Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval.* In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark, September 17-19, 2008.

[68] J. Koberstein and Y.-K. Ng. *Using Word Clusters to Detect Similar Web Documents.* In J. Lang, F. Lin, and J. Wang, editors, KSEM, volume 4092 of LNCS, pages 215–228. Springer,2006

[69] E. Gabrilovich and S. Markovitch. *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. In Proceedings of IJCAI, pages 1606–1611, 2007.

[70] Calegari S. and Sanchez E. *A Fuzzy Ontology-Approach to improve Semantic Information Retrieval.* Proceedings of the Third ISWC workshop on Uncertainty Reasoning for the Semantic Web - URSW'07

[71] Proctor, R.W., Kim-Phuong L. Vu. *Handbook of human factors in Web design*. L. Erlbaum Associates (Mahwah, NJ, London, UK), 2004

[72] Geller, J., Chun S. and Yoo Jung An. *Toward the Semantic Deep Web*. Computer, vol. 41, no. 9, pp. 95-97, Sept. 2008

[73] Kiryakov, A, Popov B., Terziev I., Manov D., Ognyanoff D. *Semantic annotation, indexing, and retrieval.* J. Web Sem. 2(1): 49-79 2004

[74] Jobber, D. *Principles and Practices of Marketing*. McGraw-Hill 2004

[75] Liyi Zhang, Mingzhu Zhu and Wei Huang. *A framework for an ontology-based e-commerce product information retrieval system* Journal of Software,VOL. 4, NO. 5, JULY 2009

[76] Engelbach, W. Gaida, D. Rybinski H., Specht T. *Ontology based search and storage of market information.* Medien- und Filmgesellschaft Baden-Württemberg, Stuttgart ; European Media Laboratory GmbH -EML-, Heidelberg 2007

[77] Hendler, J. *Agents and the semantic Web*, IEEE Intelligent Systems, 16. 2001

[78] Cost, R.S., Finin, T, Joshi, A., Peng, Y., Nicholas, C., Soboroff, I., Chen, H., Kagal, L., Perich, F., Zou, Y. and T. Sovrin. *ITtalks: a case study in the semantic Web and DAML+OIL*, IEEE Intelligent Systems, 17, 2002

[79] Lamont, J. *Semantic Web holds promise for KM*. Available at http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=19136 [Accessed October 2009]

[80] Zhu, H., Siegel, M.D. and Madnick, S.E. *Information aggregation - a value-added E-Servic*', in Proceedings of the International Conference on Information Retrieval 2001