TECHNISCHE
UNIVERSITÄT
WIEN

VIENNA
UNIVERSITY OF
TECHNOLOGY

DISSERTATION

Business Intelligence in the Logistic Domain using Visual Information Extraction

ausgeführt zum Zwecke der Erlangung des akademischen Grades
einen Doktors der technischen Wissenschaften
unter der Leitung von

o. Univ.Prof. Dipl.-Ing. Dr. Georg Gottlob
E1842
Institut für Informationssysteme
Abteilung für Datenbanken und Artificial Intelligence

Eingereicht an der Technische Universität Wien
Fakultät für Technische Naturwissenschaften und Informatik

von
M.Sc. José Aldo Díaz Prado
Matrikelnummer: 0322780
Forsthausgasse 2-8/2136, A-1200 Wien

Wien, im April 2006

To my mother
María Guadalupe Prado-Torres,
Who supports me in all my life projects.

Deutsche Zusammenfassung der Dissertation

Wir leben heute in einer Wettbewerbsgesellschaft: Unternehmen müssen noch konkurrenzfähiger als ihre Mitbewerber sein. Sie müssen genauer über das Handeln ihrer Mitbewerber Bescheid wissen, indem sie jede verfügbare Information sowohl im Internet, als auch in anderen internen und externen Datenbanken nützen. Außerdem müssen Firmenverantwortliche wissen, welche Strategien ihre Mitbewerber implementieren, und in welche Richtung sich diese in ihrem Geschäftsfeld bewegen: Wer sind die neuen Kunden? Welche Neuakquisitionen stehen an, welche Unternehmenszusammenschlüsse? Wo finden diese statt? Und was sind die neuen Strategien? Und, darüber hinaus, was passiert genau mit einem bestimmten Mitbewerber?

Unter Berücksichtigung all dieser Umstände schlägt diese Dissertationsschrift vor, die Tätigkeiten einiger konkurrierender Wirtschaftseinheiten in der Zementindustrie zu beobachten, im konkreten die der Cemex Corporation. Das Unternehmen kann (hauptsächlich mit HTML formatierte) Informationen von den einzelnen Websites seiner Mitbewerber extrahieren. Für den Extraktionsprozess wird unter Anwendung der Lixto-Software-Werkzeuge die Verwendung von Datenextraktionstechniken vorgeschlagen.

Unter dieser Perspektive sollen die Entscheidungen, die aus dem Business-Intelligence-Prozess bei Cemex – basierend auf relevanter Information – gewonnen werden, den Firmenverantwortlichen helfen, effektivere Handlungsstrategien zu entwickeln und zu verwirklichen. Diese Entscheidungen haben daher größere Implikationen für die Wettbewerbsfähigkeit des Unternehmens und in weiterer Konsequenz eine stärkere Auswirkung auf Folgeentscheidungen.

Mit diesem Ziel vor Augen ist es die Absicht dieser Dissertationsschrift zu beweisen, dass die Konzepte der Datenextraktion und der Wiederverwendbarkeit und Wiedernutzbarkeit von Information aus verschiedenen Quellen aus dem Internet die Stategien zur Positionierung in der Zementindustrie unterstützen. In dieser Forschungsarbeit wurde der Stand der Fähigkeiten und Möglichkeiten der Mitbewerber untersucht. Das Konzept der visuellen Informationsextraktion wurde verwendet, um unter Ausnutzung geografischer Informationssysteme (GIS) einige elektronische Karten zu erzeugen.

Die Integration geografischer Informationssystemen bringt einen Mehrwert für die Informationsanalyse und unterstützt außerdem die Transformation der Daten in visuelle Information.

Schlüsselwörter

Business Intelligence, Datenextraktion, visuelle Business Intelligence, Datentransformation, Web-Recycling, Web-Integration, Datenrepositorien, XML, Python, MySQL, Wissensrepositorien, Wissensmanagement, Data Warehouse, visuelle Informationsextraktion.

Abstract

Today, as we are living in a competitive business world, corporations need to be even more competitive than their market peers. They require knowing more in detail what is happening with their competitors in the business arena. The corporations will get to know more in detail about their competitors' activities and strategies through the analysis of information available on the Internet and other internal or external data bases. Also, corporations must know what strategies their competitors are implementing and in which direction they are moving in the business market. Who are the new customers? What are their new acquisitions? What are the new mergers? Where are they located? And, what are the new strategies? Moreover, what is happening with each of their other competitors?

With the previous antecedents and to develop this work, this thesis proposes monitoring the performance of some of the competitors' business entities in the cement industry, specifically the Cemex Corporation. The corporation can extract the information from each of the competitor's websites that are present in the World Wide Web. (This information is mainly formatted in HTML). For the extraction process, it is proposed to use techniques of data extraction, applying the Lixto Software Tools.

Under this perspective, the competitive business intelligence decisions in Cemex (sustained by relevant information) should help the executives in the organization to design and take more effective strategic actions. These decisions will have larger implications for the competitiveness of the company, which as a consequence will have impact on subsequent decisions.

With this perspective, the intention of this thesis is to prove that the concepts of data extraction, reusability and recycling of information from different sources on the internet will support the strategies of positioning in the cement industry. In the case of this research work, the position of competitors' facilities was investigated. The concept of Visual Information Extraction was used to generate some electronic maps using the Geographical Information Systems (GIS).

The integration of Geographical Information Systems will add value to the activity of information analysis and will also support the process of transforming data to visual information.

Keywords
Business Intelligence, data extraction, visual business intelligence, data transformation, web recycling, web integration, data repositories, XML, Python, MySQL, knowledge repositories, knowledge management, data warehouse, visual information extraction.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter the general approach of this thesis is explained under the subject of Visual Information Extraction from web sources to internal information systems of corporations. The ideas and technologies already explored in other scientific works will be integrated. Special attention will be regarded to the logistic domain of a cement company in Mexico.

It will be explained the approach to develop Strategy Business Intelligence through Visual Information Analysis. This process starts with data extraction, using the Lixto Visual Wrapper technology [62]. The process happens through the generation of repositories of information of the extracted web data. Finally, Visual Business Intelligence & Information Modeling is possible through the Geographical Information System (MapInfo) [66].

1.1 Motivation

Companies around the world are interested in getting relevant information about their main competitors. These companies demand detailed information of the different strategic areas of operation of their competitors. As an example of domain, we look into the domain of logistics. Any given corporation is interested in knowing the different geographic positioning of warehouses and distribution centers of its competitors since this information has a high strategic value for future analysis and decisions.

The main motivation in doing this research is to meet the need a cement corporation has to solve the problem of acquiring automatic information by monitoring its competitors and extracting information in real-time. The attempt is to use parallel and complementary technologies that allow the modeling of visual information as result of an intelligence process based on data extraction from web sites.

In this approach, the controlling idea is to recycle the information available in the competitors' Web sites. These web sites provide an ocean of semi/structured data which should be transformed and depicted in a structured data format. The structured data will be available to generate the repositories of information to create the models of different scenarios by means of maps with MapInfo.

1.2 Data Extraction

Before the data extraction begins, it is intended through this thesis to analyze different problems existing in the web, such as how to find and have access to relevant information, as well as to tackle the present deficiencies of historical information in the different web sources and the limitations in search engines. In addition, it is also explored the possibility of using a data warehouse as part of the information management system so as to integrate and share the information extracted from web sites.

The proposal is to do data extraction through the development and application of wrappers using the Lixto Visual Wrapper technology. After the wrapper execution generates the extracted XML, it is proposed to generate a repository of information of the different web sources of competitors. The competitors information will be stored in a MySQL database[71].

1.3 Business Intelligences First Level (BI Repository)

The first stage is denominated Businesses Intelligence First Level, which is the foundation to start the recycling and storing of information in a structured format using the relational database technology.

The information will be stored in the repository of information and it will be accessed in different forms. The final user will execute some queries to generate reports that he/she will use as raw data for future logistic applications.

In order to reuse the information as raw data, it will be necessary to make a series of transformations to the information generated from MySQL. This information will later be used as raw data in a plain text format (PTF).

1.4 Business Intelligence Second Level (BI Integration)

In the second level of Business Intelligence. the information contained in the competitors' repository will be integrated with the information systems within the company such as: CRM systems, date warehouse, data mining, logistic systems, marketing systems, and financial systems, among others.

One of the most important contributions of this thesis is here in this second level of Business Intelligence. Here, the process of information recycling starts. The information contained in the MySQL databases in structured format will be used as raw data for different computational applications within the organization.

For the purpose of this thesis, it has been considered to work with an application in the logistic area which will generate the maps from the information existing in the repository of information (Business Intelligence Repository).

1.5 Information Modeling

In order to develop the information modeling, the different kinds of models for visual representation of the information have been analyzed and it is proposed to make the modeling of information using a token type model [54]. In UML, such models are referred to as snapshot models or representation models .

This kind of token model is utilized in the case of maps. A map is a "metamodel" that can be an instance of the "class-class" type, by which the modeling of information is made by a token type model. This kind of model applies for "model driven development" or "capturing systems configuration". Token models are what people often have in their minds when talking about ordinary models.

With this information modeling, it is desired to create the data model that will be used for the databases of the information repository, MapInfo and the Corporate Information Factory. The most important variables to make the modeling of maps will be generated with this data model and will also be used also for the interaction with other external information sources through the Corporate Information Factory.

The most relevant information for MapInfo is mainly extracted from tables in the databases and it conforms to the repository of information. The information including country, city, street, number and, zip-code, will be used as raw data by MapInfo. Later, the system will automatically generate the visual model (map) in which each of the competitors' warehouses and/or distribution centers is depicted.

1.6 Information Analysis

The stage of information analysis will be generated exploiting the maps of MapInfo. For the organization's decision makers, it will be easier to visualize great volumes of information through graphs, charts or images. Information mapping provides the opportunity of identifying business opportunities at a reduced cost and in a short hitch. It also saves the effort of analyzing a large amount of information in hard copy or exploring seemingly senseless information.

The information mapping will also allow comparative analyses in visual form, contrasting different sources of information about the location of warehouses and distribution centers for two or more competitors.

It will be possible to integrate new information sources for further analysis that supplements the information extracted from WebPages. These new information sources will come from internal and external sources of the corporation. In the case of this research, it will be used a database containing statistical and geographical information about the country in analysis. This information will be part of the Corporate Information Factory.

The Corporate Information Factory plays a very important role in the integration of the information to the internal information systems in the organization. It is aimed at guaranteeing the homogenization of the different data models from the different information systems, and protecting the quality of the information contained in systems such as CRM, data warehouses, and data mining among others.

With the integration of new information sources in the maps, the decision makers can undertake a more detailed analysis of information coming from different sources. This enriches the processes of visual information extraction and of recycling information, thus, generating an added value to the organization.

1.7 Contribution of this thesis

As previously mentioned, the intention of this thesis is to develop a visual model for the information extracted from multiple competitors' web sources (Visual Information Extraction). Also, it is sought to generate strategic value to the organizations who are interested in doing Business Intelligence in a Visual Way.

The work done in this thesis will try to prove that it is possible to make an information model based on the recycling of information from multiple web sources, transforming it into a visual model for its analysis. Another contribution is to strengthen the visual modeling through the integration of additional external and internal information sources with the interaction of the Corporate Information Factory.

At the same time, it is intended to show that the World Wide Web can be a great source of information for the process of Business Intelligence through the application of technologies of data extraction such as Lixto [62].

The aim is not only to provide an information storing system -coming from a set of information in a data warehouse or information repository,- but also to go beyond these boundaries by creating visual models through recycling information. This last will facilitate the process of disseminating structured information for future analysis and strategic planning.

1.8 Organization of the thesis

In this section the outline of this thesis is presented. Also, the highlights of the chapters that follow are introduced.

This introductory chapter has provided an overview of the problem domain including descriptions of possible application scenarios. It has also taken care of the first presentation of Business Intelligence in the logistics domain, using visual information extraction which this thesis advocates and on which it will be elaborated in the following chapters.

- Chapter 2 presents an overview of the intelligence process. The chapter describes the importance of creating advantages that provide a competitive edge by means of intelligence. It also sheds light on the importance of carrying out strategic planning in corporations.

- Chapter 3 provides an overview of the competitor´s radiography. This chapter is devoted to point out the importance of sketching the competitors' profiles so as to identify their future strategies and plans, to predict their possible reactions to competitive initiatives, and to understand their weaknesses.

- Chapter 4 undertakes the analysis of the main problems associated with data extraction on the web such as: lack of productivity, lack of historical data, limitations of search engines and warehouses, and the importance of XML technologies.

- Chapter 5 evinces a description of the automated wrapper generation and its associated technology, such as the wrapper protocol and its data model for the HTML documents. In addition, an analysis of some wrapper generation tools is included.

- Chapter 6 analyzes different kinds of data modeling. This chapter also justifies the reasons why a token model is used for information modeling, in the intelligence that the information contained in the repository is utilized as well as its respective model.

- Chapter 7 displays the full Visual Info Extraction Process applying the Lixto Visual Wrapper and the Lixto Transformation Server. This chapter points out all the conceptual and practical foundation of the Corporate Information Factory which requires the interaction of two levels of Business Intelligence and a Business Intelligence repository to carry out visual analysis by means of information modeling.

- Finally, Chapter 8 is devoted to presenting the highlights of related works found in the literature. Along with the drawing of the final conclusions, an outlook to future research topics is suggested.

Chapter 2

Intelligence Processes through Information

Nowadays, as we are living in a competitive business world, corporations need to be even more competitive than their market counterparts. They require knowing more in detail what is happening with their competitors in the business arena, what kind of strategies they are implementing and what business scope they are considering. In regards to their competitors, corporations wonder and try to answer several matters such as: who their new customers are, what new acquisitions they have, in what new mergers they are involved; and what new strategies they are implementing. In general, what is happening with each of them.

Like a number of management terms, strategy is an overused word that means different things to different people. For the intention of this thesis, strategy is defined as a dynamic process supported by information. Times change; technology changes, markets change, data change, information changes and rules of competition change. Consequently, never can strategy stay still.

Michael Porter [80] stated that a company can improve its performance in regards to that of its rivals only if it can establish a difference of value to customers that can be preserved over time. On the other hand, economists refer to this as competencies. Competencies are generated from the information analysis and knowledge of customers and competitors. Competencies have also been defined as the organization's resources and capabilities, unmatched by competitors and they explain the origin of the economic value generated by a company.

A competitive advantage is the strategic way the organization is positioned in the market. To obtain an advantage over its competitors, a corporation needs to analyze all the information related to their market counterparts to try to identify such competitive advantage [78].

Organizations that identify oppotunities to create conditions of imbalance can allow the company to claim economic rewards beyond those resulting from conditions of perfect competition, and then to sustain and protect those conditions as long as possible, obtaining, protecting and exploiting strategic information.

2.1 The Competitive Advantage

Strategic planning is the business management process through which an organization defines and clarifies what their cutting edge will be (See Figure 2.1). Through strategic planning, an organization fully specifies vision, mission, goals and courses of action. Successful strategic planning entails having updated information related to competitors, markets, and products, among others.

Strategic planning is also part of a large process that has demonstrated generating empowerment within the organization. Strategic management is a way of leading the organization with the ultimate objective of developing values, managerial capabilities, organizational responsibilities, and administrative systems that link strategic and operational decision making processes at all hierarchical levels, and across all lines of authority (see Figure 2.2.)

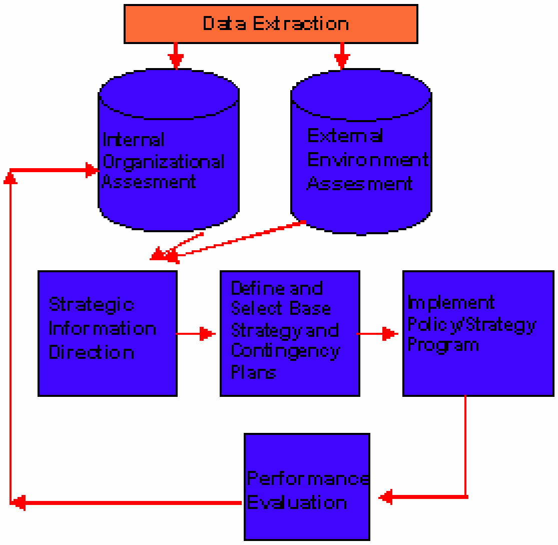
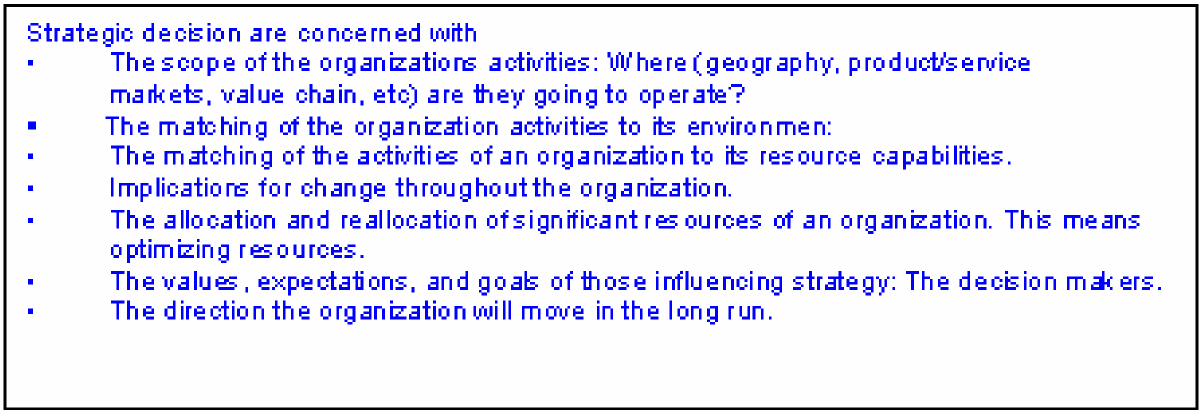

Data Extraction
Internal Organizational Assesment
External Environment Assesment
Strategic Information Direction
Define and Select Base Strategy and Contingency Plans
Implement Policy/Strategy Program
Performance Evaluation

Figure 2.1 Strategic planning process through information

The strategic management framework is supposed to help decision makers understand the set of elements that must be together in order to make effective strategic decisions, sustained with information from the business community such as financial reports from the Internet websites and also inside the corporation, internal databases and repositories (see table 2.1).

Table 2.1 Some elements of strategic decision sustained with information



Strategic decision are concerned with

- The scope of the organizations activities: Where (geography, product/service markets, value chain, etc) are they going to operate?
- The matching of the organization activities to its environmen:
- The matching of the activities of an organization to its resource capabilities.
- Implications for change throughout the organization.
- The allocation and reallocation of significant resources of an organization. This means optimizing resources.
- The values, expectations, and goals of those influencing strategy: The decision makers.
- The direction the organization will move in the long run.

The decision makers will always attempt to understand how to position the organization. The process to generate this form of understanding supports the competitive intelligence “process”: Competitive intelligence decisions should help the organization be more effective in its strategic actions, they should have larger implications for the competitiveness of the company, with a greater impact on subsequent decisions.

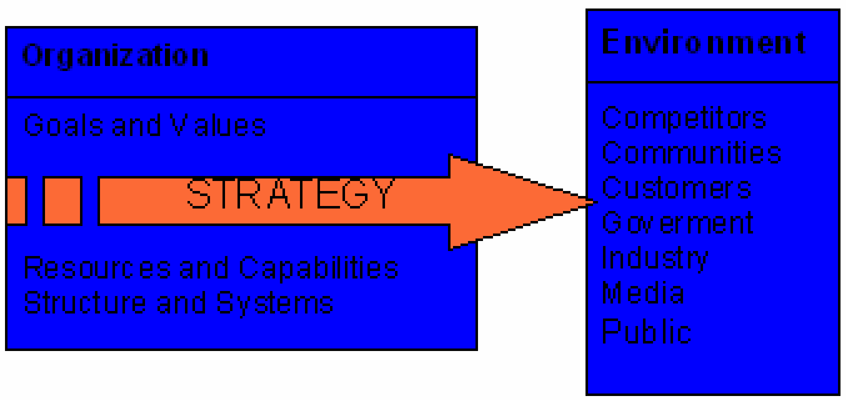


Figure 2.2 A generic strategic management and information requirements for BI.

2.2 Strategic Thinking

Some experts, including Henry Mintzberg [69], proposed that strategic planning, needs to be replaced by a capacity of strategic thinking. What is the difference between these concepts and why should decision makers take care of strategic thinking? And, how will strategic thinking improve the process of business intelligence?

In this so competed world, strategic planning in corporations impedes dynamic and innovative executive decisions and requires a more strategic business action. While many organizations are developing their strategic plans, their competitors are out in the field executing their strategies designed to win larger market slices around the world. Strategic planning also suggests the need of management's strategic thinking. What, then, is strategic thinking through information? Table 2.2 provides a useful overview.

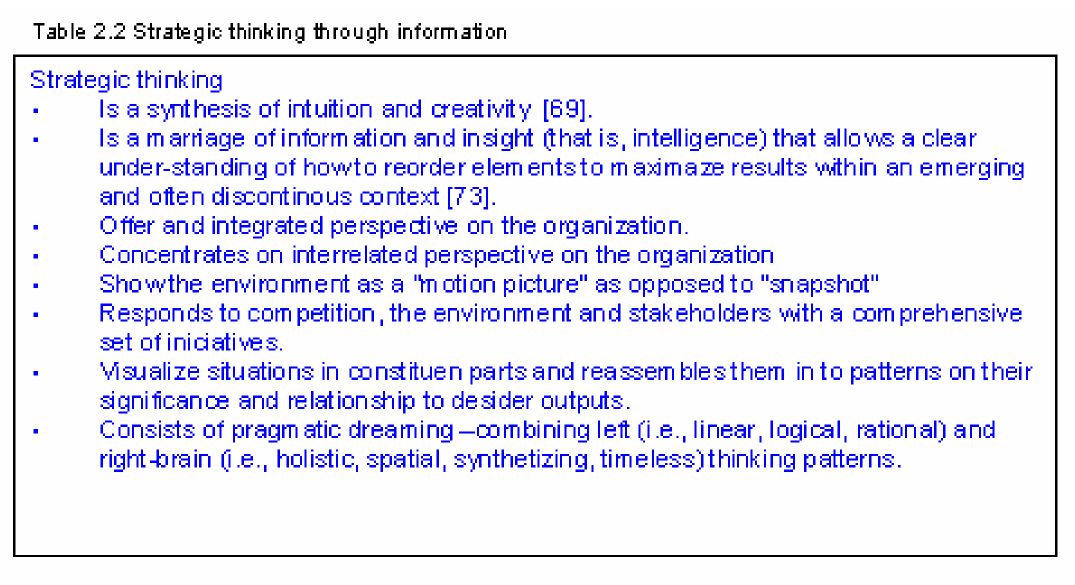


Table 2.2 Strategic thinking through information

Strategic thinking

- Is a synthesis of intuition and creativity [69].
- Is a marriage of information and insight (that is, intelligence) that allows a clear under-standing of how to reorder elements to maximaze results within an emerging and often discontinous context [73].
- Offer and integrated perspective on the organization.
- Concentrates on interrelated perspective on the organization
- Show the environment as a "motion picture" as opposed to "snapshot"
- Responds to competition, the environment and stakeholders with a comprehensive set of iniciatives.
- Visualize situations in constituen parts and reassembles them in to patterns on their significance and relationship to desider outputs.
- Consists of pragmatic dreaming – combining left (i.e., linear, logical, rational) and right-brain (i.e., holistic, spatial, synthetizing, timeless) thinking patterns.

2.3 The Competitive Intelligences Process

While strategy and planning can define the direction the organization will move to, intelligence identifies the horizon and allows the organization to get benefits. Thus, it is of fundamental importance the quality of the competitors' information available on websites for analysis.

At present, competitive intelligence (CI) -in general- is often viewed as a relatively new discipline, dating from 1980. It grew out of developments in economics, marketing, military theory, strategic management, and information systems and it has kept evolving as a separate function within the organizations.

In this work, intelligence is defined as the added value to products or activities, resulting from the collection, evaluation, analysis, integration, and interpretation of all available information (internal & external) that allows to cover one or more aspects of an executive's needs, and that is immediately or potentially significant to decision making. This definition of intelligence serves for two useful purposes in this thesis:

1. It differentiates intelligence from information (unevaluated material).
2. It captures the dynamic, cyclical nature of intelligence supported by information technologies.

Intelligence is necessary to reduce uncertainty and risk in making decisions within the organization. Often, the opinion available to an organization will depend on how early problems are identified after the information has been analyzed. Business Intelligence brings together both the threats and the opportunities coming from the environment's outside elements which could have an impact on the organization's present and future competitiveness. In this work, business intelligence focuses on a) a systematic process or cycle (see figure 2.3) from collecting and analyzing information of competitors' activities, b) one's business environment and c) the business trends facing the future in accordance with the organizational goals.

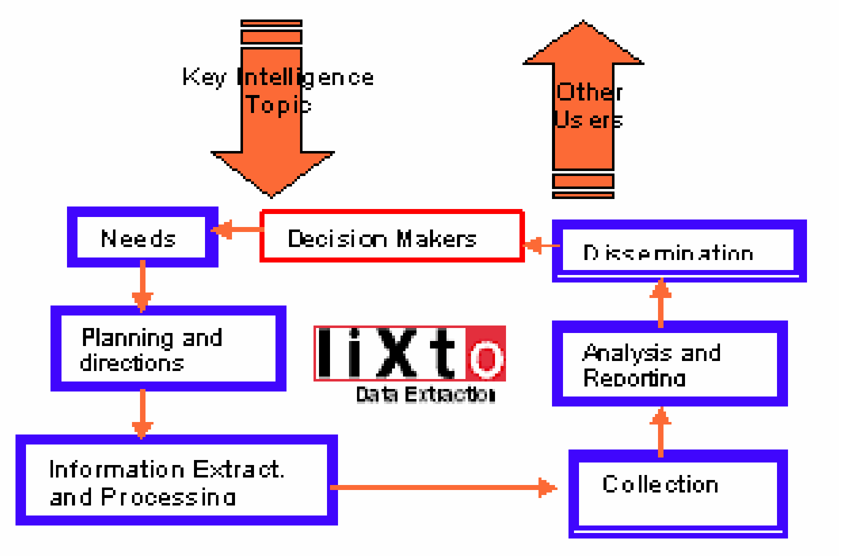


Figure 2.3 Analysis, strategy, and business intelligence cycle

The main purpose of carrying out competitive analysis is to better understand the industry and its competitors in order to make decisions and develop strategies that provide a competitive advantage in the short, medium and long terms. Such advantage must reflect on the continuity of superior results than those of the competitors'. The outcomes of the analyses should be suitable to set in action, that is, they ought to be oriented to meet future goals, to help decision-makers develop better strategies in the aggresive competition, to procure a better understanding of the market over the one competitors have and to identify present and coming competitors, their plans and their strategies.

The ultimate objective of this kind of intelligence process is to produce better business results. Extracting these results from competitive analysis has become a more important step of competitiveness in recent years because of the reasons explained below.

2.3.1 Competition

The global world has outgrown the level of competition in the majority of marketplaces. In the past, a competitor could sustain marketplace advantages by being in the right place at the right time. Geographic, physical and sociopolitical barriers served to keep competitors out of many marketplaces. Most of these barriers are falling as a consequence of the progress made in the communication systems , trade policy, technology, information technologies and the World Wide Web (see figure 2.4). New competitors quickly appear and use the information technologies in an intensive way, just right in the place and at the time the marketplace barriers fall.

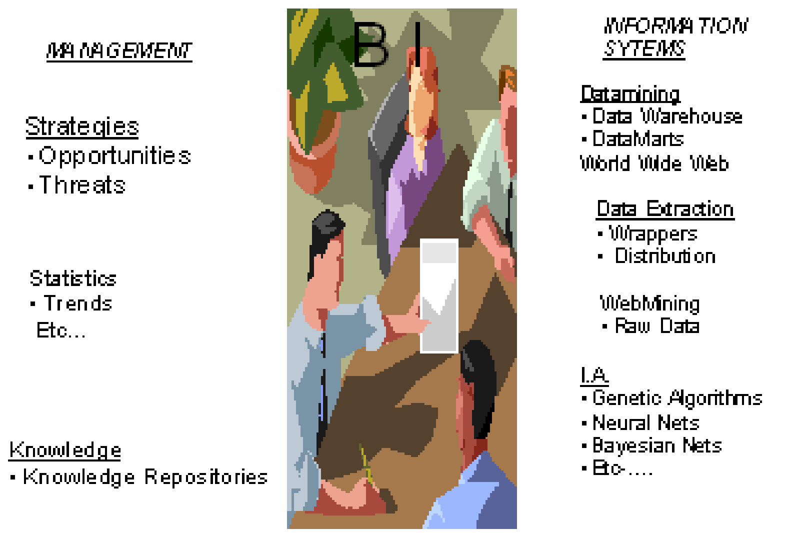


Figure 2.4 Business intelligences as a core competences between information systems and management

These new competitors will use the information technologies as a strategic tool and they may compete very differently from existing competitors. They are more oriented to use information technologies to support their decisions about customer demands and to know more about their competitors' strategies. Sometimes, the form of competition may not even appear logical, insightful, or even ethical. Because of this new global competition, they need to thoroughly understand competitors and the importance of the growth of business contexts if they want to exist in the short future.

2.3.2 Knowledge Economy

The global economy is increasingly being characterized as a knowledge economy. A paradigm shift has occurred as we move farther away from the industrial economy paradigm that dominated mostly during the last two centuries. As opposed to tangible things, services and related intangibles now constitute the large part of the GDP (Gross Domestic Product) in most of the leading economies, and services are more knowledge-based and information sustained than material-based.

Companies amass data and information, but seldom do they recognize that knowledge is not the same thing as information. Sustaining a competitive advantage requires companies not only to apply data, analyze information and transform it so as to create order out of complexity and uncertainty, but also to leverage and transfer knowledge.

2.3.3 Increased Limitation

The new economy is characterized by increasing reproducibility, by means of which corporations have greater ability than ever to quickly replicate most facets of a competitor's new product, strategies or service offering. Some companies succeed by being "quick seconds" in the market place and stressing their competitive abilities at providing an improved product or service based on the innovation of the originator.

2.3.4 Complexity and Speed.

There is increasing complexity and speed in business and information technologies. Underlying the changing market place are communication and information technologies that allows for the transfer of data at rates faster than ever before. However, human ability to process data remains essentially stable.

Intelligence and analysis must be user-driven to be of enduring value. This is fundamental to know your customers and competitors. At the same time, it is very important to integrate the major IT applications in business being the bull eye's target more aggressive business intelligence strategies.

For the intention of this thesis, the literature reviewed focuses on five basic types of intelligence analysis.

1.- Forseeing unexpected events to provide timely notice to the organization.
2.- Supporting the decision-making process.
3.- Competitor assessment and monitoring.
4.- Intelligent assessment and monitoring.
5.- Comprising a key part of collecting and reporting

Data extraction will support the third type of intelligence analysis, competitor assessment and monitoring. Inside the corporations, the use of wrappers will provide great support to the intelligence task which is backed up by information technologies in the executive, management, knowledge and transaction systems (see figure 2.5).

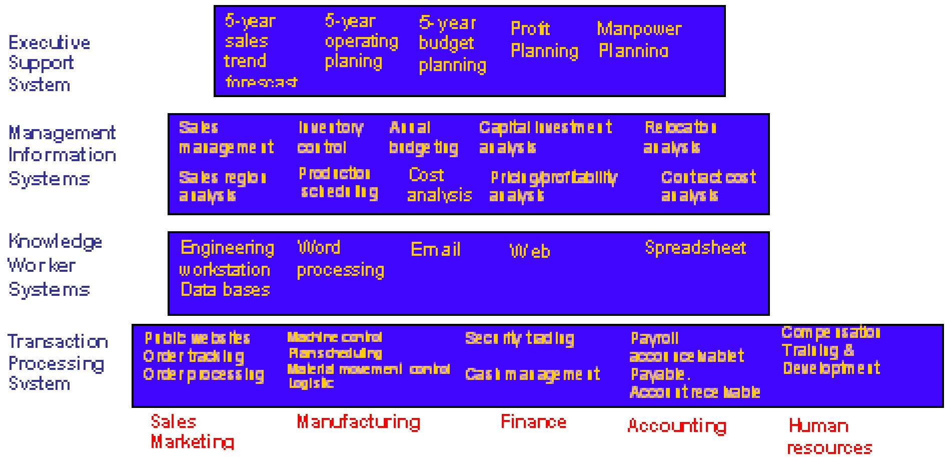


Figure 2.5 Major business IT applications in business

Good competitive and strategic intelligence requires effective analysis. This requires a profound comprehension of environments, industries and organizations. Effective analysis can be supplied only by the experienced who is not only able to get solid data and information, but who is also keen at using analytical techniques and institutional models of competition in the business field.

Chapter 3

Competitors’ Radiography

A competitor’s radiography provides a comprehensive image of the strengths (bones) and weaknesses (nerves) of current and potential rivals. This analysis provides both offensive and defensive strategic contexts through which opportunities and threats can be identified. A competitor’s radiography combines all the relevant sources of competitors analysis into a systems than can support efficient and effective strategy.

The competitor’s radiography will be used for four major purposes: to identify the competitor’s future strategies and plans; to predict its most likely reactions to competitive initiatives; to determine how well-matched its strategy is in accordance to its present capabilities; and to understand its weaknesses.

3.1 Porter’s Analysis of Competitors

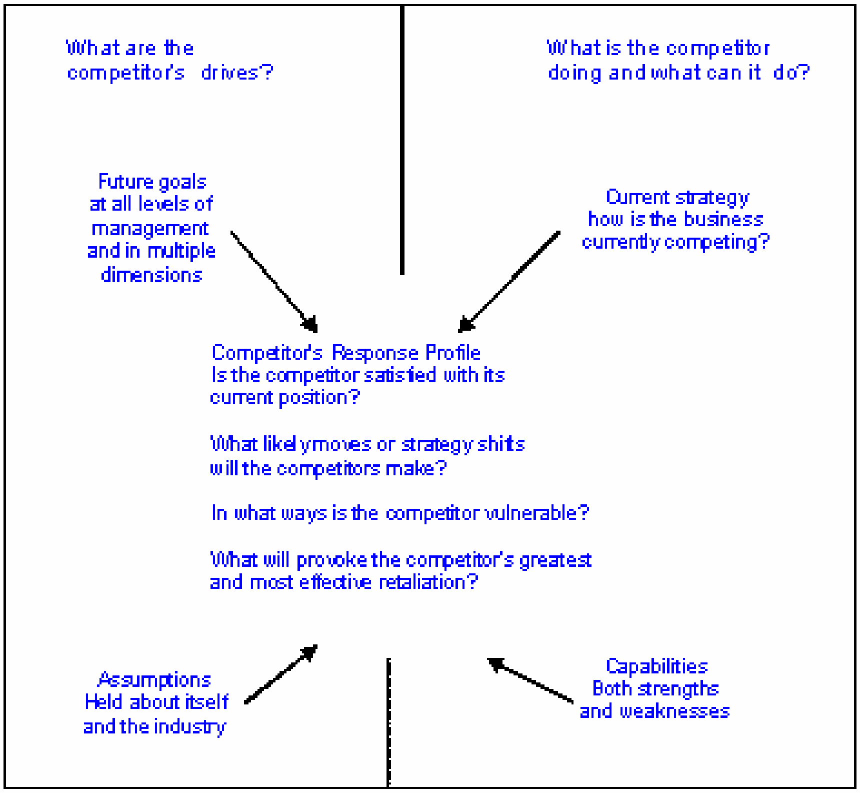
According to Porter, the analysis or radiography of competitors is an essential component of corporate strategy, Porter [79] said that most firms do not conduct this type of analysis as an organized activity. A lot of firms operate on what he calls informal intuition, acquired through the conjectures of information about competitors managers continuously receive from external and internal sources.

Porter was one of the first writers to propose a formal process to gather information about competitors and to get the benefits from the analysis and strategies generated of this information. His model is described in Figure 3.1 and Table 3.1.

The forward-looking stance of a competitor's radiography is intentional since its primary objective is to predict the future strategic actions of rivals and competitors. That is, the model motivates the analyst to use current and past information about rivals and businesses in order to predict at least three concerns: the future strategic moves that a competitor may pursue in response to the firm's own strategies; the strategies of other firms in the industry; or the transformation of the competitive environment into business strategies. Starting with this knowledge, the analyst will be in a better position to confront both defensive and offensive strategies.

Unfortunately, nowadays a large number of firms do not formally generate the radiographies of their competitors or their rivals' profiles with the systematic rigor required by the analysis of competitors' profiling. In a recent research work of nine major studies investigating current practices of competitors profiling within U.S. companies, Ram and Samir [83] came to an interesting conclusion.

Figure 3.1 The components of a competitors radiography


What are the competitor's drives?
What is the competitor doing and what can it do?
Future goals at all levels of management and in multiple dimensions
Current strategy how is the business currently competing?
Competitor's Response Profile
Is the competitor satisfied with its current position?
What likel ymoves or strategy shits will the competitors make?
In what ways is the competitor vulnerable?
What will provoke the competitor's greatest and most effective retaliation?
Assumptions Held about itself and the industry
Capabilities Both strengths and weaknesses

Source: Competiting Strategy: Techniques for Analyzing Industries and Competitors, by Michael E. Porter, 1988 by The Free Press.

The 1988 Conference Board Study found that only 3 percent of the firms analyzed had fully developed systematic competitor profiling systems in the corporation. However, it also indicated that 24 percent of those analyzed firms had a fully functional, advanced competitor profiling program.

With this study, we will see that Western firms are starting to learn some valuable lesson from their East Asian counterparts who have benefited from advanced competitor profiling for at least a decade if not longer. At the same time we know that the Asian companies have a lot of tradition in corporate benchmarking as a parameter to measure their productivity and competitiveness. In Latin America corporations do not apply these concepts as a strategic tool to achieve competitiveness; most of them only compete in a local arena and the corporation lives mainly with rumors of the industry and rivals, as it happens with the Cemex (Cementos Mexicanos) case .

Superior knowledge of rivals offers a legitimate source of competitive advantage. The real value of the competitive advantage consists in offering superior customer value. Customer value is defined as relative to the rival's offerings, making the competitor's knowledge an intrinsic component of corporate strategy. Profiling facilitates this strategic objective in three important ways:

- First, profiling can reveal strategic weaknesses in rivals that the firm may exploit.

- Second, the proactive position of competitor's profiling will allow the firm to anticipate the strategic response of their rivals to the firm's planned strategies, the strategies of other competing firms, and the changes in the environment.

- Third, this proactive knowledge will give the firms strategic agility. Offensive strategy can be implemented more quickly in order to exploit opportunities and capitalize on strengths.

Clearly, those firms practicing systematic and advanced competitors' radiographies have a significant advantage. As such, a comprehensive profiling capability is rapidly becoming a core competence required for successful competition. An appropriate analogy is to consider this advantage similar to having a good idea of the next move that your chess match opponent will make. By staying one step ahead, the checkmate is one step closer. Indeed, as in chess, a good offense is the best defense in the business game.

3.2 Competitors’ Radiography Advantages

Given the close relationship among a competitor’s radiography, the information technologies and the competitive advantage, many of the strengths of this tool are evident. However, several unique benefits arise. First, profiling encourages the firm to adopt a confident, aggressive, and proactive stance toward competitive strategy. The knowledge of rivals that profiling grants allows firms to define the parameters of strategy rather than to react to unexpected competitive sideswipes. Done well, a competitor’s radiography or analysis will become one of the firm’s core competencies, contributing to competitive advantage. Part of the objective of this work is to show how by implementing a good business intelligence process from competitors, Cemex will obtain business benefits.

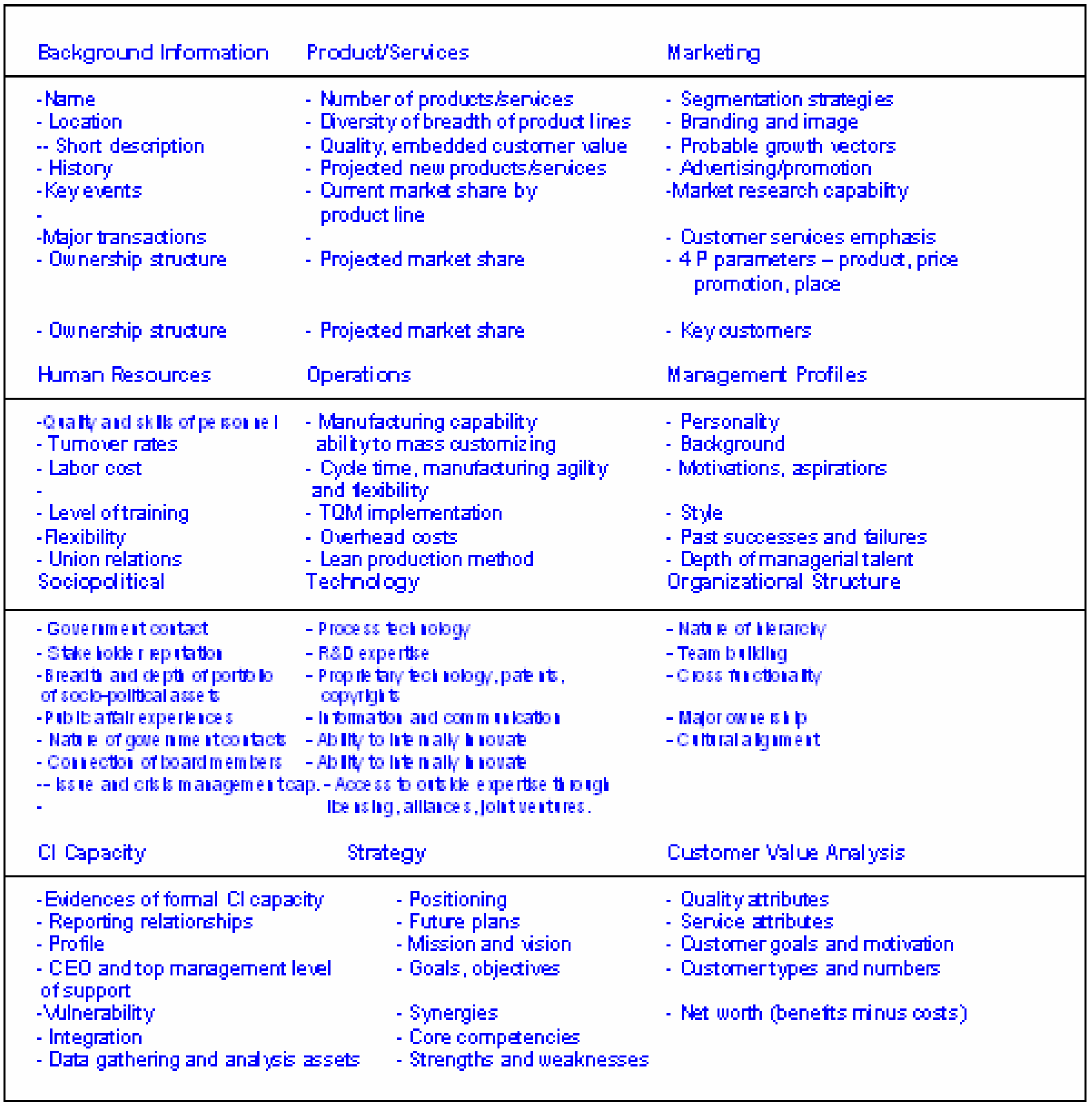
A competitor’s radiography greatly facilitates sharing insights and perspectives across traditional functional boundaries of the firm. This allows for the articulation of many unique opportunities that would have remained hidden in the absence of a formal profiling approach. Often, the process of competitors radiography acts as a rallying point for many of the firm’s employees and departments. It will start the synergy for sharing information between departments and to consolidate a unique intelligence repository as it is proposed for Cemex.

The implementation of a competitor’s radiography systematic process also creates an efficient approach to the formulation of strategies. The analytical product of a competitor’s radiography in the form of meaningful, timely, concise, and visually accessible format is an excellent vehicle to communicate the relevant factors for future strategies.

3.3 Limitations of a Competitor’s Radiography

In the attempt to execute a competitor’s radiography and in the pursue of becoming an industry leader, the firm will eventually become a follower, if it defines leadership as being closely related to current rivals. Comparing itself to rivals must always relate to the notion of customer value. Another strategy in the case of the cement industry is to make reference to the firm’s strategy to face rivals based on the idea of strategic groups or interindustry competition. This will eventually blindfold the firm to innovative approaches of potential rivals in delivering superior customer value from outside the industry. This underscores the importance of keeping an eye on potential rivals from seemingly unrelated sectors and industries in order to root out complacency and blind spots that seem to grow unstoppably. This is just too common in the cement and construction industry in which new applications and industries appear frequently.

Table 3.1 Typical categories and types of competitor radiography information



Background Information	Product/Services	Marketing
-Name - Location -- Short description - History -Key events - -Major transactions - Ownership structure - Ownership structure	- Number of products/services - Diversity of breadth of product lines - Quality, embedded customer value - Projected new products/services - Current market share by product line - - Projected market share - Projected market share	- Segmentation strategies - Branding and image - Probable growth vectors - Advertising/promotion -Market research capabilty - Customer services emphasis - 4 P parameters – product, price promotion, place - Key customers
Human Resources	Operations	Management Profiles
-Quality and skills of personnel - Turnover rates - Labor cost - - Level of training -Flexibility - Union relations	- Manufacturing capability ability to mass customizing - Cycle time, manufacturing agility and flexibility - TQM implementation - Overhead costs - Lean production method	- Personality - Background - Motivations, aspirations - Style - Past successes and failures - Depth of managerial talent
Sociopolitical	Technology	Organizational Structure
- Government contact - Stakeholder reputation - Breadth and depth of portfolio of socio-political assets - Public affair experiences - Nature of government contacts - Connection of board members -- Issue and crisis management cap. -	- Process technology - R&D expertise - Proprietary technology, patents, copyrights - Information and communication - Ability to internally innovate - Ability to internally innovate - Access to outside expertise through licensing, alliances, joint ventures.	- Nature of hierarchy - Team building - Cross functionality - Major ownership - Cultural alignment
CI Capacity	Strategy	Customer Value Analysis
-Evidences of formal CI capacity - Reporting relationships - Profile - CEO and top management level of support -Vulnerability - Integration - Data gathering and analysis assets	- Positioning - Future plans - Mission and vision - Goals, objectives - Synergies - Core competencies - Strengths and weaknesses	- Quality attributes - Service attributes - Customer goals and motivation - Customer types and numbers - Net worth (benefits minus costs)

3.4 A Competitor's Radiography

The process of making a competitor's radiography for Cemex is made up of nine stages to be defined:

1. Who are its competitors?
2. Who could its potential competitors be?
3. What information about these rivals is required?

4. Build a competitor analysis capability in order to secure this information.
5. Conduct a strategic analysis of the information gathered.
6. Present the information in an accessible format (graphs, charts. images, etc).
7. Ensure that the right decision makers get the right information opportunely.
8. Develop strategies based on the information analysis.
9. Monitor rivals and scan for potential rivals continuously. (It must be a never ending process.)

3.4.1 Who Are Cemex's Competitors?

The first steps are very closely related. Determining the current Cemex competitors is usually quite obvious upon cursory observation. A typical criterion includes those firms that serve the same customer base. However, upon deeper analysis this distinction will probably become diffused. What exactly is the customer base? Does it refer to customers of the same product? Is it the same product category? Ultimately, all firms are rivals in the sense that they are all trying to attract the same discretionary income. In the case of the cement industry a competitor is defined as the company that is providing the same cement products in the same business arena.

Given the industry migration and value chain erosion that is commonplace nowadays, this critical assumption is important to the cement industry (focus of this thesis) at the onset in order to prevent the analysis from becoming too narrowly focused.

There are two very distinct ways to define competitors. The traditional method has taken a supply side perspective centered on defining strategic groups. Strategic groups are closely related firms that are usually premised on relatively similar strategies, use similar links on the industry's value chain, and share similar resource capabilities. As such, this method is more adept at identifying current competitors within strategic groups or within industries.

Operationally, strategic groups can be plotted on a two-axe graph into which all of the traditional rivals are represented on the space relative to some sets of distinctions along each axis. Next, the rivals occupying niches within the market are plotted. The final plot locates those potential rivals that are currently operating on the fringe of the industry's conventional parameters such as firms employing substitutable technology, suppliers who may choose to be integrated, and customers who may eventually choose to be reintegrated. At a minimum, every member of the industry's value chain should have been included when the strategic map is finished. It is important to be as creative as possible at this point because potential rivals are often not that evident.

3.4.2 Who Could the Potential Competitors Be?

Other methods of defining markets explicitly address the fact that potential rivals are evident. However, they are hidden and are actually developing innovative ways of delivering customer value on new and competitive platforms. Also, the current leading companies are oblivious of these competitors who could take over in the market by focusing on customer value.

Which companies do your customers see as being your major competition? Firms can define potential rivals according to their provision of comparable customer value through different platforms. Here, the focus of the analysis is on defining potential rivals based on changing customers' preferences, motivations, deployment of products or services and/ or technological innovation.

Generally, the most valuable sources of information regarding the identification of both current and potential competitors will be the firm's customers, sales staff, marketers, operations managers and Internet web pages. Other, less valuable sources may be found in industry directories, trade associations, and other secondary resources on the Internet.

3.4.3 What Information About These Rivals is Required?

Those in charge of making decisions in the firm will have all the elements required to profile accurately what kind of information regarding competition is best for the analysis. It is important to ensure that the competitive intelligence (CI) efforts around the competitor's radiography are user-oriented. To facilitate this objective, the efforts in gathering information should be demand-oriented from the beginning The decision makers must determine what kind of information they need for their analysis and intelligence processes.

It is very important that the information requirements are closely related to the internal demand of CI. It is important to keep the information gathering activities relevant to the external demand of those competitive parameters that will impact on the future customers values. Table 3.1 describes the types and categories of information that could be considered during this stage.
Besides this main requirement, helpful ideas can be gathered from the various CI benchmarking initiatives that have been concluded. An excellent summary of nine major studies can be found in the Ram and Samir [83] study cited previously. For

example, a study conducted by the Conferences Board asked 308 responding companies to rate the most useful types of information (see tables 2.2 and 2.3). This may give the analyst some ideas regarding potentially useful types of information that can change over time.

3.4.4 Analysis Capability

The concept of the intelligence cycle is applicable to Competitive Intelligence capability. The structure of the CI system can include competencies based on four distinct organizational skills: collection, processing, analysis, and dissemination. An important item to keep in mind is the fact that, contrary to intuition, most of the information required already exists inside the firm. Sales people, marketing, staff, operations, Internet, everyone and everything in the corporation is probably in possession of valuable strategic information of competition. Figuring prominently within these primary sources of information are the firm's customers and suppliers. Tables 3.4 and 3.5 rank the common sources of information of competitors.

3.4.5 Gathered Information

Porter's framework depicted in figure 3.1 can be used as a helpful guide to carry out the analysis of the gathered information.

3.4.5.1 Future Goals.

The future goals of rivals will help Cemex forecast its competitors' strategies to face its counterparts as well as the firm's analyst's designed strategies. To know where a competitor is headed, it is important to learn about the direction they are implementing to increase their share of market, their profitability, and the organization's performance

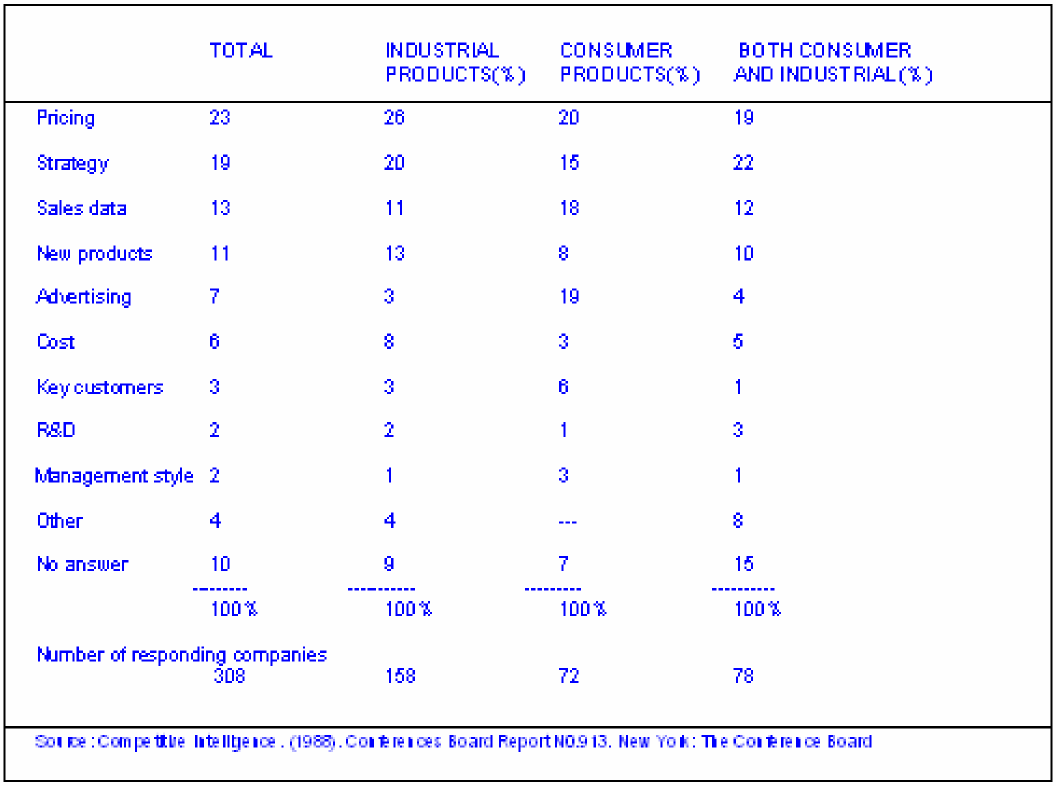
3.4.5.2 Current Strategy

In trying to infer the competitor's current strategy, it is very important to determine which of the four generic strategies (differentiation, low cost, market or focus) the firm is pursuing . The next step is that the analysis be used to outline the strategic implications for each functional area of the rival's business.

A competitor's current strategy may be identified on the basis of what it is doing at present, what the firm says to the newspapers, and what financial reports they are publishing in the mass media.

What are its stated short-term goals? Start by identifying the differences between future goals and what it is currently doing. Are its short-term activities in line with its future goals? Remember, in the absence of particular forces for change, it can be assumed that a company will continue to compete in the future in the same way it has competed in the past.

Table 3.2 Most useful type of information (by type of market)



	TOTAL	INDUSTRIAL PRODUCTS(%)	CONSUMER PRODUCTS(%)	BOTH CONSUMER AND INDUSTRIAL(%)
Pricing	23	26	20	19
Strategy	19	20	15	22
Sales data	13	11	18	12
New products	11	13	8	10
Advertising	7	3	19	4
Cost	6	8	3	5
Key customers	3	3	6	1
R&D	2	2	1	3
Management style	2	1	3	1
Other	4	4	---	8
No answer	10	9	7	15
	100%	100%	100%	100%
Number of responding companies	308	158	72	78

Source :Competitive Intelligence . (1988). Conferences Board Report NO.913. New York: The Conference Board

3.4.5.3 Capabilities.

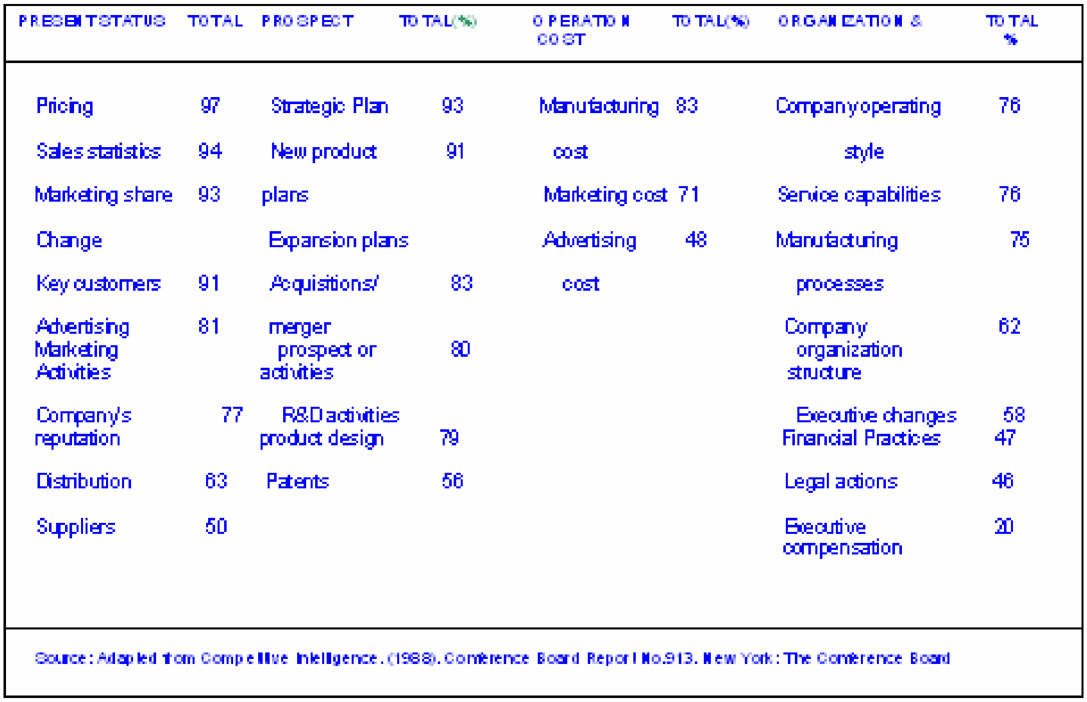
The collected information will serve well to lead a SWOT analysis (strengths, weaknessed opportunities, and threats) of each competitor. The main objective is to identify what the competitor is doing and what it is planning to do. Although a competitor may announce its strategic intentions, these may be different from its current capabilities. Thus, the analyst must pose to himself and answer some questions about how the company is going to make ends meet.

3.4.5.4 Assumptions.

The industry, as well as rivals, will yield many useful insights regarding any potential incorrect assumptions. Frequently do these assumptions supply competitive opportunities in the short run. What assumptions does the competitor hold about its world? Are these reflected on its strategies, both current and future?

Assumptions can be identified by the inconsistencies between capabilities, current strategies and future goals. On the other hand, a company that has all three areas (capabilities, strategies and goals) in sync may be an unbeatable competitor. However, all companies hold assumptions about the world and the future; thus, these assumptions need to be uncovered.

Table 3.3 Most types of information rated as useful or fairly useful



PRESENT STATUS	TOTAL	PROSPECT	TOTAL(%)	OPERATION COST	TOTAL(%)	ORGANIZATION &	TOTAL %
Pricing	97	Strategic Plan	93	Manufacturing	83	Company operating	76
Sales statistics	94	New product	91	cost		style	
Marketing share	93	plans		Marketing cost	71	Service capabilities	76
Change		Expansion plans		Advertising	48	Manufacturing	75
Key customers	91	Acquisitions/	83	cost		processes	
Advertising Marketing Activities	81	merger prospect or activities	80			Company organization structure	62
Company's reputation	77	R&D activities product design	79			Executive changes Financial Practices	58 47
Distribution	63	Patents	56			Legal actions	46
Suppliers	50					Executive compensation	20

Source: Adapted from Competitive Intelligence. (1988). Conference Board Report No.913. New York: The Conference Board

The most critical issue underlying the competitor's radiography is understanding the key assumptions made by the competitors' management team. This identifies fundamental weaknesses in how they compete and it also provides a framework on how they see their marketplace. Answering questions such as "Are they satisfied with this position in the market place?" "What are their vulnerabilities?" "What are their future expansion plans?" can provide the necessary strategic input and understanding to take on competitors .

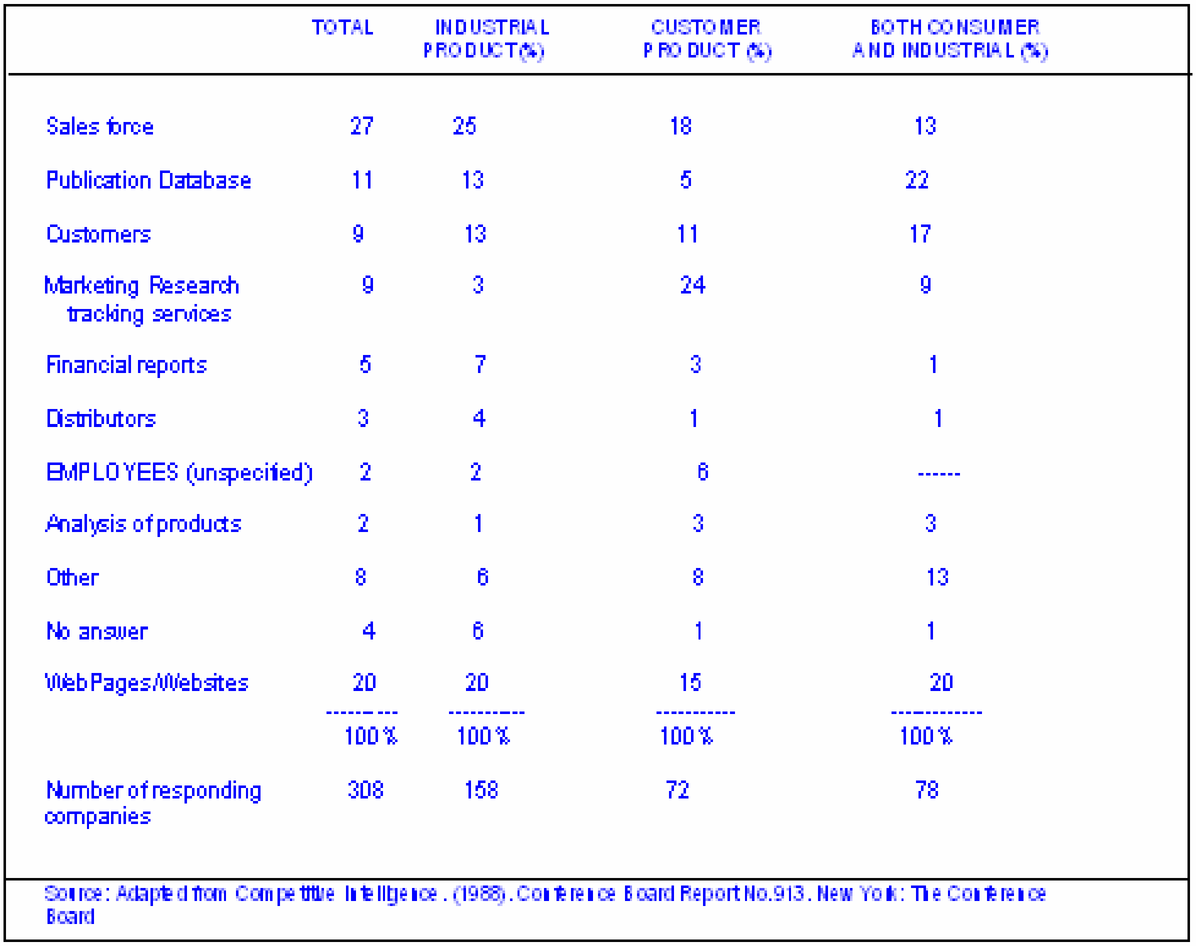
All the previous analyses will be integrated into a competitor's radiography report. The purpose of this integration is to forecast with reasonable accuracy what a rival will bring forward or how it will respond to various competition pressures in the business arena:

1.- The offensive status of rivals helps predict any proactive moves they will take.

2.- The defensive status of rivals helps forecast how a rival will react to various competition pressures.

It is important to point out that for the process of making a competitor's radiography the qualitative factors of the information are often more predominant than those of a more traditional quantitative approach of business analysis. Sometimes the statistics are also relevant to complement the analysis, though.

Table 3.4 Most useful sources of information



	TOTAL	INDUSTRIAL PRODUCT(%)	CUSTOMER PRODUCT (%)	BOTH CONSUMER AND INDUSTRIAL (%)
Sales force	27	25	18	13
Publication Database	11	13	5	22
Customers	9	13	11	17
Marketing Research tracking services	9	3	24	9
Financial reports	5	7	3	1
Distributors	3	4	1	1
EMPLOYEES (unspecified)	2	2	6	------
Analysis of products	2	1	3	3
Other	8	6	8	13
No answer	4	6	1	1
Web Pages/Websites	20	20	15	20
	100 %	100 %	100 %	100 %
Number of responding companies	308	158	72	78

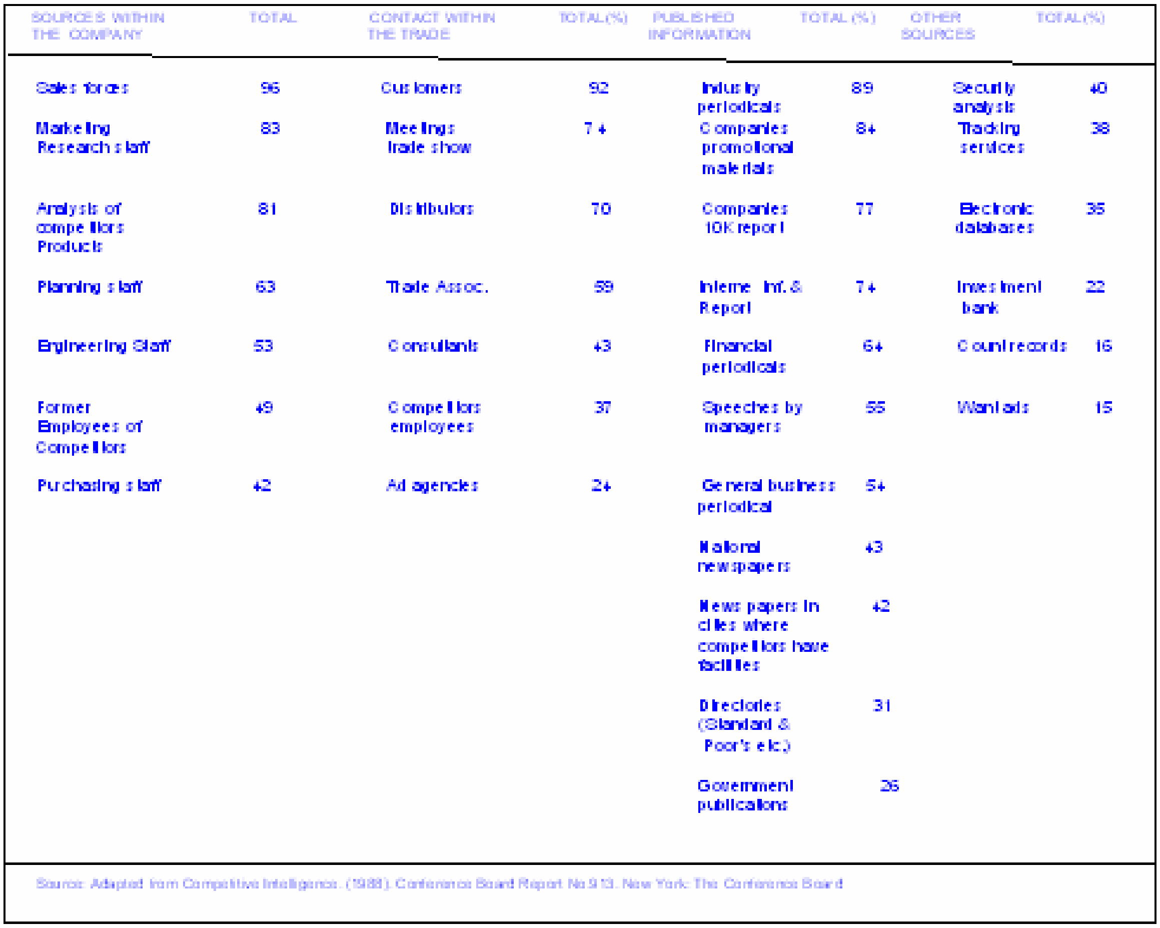
Source: Adapted from Competitive Intelligence. (1988). Conference Board Report No.913. New York: The Conference Board

3.4.6 Information in an Accessible Format

There are different formats through which to communicate the analysis of the competitor’s radiography. Most of them are with visual-oriented as opposed to written reports (hard copy).

3.4.6.1 Comparison Grids.

For positioning, the plot rival position will be used. It includes performance, capabilities, key success factors, and location, among others. All of them will be represented on high/low dependent and independent variable cross-line axes. Depending on the application, the referent points are either the firm’s performance figures or the industry averages. Comparison grids provide nice snapshots of the relative performance across two competitive parameters. (See figure 3.2).



SOURCES WITHIN THE COMPANY	TOTAL	CONTACT WITHIN THE TRADE	TOTAL(%)	PUBLISHED INFORMATION	TOTAL (%)	OTHER SOURCES	TOTAL(%)
Sales forces	96	Customers	92	Industry periodicals	89	Security analysis	40
Marketing Research staff	83	Meetings trade show	74	Companies promotional materials	84	Tracking services	38
Analysis of competitors Products	81	Distributors	70	Companies 10K report	77	Electronic databases	35
Planning staff	63	Trade Assoc.	59	Interne Inf. & Report	74	Investment bank	22
Engineering Staff	53	Consultants	43	Financial periodicals	64	Count records	16
Former Employees of Competitors	49	Competitors employees	37	Speeches by managers	55	Want ads	15
Purchasing staff	42	Ad agencies	24	General business periodical	54		
				National newspapers	43		
				News papers in cities where competitors have facilities	42		
				Directories (Standard & Poor's etc.)	31		
				Government publications	26		

Source: Adapted from Competitive Intelligence. (1988). Conference Board Report No.913. New York: The Conference Board

Table 3.5 Information sources rated very important or fairly important

3.4.6.2 Radar Chart.

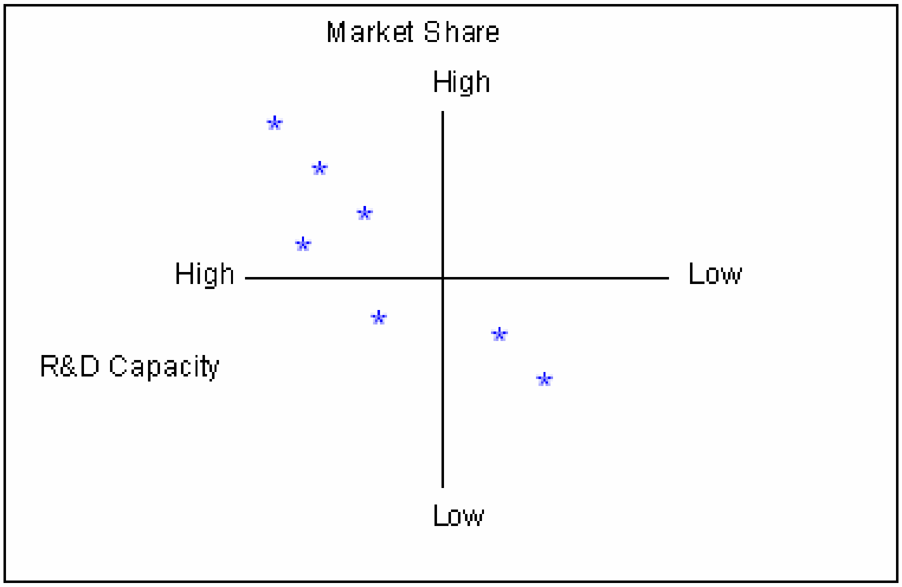
Radar charts are often used to communicate profiling or radiographic analysis. Radar charts are composed of an underlying circle with several points on the circumference representing the industry average around relative competitive parameters.

Superimposed over this circle are geometric shapes representing the performance of the firm or rivals under analysis. Depending on superior or inferior performance, the resulting geometric shape of the overlay will describe a concise visual of relative performance.

3.4.6.3 Color-Coded Competitor's Strength Grid.

As described by Aaker [1], in order to characterize how rival firms ourgrow others in a number of parameters, the strength grids and maps consitute a powerful and simple way to go by. By assigning a spectrum of colors to represent relative competitive inferiority, parity, and superiority, the graph will effectively describe the spectrum of relative competitive advantages among rivals.

Figure 3.2 Competitors' Comparison Grid



3.4.7 Decision Makers Get the Right Information on Time

As it is well known, not only is the world changing very fast, but also competitiveness. Therefore, competitor’s intelligence is only worthwhile if the relevant strategic decision maker receives it opportunely. In this respect, timelines and relevance must match complete accuracy.

3.4.8 Strategy Based on the Analysis

At this point of the analysis, competitors’ radiographies are used to develop strategies around several relevant considerations of competitiveness such as determining the probable rules of engagement within the strategic position; and choosing the business arena – where, how, and against whom the firm will compete by developing a strategy that leverages the firm’s strengths, exploits rivals’ weaknesses, neutralizes competitive threats, and defends against own weaknesses.

3.4.9 Continuously Monitoring Rivals

The decision makers and analysts should always assume that competitors are simultaneously performing similar competitors’ radiographies for their own firms, using strategic information at present available on the internet (the most important database in the world) or other information sources. Volatile markets, hypercompetition and industry migration give ample reasons to continuously monitor and update a competitor’s radiography.

During the process of monitoring, the technology associated with Business Intelligence is the marrow of the process of collecting and refining information from many sources. It must be bore in mind that the objective is to analyze and present it in a useful way in order for the corporation insiders to make better business decisions.

The rapid pace of today’s business environment has made the Business Intelligence systems indispensable to an organization success. Identifying trends, analyzing customer preferences and making opportune business decisions are some of the profits of turning a company´s raw data into employable information through Business Intelligence. Over the past few years, business intelligence systems have only been used to understand and address back office needs such as efficiency and productivity. As important as they are, business intelligence systems should be capitalized in terms of facing competition.

Chapter 4

Data Extraction on the Web

The World Wide Web (WWW), also known as the Web, was introduced in 1992 at the Center for European Nuclear Research (CERN) in Switzerland [13]. At the beginning, the World Wide Web started facilitating data-sharing in different forms among physicists at the CERN which today is a repository of information that has revolutionized the information age.

At present the Web is organized as a set of hypertext documents connected by hiperlinks, used in the Hypertext Markup Language (HTML) to enable the links between different web documents. Prior to the Web, the Internet was a massive interconnected network of computers, which was mainly text oriented and used primarily for sharing scientific information (see figure 4.1).

Since the begining of the World Wide Web, the Internet has had an expansive growth in the number of users, and servers and in the amount of information available on it. This information or context includes images, graphs, sound, video and text. The ability and perfomance of the web for collection and dissemination of information has arguably transcended in just a few years.

When automatically extracting information from the World Wide Web, most established methods focus on single HTML documents and other methods focus on Portable Document Format (pdf documents.) However, the problem of analyzing complete web sites has not been utterly explored yet, in spite of its importance for various applications.

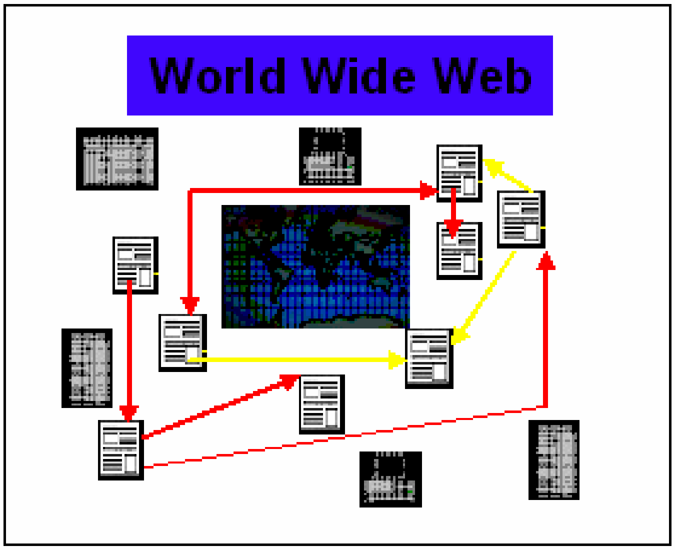


Figure 4.1. The World Wide Web

4.1 World Wide Web Trends

The expansive growth of the Internet has dramatically changed the way in which information is managed and accessed. Actually more people are using the web as a tool to disseminate information and knowledge. What makes the Web so exciting is its power and capacity to cross geographical borders and to bring information of millions of topics in an instant directly to the desktop of people all over the world.

In a recent report about the future of database research known as the Asilomar Report [77], it was predicted that in ten years, the majority of the human information will be available on the Web.

As it can be seen, corporations such as Google [39], Yahoo [107] and others are evolving in an aggressive competition to become the Web market leader. Not only are they doing intensive research work around the Web technologies, but they are also making different technologies available to the internet user for free. Some of these new revolutionary technologies support for example the concepts of the Google book. Apparently, this and other technologies will complement the predictions made in the Asilomar Report[77].

To manage these trends and the complexity of the World Wide Web, three important things have been done: a) traditional information retrieval and wrappers have been applied to the information and document collection on the Internet, b) a set of search engines (Alta Vista[5], Excite [33], Magallan and WebCrawler have been acquired by Excite, HotBot [48], Infoseek [109], Lycos [63], Yahoo [107], Google [39a]) and c) tools have been proposed and implemented. As it can be seen in Figure 4.2, typically, the architecture of a search engine has four major components at two levels. At the software level it has the querying interface and the web crawler. At the hardware level it has the search index the hosting hardware. The arrows indicate the interaction and dynamism between components.

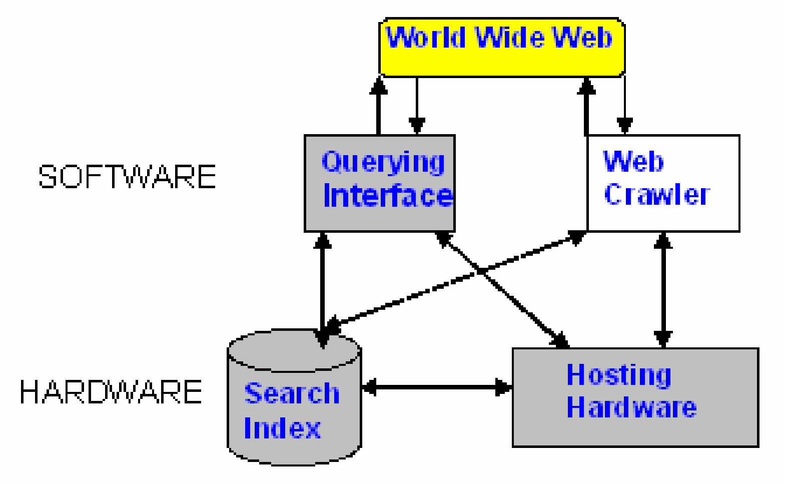


Figure 4.2. The architecture of a search engine.

4.2 Problems with the Data

In the process of business intelligence, the lack of credibility on the Web data is one of the big problems. In the case of this thesis it is also a problem. but to reduce uncertainty, the websources are only the competitors' websites so as to guarantee that the information comes from credible sources.

To illustrate this problem suppose that a company wants to find specific information or news about a certain competitor as published in the mass media. The company finds information in different sources on the Internet, but each of the documents has different information. Evidently, this will puzzle users who need to make decisions, and if they do make them, they will probably wonder about having made the right decision since they are unknowing of sources reliability.

Such crises are big problems on the Web and the major reasons for this kind of problems are: (1) The information does not come from one single source, (2) the web sites have an autonomous nature, (3) there are different alternatives to make the selection of Web sites used as information sources, and (4) there is no common source of data, so the information is confusing.

4.2.1 Poductivity Web Problems

Credibility is not the only major problem associated with Web Data. The productivity achieved while searching for relevant information on the Web is a disaster, because at present technology is not in the position to help users find the right information through the available web tools. The first task is to locate relevant Web sites to answer the queries with the search engines. Corporations are interested in finding the right websites for the process of making the competitors' radiographies or to get some information from there. To do this, many websites and their content must be analyzed. There are also several other factors to consider, as well. Such information is rarely provided by a single Web site and thus, it is hidden in different Web sites. Moreover, not all such web sites are relevant to the user.

4.2.2 Historical Data Web Problems

Web data are volatile by default. This means that web data can change any time. A major Web document reflects the most recently modified data at any time. For instance, most newspapers, magazines and websites do not file news reports that are over four months old.

Observe that once data on the Web change there is no way of retrieving the previous data that had been shown or exhibited. The importance of previous information of Web data is not only essential to analyze Web data over a period of time, but also to address the problem of broken links ("Document not found" Error). This common problem arises when a document pointed to by a link does not longer exist.

4.2.3 From Data to Information

There is another major problem with Web data that is the inability to go from data to information. To resolve the above limitations so that Web data can be transformed into useful information, it is necesary to develop effective tools to peform such operations. Currently, there are two types of tools to gather information from different sources: search engines enable us to retrieve relevant documents from the web; and conventional data warehousing systems can be used to integrate information from different sources.

4.3 Limitations of Search Engines

Information on the Web may be found basically by two technologies: browsers and search engines. These technologies offer limited capabilities for retrieving the information of interest. The Web still lacks standards that could facilitate automated indexing. Documents on the Web are not structured, thus, programs can not extract the routine information that a human might find through inspection: author, date of last modification, length of text, and subject matter (these pieces of information are known as metadata). As a result, search engine have made so far little progress in exploiting the metadata of Web documents. Maybe, in the short future, it will be possible to use a new generation of browsers and search engines that will apply ontologies or semantic webs for better and more intelligent performance on the Web search process.

Search engines fail to determine whether the last modification time of a web site is more recent compared to another site. It may also fail to determine those web sites containing relevant information about a particular subject.

Once the relevant information has been detected, the process to integrate relevant data from different Web sites can be done using the methods that follow:

1.- Retrieving a set of Web sites containing the information and navigating through each of them to retrieve relevant data.

2.- Using the search facilities, if any, provided by the respective Web sites to identify potential relevant data and then compare it to compute the desired information.

3.- Using the search facilities on the Web site to get a list of potentially relevant data and then write a program to compare the data retrieving information about the requeriments.

Queries in search engines are evaluated on index data rather than on the up-to-data. Moreover, the search engines do not store historical data, thus, they are not satisfactory tools for converting data to information. For this research work the first method is applied. Each cement corporation has its own web portal, because most of them are public companies and they need to report their performances to the stockholders and public community.

4.4 Traditional Data Warehouse

Data warehouses can be viewed as an evolution of management information systems [89] and as a technique to convert data in information through data extraction. The data warehouses integrate repositories that store information which could originate from multiple and possibly heterogeneous data sources.

The technology of data warehousing has successfully been deployed in many areas such as: manufacturing (for shipment and customers support), retailing (for user profiling and inventory management), financial services (for claim analysis), transportation (for fleet management), telecommunications (for call analysis) utilities (for power usage analysis), health care (for patient analysis) and logistics (for intelligence competitors, warehouse locations and analysis, etc) [89].

4.4.1 Translation of Web Data

The proposal to use traditional data warehousing techniques to analyze semistructured web data has severe restrictions. In a traditional web housing system, each information source is connected to a wrapper [90, 104, 156] which is reponsible for translating information from the native format of the source into the format and data model used by the warehouse system.

Identifying trends, analyzing customer preferences and making opportune business decisions are some of the profits of turning a company´s raw data into employable information through Business Intelligence. However, such a technique becomes very tedious when translating Web data to a conventional data warehouse. This is because for each of the web sites different wrapper components are needed, since the functionality of the wrapper is dependent on the content and structure of the Website. For each relevant Web site, it is necessary to generate a wrapper component (see Figure 4.3.)

If the content and structure of the site change, the wrapper component has to be modified or adapted to the new web site structure. The process of reconfiguration of the previous wrappers will be an extremely tedious and undesirable activity for retrieving relevant data from the Web. In this thesis for the task of wrapper generation, it is proposed to use the Lixto Visual Wrapper and the Lixto Transformation Server for the integration of the multiple wrappers through pipes [85]. With this technology it is possible to guarantee getting good performance and updating of each of the wrappers for data extraction.

4.5 Warehousing the Web

With the previous discussion, it is evident that there is the need to develop or to use novel techniques for managing web data in such way that they are then able to support individual needs of users inside the corporations as well as the needs for business organization in the decision making process. Under the concept of traditional data warehousing techniques, it is proposed to use a special data warehouse design for Web data (a web warehouse) [30,104] to address the needs of information users to support the decision making process inside the corporation (see Figure 4.4).

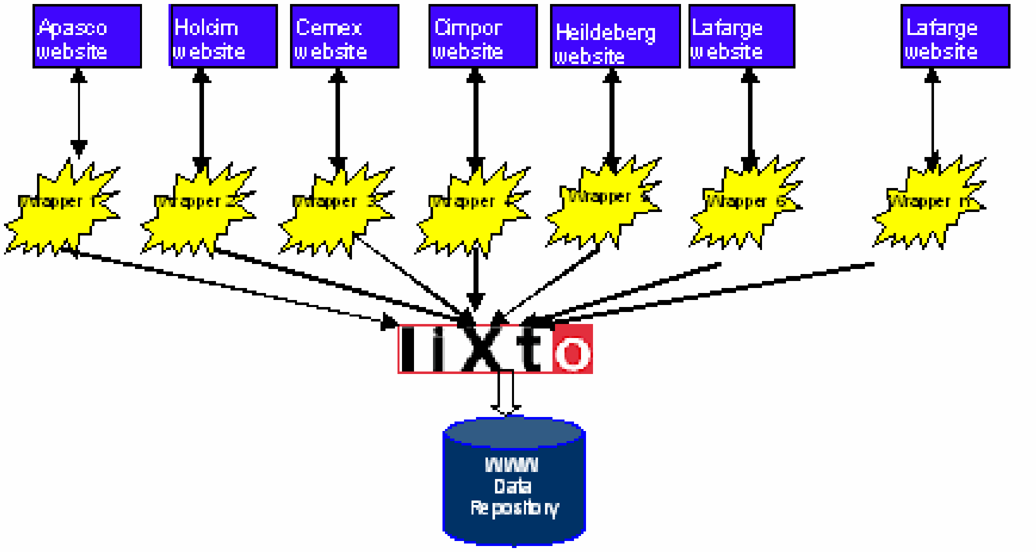


Figure 4.3. WWW data repository for multiple cement industry webportal

It is possible to compare a data warehouse and a web warehouse in that they are both repositories. The most importance difference relies on the fact that the information comes from within the corporation (data warehouse) or from the outside (web warehouse). This integrated data will be available for querying and analysis to generate added value to the decison makers.

The data from different Web sites are extracted and translated into a common data model inside the corporation, and are integrated to existing data at the warehouse. The queries can be executed at the warehouse, and as a result, the data analysis will be performed in a fast and efficient fashion. This process is the same as the one of a conventional data warehouse. To access data at a Web warehouse does not involve costs that may be associated with accessing data from web sources. Another advantage of the web warehouse is that this may also provide access to data when they are not available directly from the Web anymore (document not found or Error!). We can have the historic information that had previously been taken from the competitors' Web sites and Web sources.
Speed of access, availability, and data quality tend to be the major problems for data warehouses in general, and web warehouses in particular. The last problem is

a particularly hard problem. Web data quality is vital to properly manage a web warehouse environment.

With this problem, the quality of the data will limit the accuracy of the decision makers when making informed decisions. Data quality problems usually occur in one of two situations: when data are retrieved and loaded into the web warehouse, or when the web sources themselves contain incomplete or inaccurate data. (It is impossible to guarantee data of quality at the beginning of the process), so it is important to trust the web data contained on the Web sources and also to trust the data extraction process. After that process, it will be possible to purge the data in the data repository or web data (This process is expensive).

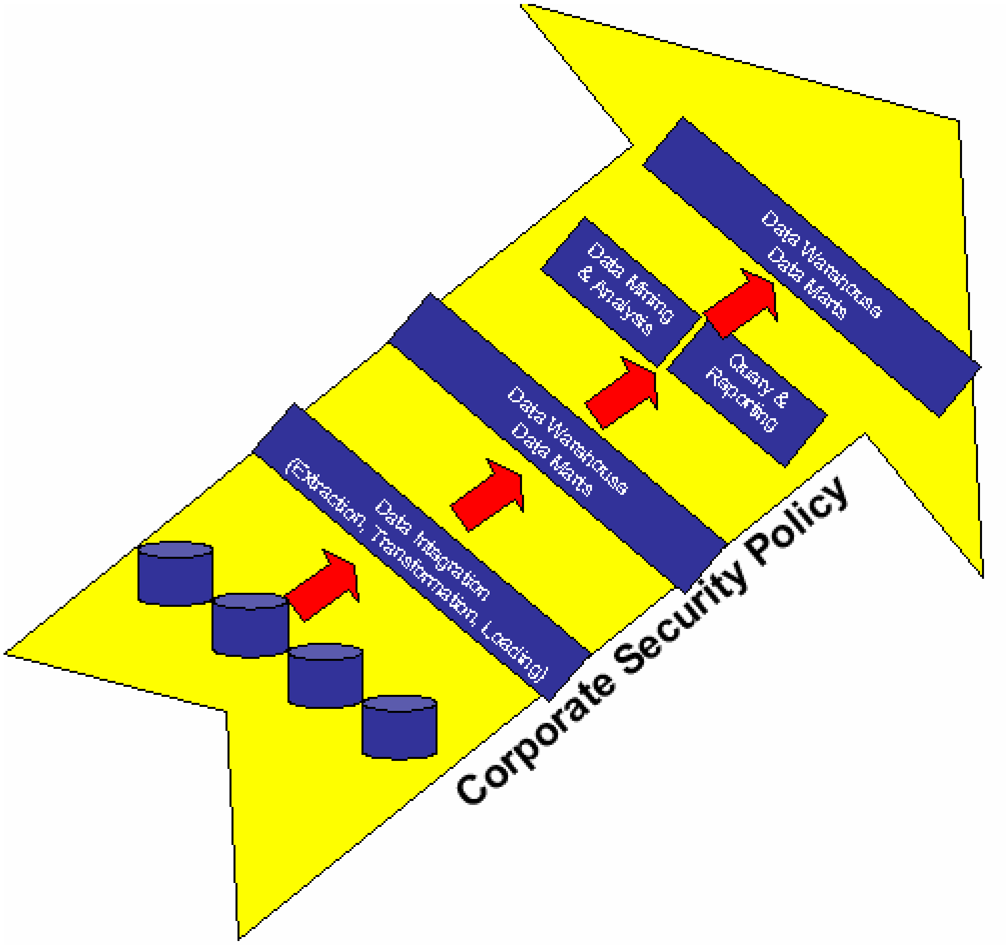


Figure 4.4. Business intelligence and data warehousing integrating data from web sites

The problem of innacurate data on the Web sources is the most difficult to face. In this case, there is nothing –no tools, no techniques- to improve the quality of data.

Hence, improving the data quality of the Web sources cannot be done in the warehouse. However, this thesis proposes a set of suggestions to bear in mind to try to improve the quality of data in the data repository when data are retrived and loaded into the web warehouse.

- Retrieving relevant Web data from a couple of web sources may also bring along irrelevant information. The existence of irrelevant information enlarges the size of the warehouse.

- Through the World Wide Web, there is access to a large amount and varied information from innumerable sources. Moreover, the Web has no restrictions, but the information can disappear or change whenever. These switches take two general forms. The first is related to presence. Web pages and sites show a variety of long-lived patterns. The second is modification of structure and content: web pages replace their antecedents, usually leaving no trace of the previous document, the wrappers for the data extraction on the new web source cannot work with the web structure of the new web source.

- There are duplicated documents. Note that these duplicated documents may have different URLs. This makes it harder to identify such replicated documents autonomously and remove them from the web warehouse. One expensive technique is to compare the content of a pair of documents.

A disadvantage of the warehousing approach is that the group of decision makers needs to identify relevant Web sites from which the warehouse will be fed. The process to find crucial websites starts with a keyboard search using a popular search engine or by querying the web directory services.

Business Intelligence is a systematic evolution of the Data Warehousing (a data repository designed to support an organization´s decision making processes), making it cost-effective to store and to manage the warehouse data, critical to any BIDW solution. If the organization does not have a capable data warehouse, it will not be possible for analysts to get information bound to analysis so as to generate opportune decision making.

It is a fact that the capacity to obtain information in real time has increasingly become more critical. Considering that information systems are more efficient now, the time span of the decision-making cycle has reduced drastically. The pressure of competition requires from businesses to not only make intelligent decisions based on their incoming data but also to make them quickly. As simple as it could sound, the ability to turn raw data into fruitful information opportunely can add hundred, thousands or even millions of dollars to an organization's bottom line.

4.6 Web Query Systems

With the complexity that generates the large amount of information available on the World Wide Web on a daily basis, it is more complex to locate relevant information. The large amount of information that is posted on the Web everyday makes it a complex, unstructured World Wide Web network. Thus, it is evident that there is a need for immediate, effective and efficient tools for information consumers. These tools must help the user easily locate disparate information on the Web, ranging from unstructured documents and pictures to structured, record-oriented data. When doing this, one cannot just ask for the information of interest. If the information were to be found, it should allow to be located wherever. An alternative option is to query the data on the Web. This has led to an increasing research effort in developing efficient query mechanisms for the Web. The most popular form of querying systems on the Web is the set of available Web search engines.

4.6.1 Search Strategies with Search Engines

Some of the search strategies with the search engines will be applied when the user submits keywords to databases of Web pages, and gets back a different display of documents for each search engine. Results from submitting comparable searches can differ widely, but they can also contain some duplicated documents depending on the search techniques of each of the search engines.

Some of the characteristics of popular search engines [50] are shown in Table 4.1. Note that the feature "Other sources" means that the search service may search Internet sources other than Web pages: most commonly the message archives of Usenet newsgroups. Many also link to services providing e-mail address directories and may have a directory of internet resources arranged by subject. Such is the case of the new Gmail technology of Google. The words "OR, AND" indicate search keywords. Automatically all the ORs will look for pages with any of the words and the ANDs will look for pages containing all of the search keywords. The feature "+ and –" means that the term that must be present can be prefixed with '+' (required). Those not required can be indicated with '-' (rejected). The words "AND, OR, NOT" specify that the term can be linked with

AND to show all words that must be present. With OR, the search can look for synonyms and with NOT it will exclude words.

Some services have options such as “any words” or “all these words” which have the same purpose. Next, the feature “fields” indicates that in some search engines It is feasible to search either the URL or the title of the document. “Truncation” allows the system to search for a number of words begining with the same word root, eg., “ceme*” would find cement products and cemetery. The parameter of “adjacency” means that the words need to be close to each other or followed by another word from a set of words. Finally, the feature “proximity” ensures that the search words are near each other, say, in the same sentence.

Nowadays, search engines are offering limited capabilities for retrieving information of interest. Search engines as Yahoo and Google also provide a lot of alternative tools and applicatios to cover their deficiencies. We can foresee in the short future more robust search engines. The following are some of the search engine shortcomings identified by the Internet community:

- **Shortage of support for metadata queries**: Because of insufficient Web support, at present it is not totally possible to get automated indexing metadata such as author, last date of modification, anchor keywords associated with hyperlinks and URLs, length of text, or summary of documents.

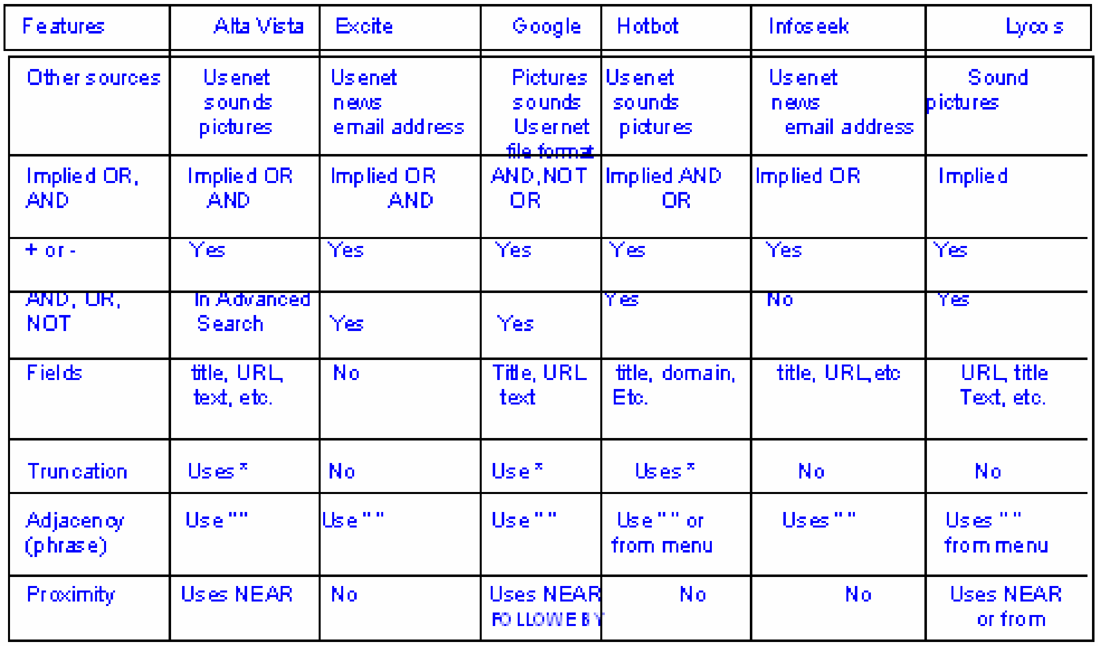
- **Shortage of support for querying interlinked documents:** So far, search engines have made little progress in exploiting the topological nature of documents on the Web. Most of the search engines always return a set of disconnected documents as their search results. They do not allow users to input constraints on the hyperlink structure of a set of documents.

- **Shortage of support for structural queries:** Search engines do not exploit the structure of Web documents. One cannot express conditions based on the hierarchical structure of the tag elements in a query. For instance, consider the query to retrieve documents that include the list of feautures of a digital camera in the form of a table on a Web page. A search engine may return thousands of documents containing a wide variety of information related to digital cameras, most of which are not necessarily structured in the form of a table. There are a lot of oppotunities for the seacrh engines to explore the structure of HTML, XHTML and the XML technologies.

- **Search engines recognize text only:** The search is limited to string matching and text associated with images. Moreover, most search engines recognize text only. Numeric comparisons, as in conventional databases can not be done. For example, the following query can not be expressed: Find the Apasco cement product sold for less than $15 USD

- **Existence of duplicated documents:** Due to the large numbers of replicated documents on the Web, the query results of the search engines will display the most unavoidable, identical existing Web documents. These identical documents may be scattered gradually in the query result returned by the search engines. The only way to find these duplicated documents is by manually browsing each Web document and comparing them to each other.

- **Limitation of querying dynamic Web pages:** Nowadays, search engines retrieve content only for a portion of the Web, called the publicly indexable Web [91]. This refers to the set of web pages reachable in the public pages on the Web, ignoring search forms and pages requiring authorization or registration. A great portion of the Web is in search-hidden forms. (Lots of databases are available only through HTML, XHTML and XML forms). This portion of the Web was called the *hidden Web* [28] and the *deep Web* [65]. To get access to the data that is taken from the deep Web, it is necessary to submit some queries, and as a response, the data are showed.These dynamically generated pages to queries submitted via the search forms cannot be analyzed or indexed by Web crawlers. Thus, search results ignore such dynamic Web documents.



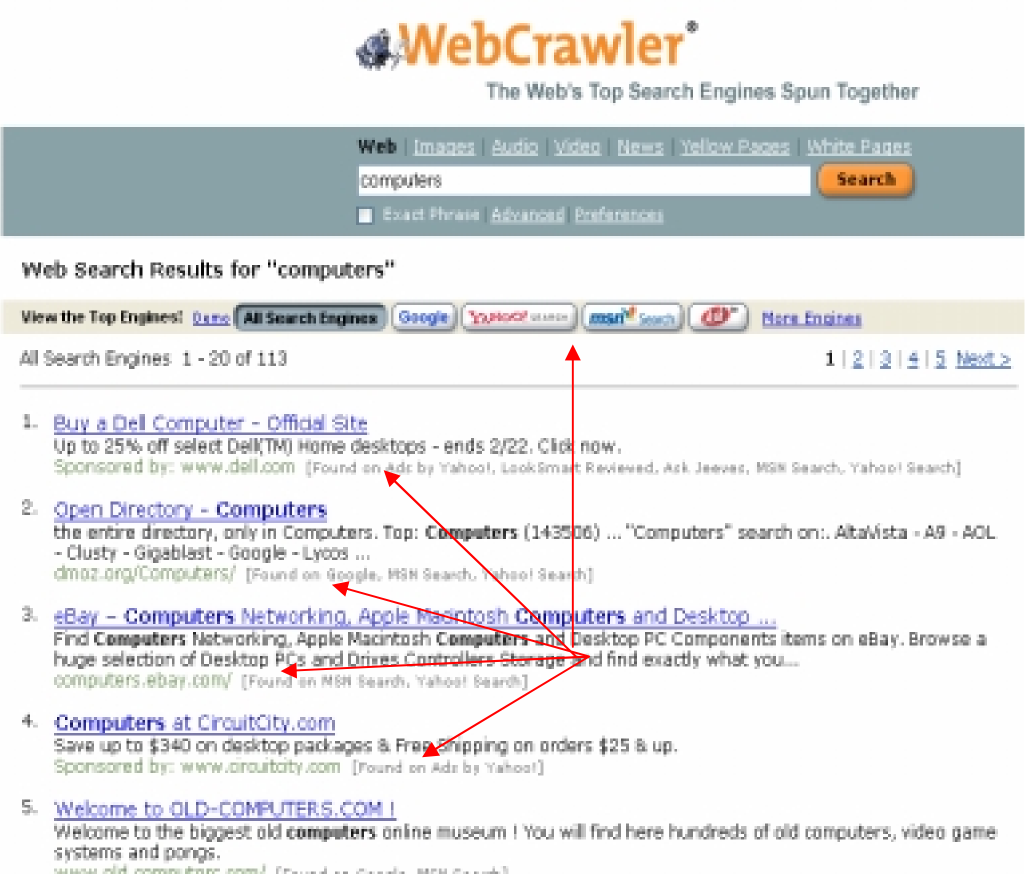
Features	Alta Vista	Excite	Google	Hotbot	Infoseek	Lycos
Other sources	Usenet sounds pictures	Usenet news email address	Pictures sounds Usernet file format	Usenet sounds pictures	Usenet news email address	Sound pictures
Implied OR, AND	Implied OR AND	Implied OR AND	AND,NOT OR	Implied AND OR	Implied OR	Implied
+ or -	Yes	Yes	Yes	Yes	Yes	Yes
AND, OR, NOT	In Advanced Search	Yes	Yes	Yes	No	Yes
Fields	title, URL, text, etc.	No	Title, URL text	title, domain, Etc.	title, URL, etc	URL, title Text, etc.
Truncation	Uses *	No	Use *	Uses *	No	No
Adjacency (phrase)	Use ""	Use " "	Use " "	Use " " or from menu	Uses " "	Uses " " from menu
Proximity	Uses NEAR	No	Uses NEAR FOLLOWEBY	No	No	Uses NEAR or from

Table 4.1 Some of the features of popular search engines

- **Shortage of support for querying XML data:** Currently, none of the search engines support querying for XML documents. They do not provide the facility to specify query conditions in the form or path expressions, either. This conditions play an important role for querying XML data.

4.6.2 Metasearch Engine

A new kind of technology of metasearch engines is supporting the search activities on the Web. Some of these metasearch engines are: AlltheWeb (www.alltheweb.com), Inference Find (www.infind.com) and MetaCrawler (www.metacrawler.com). With these tools a query is submitted in the form of keywords in its search box, and it is trasmitted at the same time to several individual search engines and their databases. The search then returns results from various search engines combined all together. For instance, WebCrawler submits a query simultaneously to the following search services: about.com, Alta Vista, Excite, Infoseek, Looksmart, Lycos, Thunderstone and Yahoo.


WebCrawler®
The Web's Top Search Engines Spun Together
Web | Images | Audio | Video | News | Yellow Pages | White Pages
Search
Exact Phrase | Advanced | Preferences
View the Top Engines! Demo
All Search Engines
Google
YAHOO! SEARCH
msn Search
More Engines
1 | 2 | 3 | 4 | 5 Next >

computers

☐ Exact Phrase

Web Search Results for "computers"

All Search Engines 1 - 20 of 113

1. Buy a Dell Computer - Official Site
Up to 25% off select Dell(TM) Home desktops - ends 2/22. Click now.
Sponsored by: www.dell.com [Found on Ads by Yahoo!, LookSmart Reviewed, Ask Jeeves, MSN Search, Yahoo! Search]

2. Open Directory - Computers
the entire directory, only in Computers. Top: Computers (143506) ... "Computers" search on:. AltaVista - A9 - AOL - Clusty - Gigablast - Google - Lycos ...
dmoz.org/Computers/ [Found on Google, MSN Search, Yahoo! Search]

3. eBay - Computers Networking, Apple Macintosh Computers and Desktop ...
Find Computers Networking, Apple Macintosh Computers and Desktop PC Components items on eBay. Browse a huge selection of Desktop PCs and Drives Controllers Storage and find exactly what you...
computers.ebay.com/ [Found on MSN Search, Yahoo! Search]

4. Computers at CircuitCity.com
Save up to $340 on desktop packages & Free Shipping on orders $25 & up.
Sponsored by: www.circuitcity.com [Found on Ads by Yahoo!]

5. Welcome to OLD-COMPUTERS.COM !
Welcome to the biggest old computers online museum ! You will find here hundreds of old computers, video game systems and pongs.
www.old-computers.com/ [Found on Google, MSN Search]

Figure 4.5. Webcrawler a metasearch engine

For this research work it is proposed to work with with Web Crawler (www.webcrawler.com). It consolidates the results in one large list, ranked by a score derived from the rankings of each of the search engines, and it also lists each site (see Figure 4.5).

Metasearch engines are useful when a user is looking for a single term or phrase contained in the database of several search engines. However, the following are some of the shorthcomings of the metasearch engines [91].

- The Metasearch engines explore multiple databases of different search engines and get only part of the information that is provided by each of these search engines after being queried.

- Since metasearch engines retrieve results from different search engines, the limitations of these search engines -as discussed above- also appear in the results of metasearch engines.

- If the query contains more than one or two keywords or it has a very complex logic, then most of the keywords will be lost while searching the relevant databases. This kind of operations will be possible if the search engine can sustain the logic of the query. (See Table 4.1)

The above limitation of search and meta-search engines has inspired considerable research activity in the academic community to build efficient web query systems for querying HTML and XML data [28].

4.7 Web Site Mining

The problem of finding new web pages of special interest to a user has not been solved adequately yet. Since companies need to know about their potential competitors, suppliers and customers, the web-crawlers and spiders have to solve the problem of finding those pages of interest to a specific project. The single main pages are not the main focus. Some techniques have been created to utilize the hyperlinks for a better categorization of pages as follows [17,108,15]:

1- Super pages. A web site is represented as a single virtual web page consisting of the union of all pages.

2- Topic vectors. A web site is exemplified as a vector of frequencies.

3- Classification of web site trees. A web site is represented by a tree of pages.

The performance of the web page classification crucially depends on the number of web pages that are to be analyzed. A web site as a so called web site tree and the classifier in this work will be based on the paths within these trees (see Figure 4.6).

Mining a web site is different from mining simple web pages. The web sites may vary in size, structure and building technique. To represent the structure within a website, the use of a tree pattern –based on the graph theory- is common. The website of a corporation is represented as a tree which will have specialized trees and subtrees. (See figure 4.7).

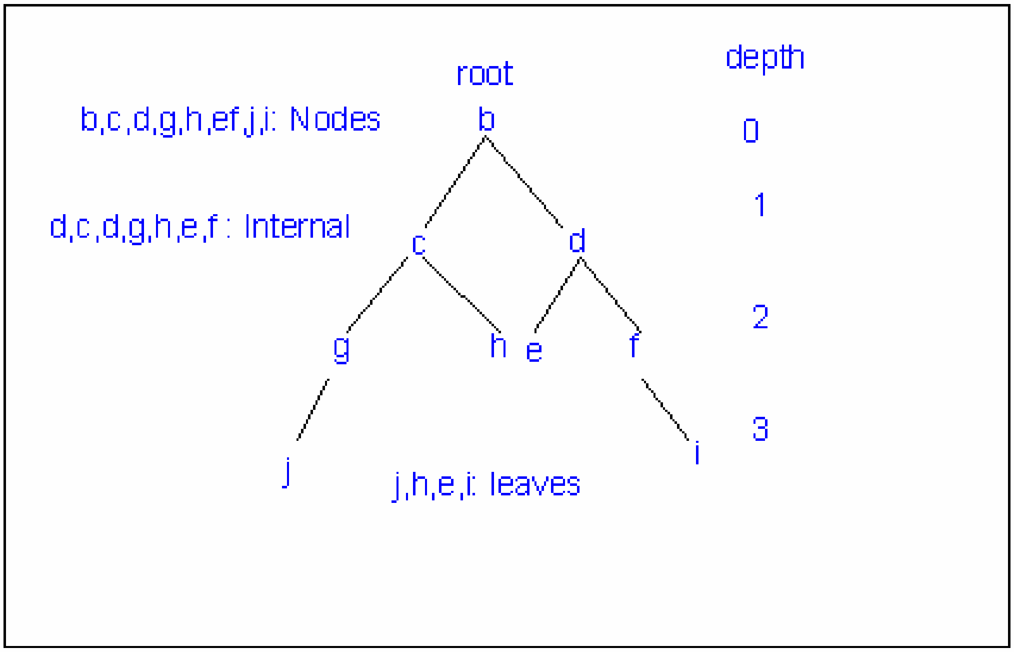


Figure 4.6. Typical binary tree

This tree structure is of much help for data extraction since it locates relevant data on the Web site considering that it is known what data is required and where this information is located inside the Web Source. Most of the competitors' Web sources are big portals that the analyst will represent as web site

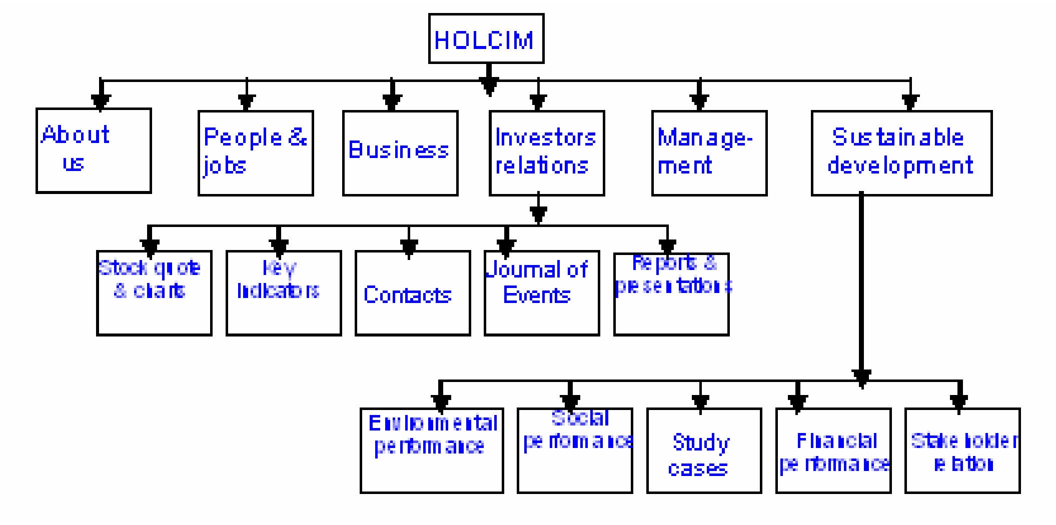


Figure 4.7. The tree structure of the web site of Holcim Corporation (Competitor)

4.8 Web Data Extraction

After the structured sets of the website trees under study are known, the next problem to be faced is to extract data from the website. This has opened the opportunity for users to benefit from available data in many interesting ways[17]. Usually, the end users retrieve information from browsing and keyword searching from the search engines as mentioned previously. These are intuitive forms of obtaining data on the Web.

These search strategies do pose some limitations. Sometimes, the browsers are not the best tools to try to locate specific information. The main problem is to try to follow the multiple links, fact that could generate the loss of some information or data. The keyword searching is often more efficient than browsing. This technique returns a vast amount of data that users can manipulate.

For the manipulation of Web data in a more effective way, the academic community has resorted to ideas taken from the database area [92]. As we know databases, require structured data, and therefore, traditional database technique cannot be directly applied to Web data.

The arrival of XML [16] as a standard for structuring data on the Web has brought some ideas to solve this problem. However, this technology doesn´t provide an ordinary solution for the manipulation of the existing web data. Nevertheless, the volume of unstructured or semistructured data available on the Web is vast and ever-growing. A possible strategy to solve this problem is to extract the relevant information from webpages of interest and to store it in a database(s) for later manipulation.

The most traditional technique for data extraction from Websources is to develop specialized programs, called wrappers. This specialized programs detect the data of interest and map them into a suitable format as, for instance, XM.

The most interesting aspect of wrappers is that they must be able to recognize the data of interest within a large amount of text. These data will have a flat structure, will be complex and will present an implicit hierarchical structure. The data could even have structural variations that must be tolerated.

There are some disadvantages when developing wrappers manually. The two most important are the difficulty in writing and the maintenance they must receive since the information and the structure of the web pages changes almost 24/7. Nowadays, there are many tools that have been proposed to better address the issue of generating wrappers for Web data extraction [11,4,6,19,25,26,32,37,44,49,56,61,70,84,93]. All these instruments were developed under different techniques such as: declarative languages, natural languages, HTML data modeling, machine learning, ontology, semantic web, etc among others (See Figure 4.8.).

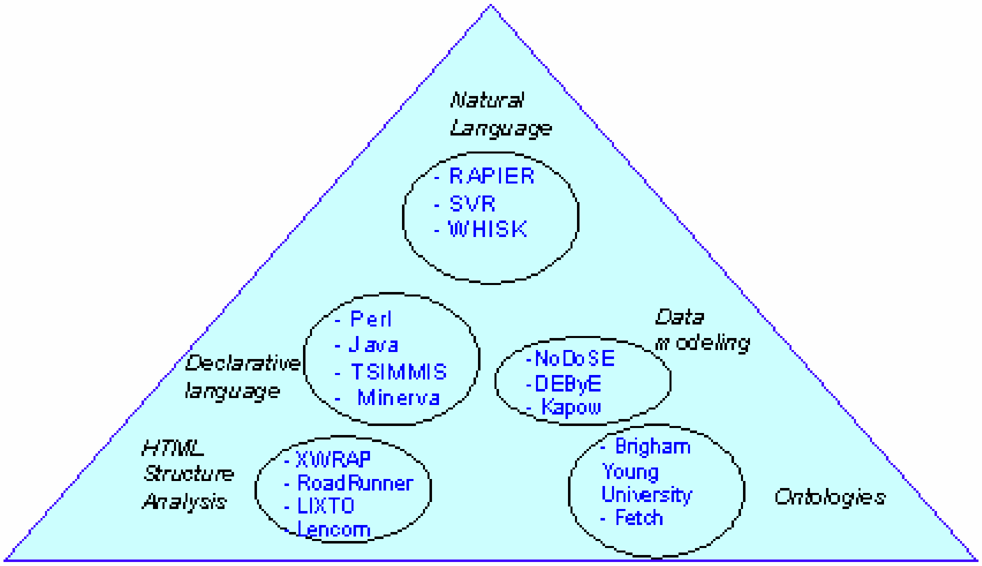


Figure 4.8. Wrapper tools and its techniques

After localizing the data to extract from the Web pages, these will be analyzed as follows. The web page will be represented as a *P*. This page can include one or more objects. The mapping *M* provides the information to the repository *R*, including the object in *P*. At the same time the mapping M can recognize and extract data from other page P´ , similar to page P. In this context, a wrapper is the program that executes the mapping *M*. The common goal when generating wrappers is to create those that are accurate and logical in extracting data. (see figure 4.9).

A very important feature of any data extraction tool is its degree of automation. This is related to the amount of work left to the user during the process of generating a wrapper for Web data extraction.

Most of the data available on the Web implicity present a complex structure. Typically, this structure is loose and has degrees of variation common to semistructured data [2]. Thus, wrapper generation tools are expected to deal properly with such complex objects.

The main objective associated with the data extraction tools is to make easy the generation of the wrapper. The tools or software devised for the wrapper generation is computer-coded writing which uses general purpose languages such Perl, Phyton and Java. Another important characteristic of these tools is the interaction with XML and the technologies associated with it. To help the user develop wrappers for Web data, some tools provide a friendly graphical user interface (GUI) with the objetive to make this task easier, as is the Lixto case.

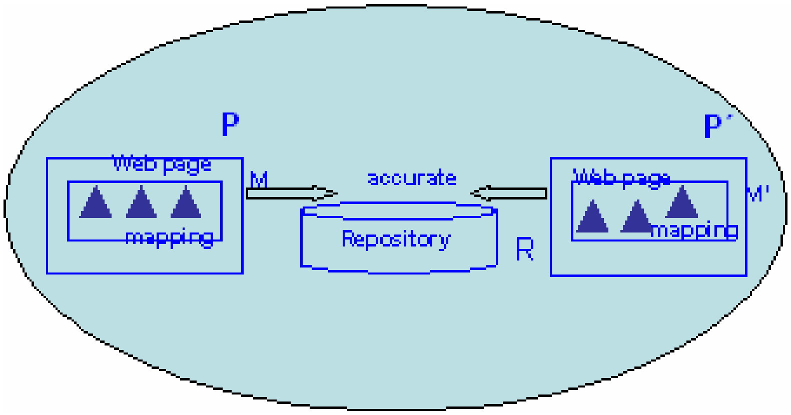


Figure 4.9 Robust wrapper generation

4.9 Semistructured Data

The concept of semistructured data (also called unstructured data) is relatively new research in the computer science environment. The aforementioned stands for the intersection of new forms of representation and querying data that do not fully match the conventional data models primarily available on the Internet. (See Figure 4.10).

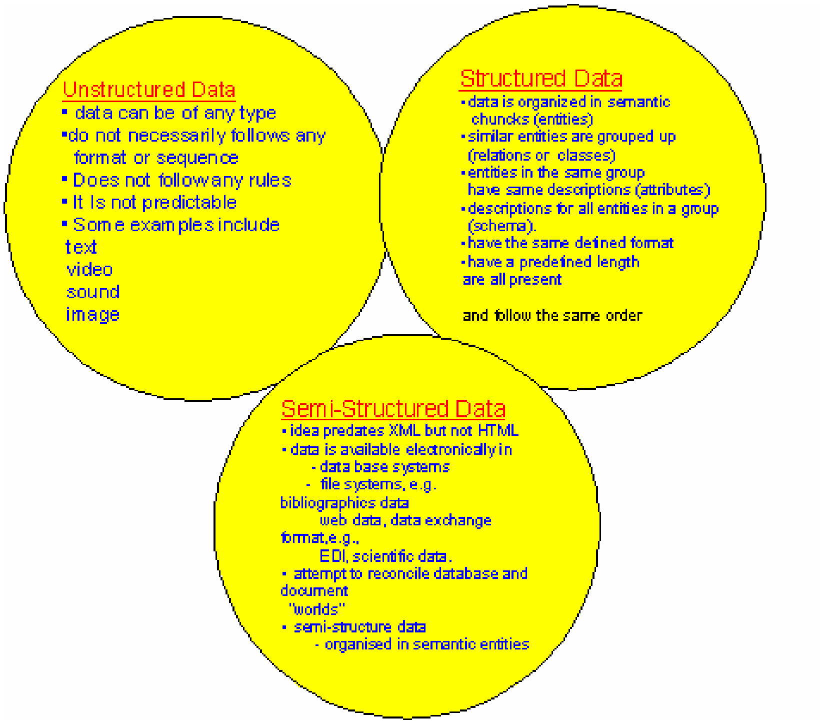

Unstructured Data
▪ data can be of any type
▪do not necessarily follows any
format or sequence
▪ Does not followany rules
▪ It Is not predictable
▪ Some examples include
text
video
sound
image
Structured Data
▪data is organized in semantic
chuncks (entities)
▪similar entities are grouped up
(relations or classes)
▪entities in the same group
have same descriptions (attributes)
▪descriptions for all entities in a group
(schema).
▪have the same defined format
▪have a predefined length
are all present
and follow the same order
Semi-Structured Data
▪ idea predates XML but not HTML
▪ data is available electronically in
- data base systems
- file systems, e.g.
bibliographics data
web data , data exchange
format,e.g.,
EDI, scientific data .
▪ attempt to reconcile database and
document
"worlds"
▪ semi-structure data
- organised in semantic entities

Figure 4.10. Types of data on the Web

The typical example of data that can not be constrained in a schema is the World Wide Web. Object-oriented terms do not thoroughly characterize the relationship between data and schema. A user cannot write a database query without deep knowledge of the schema.

Under the concept of semi-structured data, it is possible to represent data in a tree-like structure. These graphs will generally be referred to as trees. Some characteristics of semi-structured data are as follows: a) the structure is atypical, b) some parts lack data structure, c) some may yield little structure, d) the schema could be prominent, disregarded, or developed quickly.

It is also important to highlight the advantages and disvantages that the semi-structured data model offers:

- Advantages
  - discover new data and load them
  - incorporate miscellaneous data
  - query without knowing data types

- Disadvantages
  - loses the kind of information associated to the data chunk
  - creates optimization arduous

4.10 XML Technologies

After the data extraction, it is necessary to use another important technology for the data standardization. XML [16] is becoming the most important standard for data representation and exchange on the Web. For this reason, it is considered as an important feature if the data extraction tool provides outputs in XML.

A vast amount of semistructured data stored in electronic form is not present in HTML pages, but in text files, such as e-mail messages, program codes and documentation, configuration files, system logs, etc. Therefore, it is very important that the data extraction tools might be able to handle such data sources in XML.

XML was developed by a consortium of organizations and companies formed under the protection of the World Wide Web Consortium (W3C) in 1996. At the beginning, XML was developed with the idea to simplify the Standard Generalized Markup Language (SGML) and to make it applicable for general purposes. XML is not a markup language, it´s a toolkit for creating and using markup language. The recommendations for the XML are - among other things- the design goal for XML [105].

The design goals for XML are: will be usable over the Internet, will support a variety of applications, will be compatible with SGML, It will be relatively easy to write programs which process XML documents, XML documents should be comprehensible and clear for humans and computers, XML designs will be concise, XML design should be prepared rapidly and XML documents will be easy to create and share.

4.10.1 Validation of an XML Document

The main elemnt of XML information is the *XML document*. An XML document is integrated of markups and other elements. The XML´s markup divides a document into separate information containers called elements. These elements nest inside each other and conform the content of the document. Each document has a *root element* at the top level which contains other elements in the subsequent levels. An XML document will be represented as a hierarchical tree.

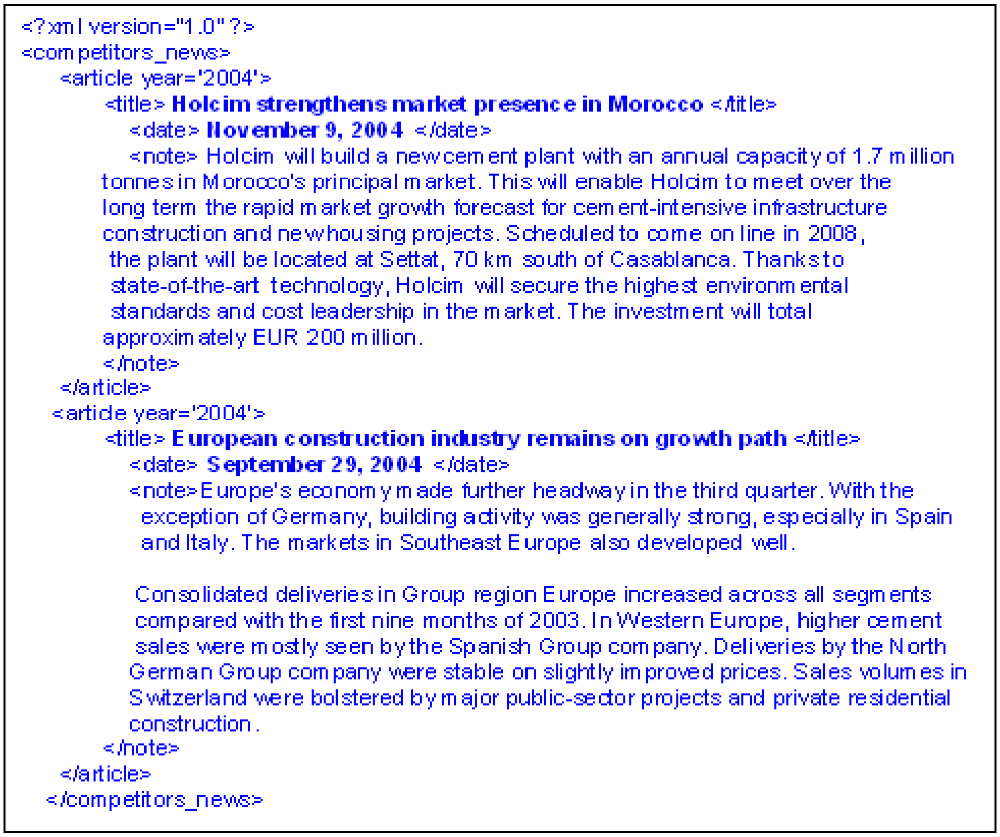
At the top of the document is the XML declaration, <? xml version=”1.0”?>. This helps an XML processing program identify the XML version, and the kind of character encoding it has. Likewise, it helps the XML processor to get started on the document. It is optional , but a good thing to include in a document.

If XML markup is a structural skeleton for a document, the tags are the bones. They mark the boundaries of elements allow insertion of comments and special indtructions, and declare settingsd from parsing environment. An XML document is a sepcial construct designated to archive data in a way that is most convenient for parsers. The XML document has two parts. First is the document prolog. The second is an elemnt called the document element, also called the root element.

Elements are building blocks of XML, dividing a document into a hierarchy of regions, each serving a specific purpose. Some elements are containers, holding text or elements. Other are empty.

The syntax for a container element will begin with a start tag consisting of an angle bracker followed by a name. The start tag may contain some attributes separated by white spaces, and it ends with a closing angle bracket.

Some elements could include additional information in the form of attributes. An example of an XML document and its name-value pairs is when someone gets the information of a set of articles on a website and stores this information. This XML format looks as shown in Figure 4.11.



```
<?xml version="1.0" ?>
<competitors_news>
    <article year='2004'>
        <title> Holcim strengthens market presence in Morocco </title>
           <date> November 9, 2004 </date>
           <note> Holcim will build a new cement plant with an annual capacity of 1.7 million
        tonnes in Morocco's principal market. This will enable Holcim to meet over the
        long term the rapid market growth forecast for cement-intensive infrastructure
        construction and new housing projects. Scheduled to come on line in 2008,
         the plant will be located at Settat, 70 km south of Casablanca. Thanks to
         state-of-the-art technology, Holcim will secure the highest environmental
         standards and cost leadership in the market. The investment will total
        approximately EUR 200 million.
        </note>
    </article>
   <article year='2004'>
        <title> European construction industry remains on growth path </title>
           <date> September 29, 2004 </date>
           <note>Europe's economy made further headway in the third quarter. With the
             exception of Germany, building activity was generally strong, especially in Spain
             and Italy. The markets in Southeast Europe also developed well.

            Consolidated deliveries in Group region Europe increased across all segments
            compared with the first nine months of 2003. In Western Europe, higher cement
            sales were mostly seen by the Spanish Group company. Deliveries by the North
           German Group company were stable on slightly improved prices. Sales volumes in
           Switzerland were bolstered by major public-sector projects and private residential
           construction.
        </note>
    </article>
  </competitors_news>
```

Figure 4.11. Example of competitor news XML document

It´s also worth noting that a document is not a necessarily the same as a file. A file is a package of data with a physical structure. An XML document can exist in one or more files, some of which may be in other systems. The XML uses a special markup to integrate the content of different files to create a single entry, described as a logical structure.

Strictly speaking, the XML is not a markup language. A language has fixed vocabulary and grammar, but the XML does not actually define any elements on which a language of its own can be built. So, the XML can be called a markup language toolkit.

If we need a markup language is possible to build one quickly using XML, and it will automatically be compatible with XML tools. The most important aspect is that XML gives the ability to tailor a markup language. With these rules there are two ways to create a markup language with XML:

1.- Well-formedness

> Because the XML does not have a predetermined vocabulary, it is possible to invent a markup language. As a consequence, there are some of these rules to form and implement tags. An example of such rules can be that each start tag must have an end tag, and so forth. However, there is no restriction on the ordering or the labeling of the tags. Free-form XML, is perfectly legal as it´s well formed. In other words, as long as you can spell tags correctly, use both start tags and end tags, and obey all the other minimal rules, it´s a good XML.

2.- DTD´s and Schemas

> Under the definition and rules of XML, there are two document modelling standards used with XML, Document Type Definition (DTD) [86] and XML Schema [87]. Both of them are used to create specifications to layout a document.

DTDs are built in accordance with the XML 1.0 specifications. They are usually independent documents to which other documents can refer although parts of the DTDs can also reside within the document. A DTD is a collection of rules, or declarations, describing elements and other markup objects. An element declaration will adds a new element type to the vocabulary and defines its content model, what elements can contain and in wich order. Type of elements not declared in the DTD is illegal. Any element containing something not declared in the DTD is also illegal.

There are to ways to define a DTD: The First one, when it can be included in the XML document it is called an *internal* DTD. The second one, the DTD is defined in a external file, it is called an *external* DTD.

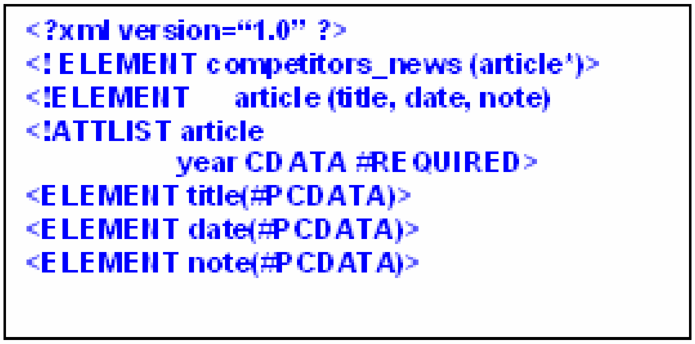
Schemas are a more recent invention. They offer more flexibility and a way to specify patterns for data, wich is absent from DTD. With an XML schema is possible to develop a more robust document definition than with a DTD. In a general sense of the world, a schema is a generic representation of a class of things.

Some of the reasons why the W3C recommends the use of XML Schemas as follows:

- It is easier to learn that DTDs.
- It is extendable to future additions.
- It is richer than DTDs.
- It supports namespaces.
- It is written in XML.

XML schemas improve DTDs in several ways, including the use of XML syntax and support for datatypes and namespaces, both of which are invaluable for interfacing XML and databases.

A posible DTD for the previous example is shown in Figure 4.12.



```
<?xml version="1.0" ?>
<! ELEMENT competitors_news (article*)>
<!ELEMENT     article (title, date, note)
<!ATTLIST article
              year CDATA #REQUIRED>
<ELEMENT title(#PCDATA)>
<ELEMENT date(#PCDATA)>
<ELEMENT note(#PCDATA)>
```

Figure 4.12. DTD for the XML document of competitors´ articles

One of the advantage of using XML Schema is that it follows XML syntax rules and therefore can be parsed by an XML parser. The sintax rules are: All nonempty elements must have an opening and closing tag, all empty elements must be terminated: <empty/>, all atributes must have values , and thos values must be in quotation marks, all elements must be nested correctly, elements and attributes are cas-sensitive.

4.10.2 Other XML Specifications

There are technologies that complement the XML, such as DOM, SAX, XSLT, Xpath, Xlink, Xqueries. They are other applications of XML that build relations to other formats the work and transformations with data. All the names and acronyms may be somewhat overwhelming, but it is worthwhile getting to know this growing family.

Transformation is one of the most important and useful techniques to work with XML. In an XML transformation, the structure, the markups, and perhaps, the content of the document will change. A transformed document may be altered or profoundly changed. The process is shaped with a configuration document sometimes called either a stylesheet or a transformation script. Documents can be transformed by using XSLT.

There are many reasons to transform an XML. (Most often, the reach of an XML document is stretched to new areas by converting it into a reader-friendly user format, comprehensible for both, humans and computers. Alternatively a transformation could alter the content of a document, such as extracting a section, or adding a table. The ability to transform XML documents increases flexibility:

1. It allows to store data in one format and display it a different one.

2. It consents the saving of data in a specific format and show it later on in another format. Using XSLT gives the possibility of extracting only the bits of information needed, transforming them into other formats and finally, storing them in a different document, file or database.

4.10.3 Some Restrictions of XML

XML is not the perfect solution for every task. XML is limited in terms of the data types it supports: XML has no means to support complex data types such as multimedia data. The reason is that XML is a text-based format.

Also, XML is limited in terms of security: XML has no facilities to ensure a secure exchange of XML documents. XML documents are only simple text documents without any encrypting facilities [87].

Chapter 5

Wrapper Generation for Information Web Sources

The World Wide Web is the international standard for information exchange. The main principle for the wrapper generation is to split information into chunks that can be transmitted and processed with other entities. Nowadays, the unit of information on the Web is typically a document that is created by a user and shared with others in the form of URL (Uniform Resource Locator). Other users and systems can keep the URL to retrieve the file or document when required. The sucess of the World Wide Web came after the development and incursion into the HTML, which at the beginning was introduced in order to structure the text for visual presentation. It also describes both an intradocument structure (the layout and format of the text), and an interdocument structure (references to other documents through hyperlinks).

5.1 Wrapper Definition

As previously mentioned, wrappers are complex and specialized computer programs that will mechanically extract data from the internet websites and then convert the information into a structured format. For this reason, it can be said that wrappers have three main functions [41].They must be able to:

- Download the HTML code from a specific webpage or website.

- Search for, recognize and extract specified data.

- Save these data in a suitably structured format to enable further manipulation.

As it will be seen, the data can be imported onto another application for additional processing. Sahughet [93] identified that around 80 percent of the published information on the World Wide Web is acquired from databases that are running in the background. After the collection of this data is placed onto an HTML document, the structure of the hidden databases is completely lost. The wrappers intent to invert this process by changing the information to a structure format [94]. Under this concept and by using suitable programs, it seems possible to use the Web as a big database with a universe of hidden information.

With the execution of some wrappers from the various information sources of the World Wide Web, it is possible to retrieve data that can be made available in an appropriately structured format [31]. As a rule, a special development wrapper is required for each individual data source because each of the web sites has a different HTML structure.

Another important problem in the data extraction process is that the web is also extremely dynamic, which results in frequent changes in the structures of websites. Under this scene, it will be constantly necessary to update or rewrite the wrappers in order to preserve the data extraction capabilities for a requiered webpage or website[3].

It should be remembered that the main interest is to get the information a competitor publishes every day as well as the information about its addresses of warehouses and distribution centers published on the Web. The source for this information sometimes is a relational database, and Web pages are generated on demand by invoking an SQL query and formatting its output into HTML. Also, for another kind of analysis, it is needed to obtain some financial information which is only accessed through HTML pages. Here the solution is to write software to parse the HTML and convert it into a structure suitable for the analysis software.

In contrast to conventional database management systems, communication with data on the Web presents an essential shift of paradigm. To start with this kind of paradigms, it is analyzed that the standard database is based on a client/server architecture (See figure 5.1) The client (computer program or a computer) executes a query that is processed and compiled. After the program execution has been completed, an answer data is returned by the server as a response from the query.

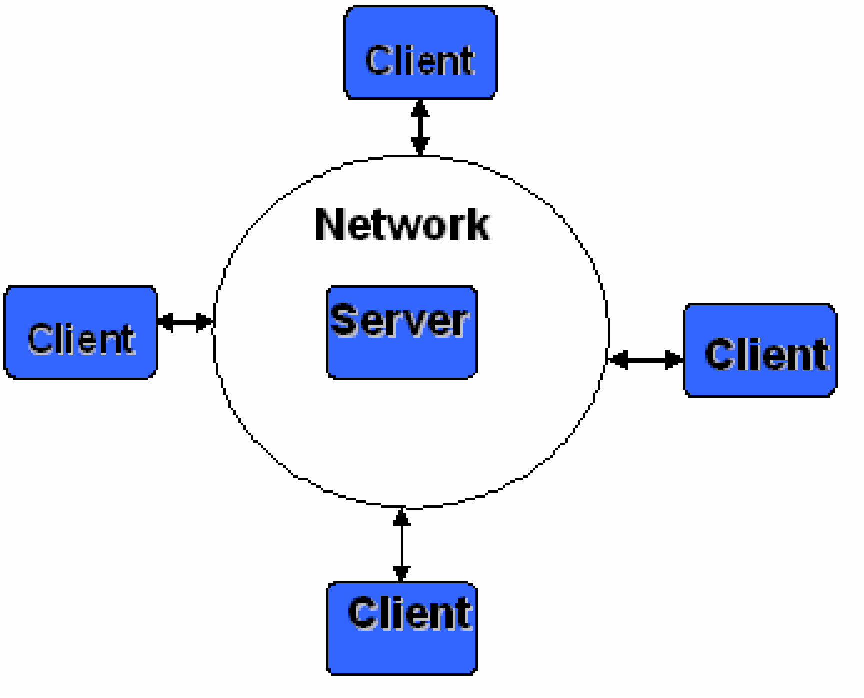


Figure 5.1 Traditional client/server database architecture

In contrast, data processing in a Web context is based on a “multitier” architecture (see figure 5.2). In this kind of architecture, the lower tier consists of ***data sources,*** also called servers. These servers may be conventional database servers; they may also be file servers or other kind of applications that produces data. Regardless of the kind of data a source stores, they will be translated into a common logical data model and a standard format, such as an XML.

The ***client tier*** is composed of a set of applications -such as an analysis package/ that consumes data. Between the client and the server, there is a collection of intermediate tiers. This tier set is often called *middleware.* A middleware is a software that transforms, integrates and adds value to the data.

Considering the simplest scenario in the architecture of a client/server, it is not possible to include a middle layer and thus, the interaction between client and server is direct. After the information has been processed, the data flow to bring the results will be from servers to clients. When new data are required, the queries are shipped on the opposite direction. The query processing in the server consists of ***translating*** the query into the server’s own data model, then processing it through a query engine and finally, translating the results into the common logical data model.

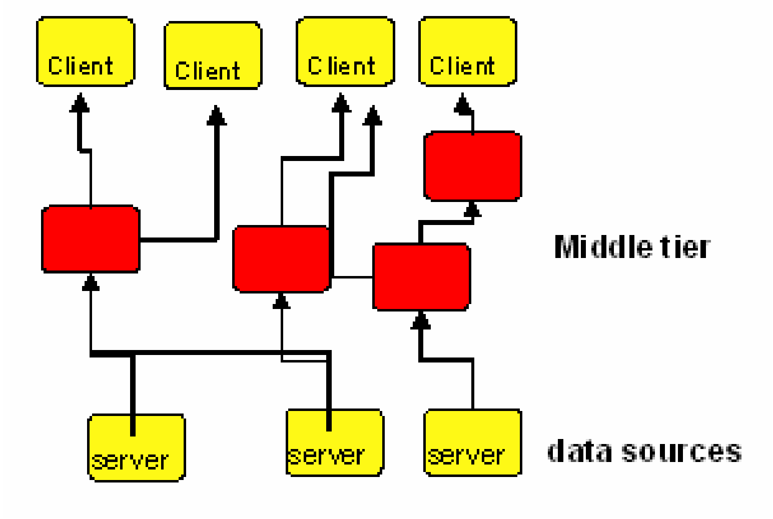


Figure 5.2 Web-based architecture

The database community that is interested in data integration is working on the topic of middleware. The middleware obtained the data from the source and stores it in a intermediate database (the warehouse), this database will be queried by the client. One of the main problems with this approach is to keep the database current when the source are frequently updated.

5.2 Wrapper Development

The process to develop a manual wrapper implies that the wrapper is written by hand using some sample pages. This process is time-consuming, bound to generate a lot of errors, and requires a lot of work. The smallest change in the page structure demands the re-writing of the extraction rules.

The approach of an automatic wrapper development uses some machine learning techniques. These can be based on heuristics of domain knowledge. The heuristic approach tries to make use of the domain knowledge to build powerful wrappers.
In order to obtain information from a lot of sources available on the World Wide Web, it is necessary the use of some tools to build information mediators which extract and integrate data, as previously mentioned.

Mediators are multifunctional. Some of the functions mediators have are as follows:

a) They obtain and (integrate) information from distributed database systems.
b) They take over the problems caused by different locations, query languages and database protocols.

c) They integrate information from multiple Web sources.

The mediator approach implies to pick out Web sources of interest, such as the competitors' Web sources, and to provide integrated access to those pages through a mediator [39,106].

In the architecture of a mediator, the wrapper is a very important component for each individual data source (See Figure 5.3) Each wrapper accepts queries from the mediator, translates the query into the appropriate query from the individual source, performs some additional processing and then provides the results to the mediator.

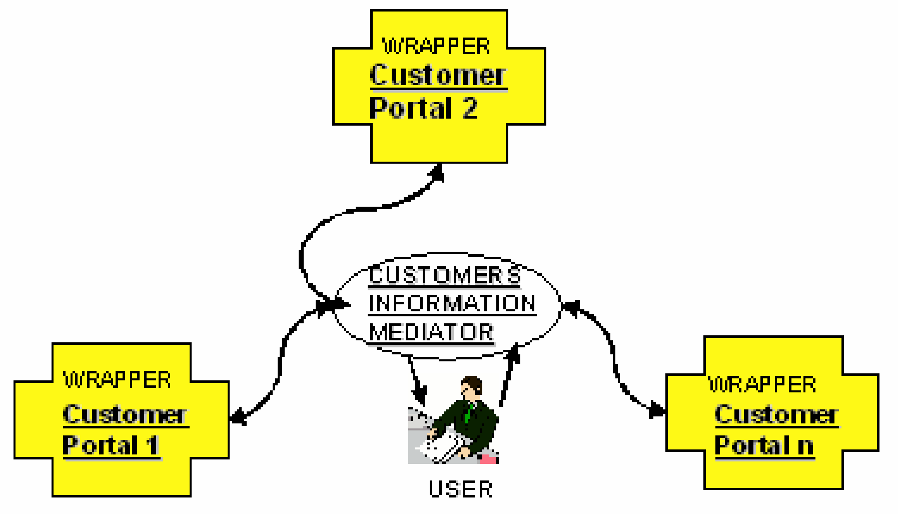


Figure 5.3 Wrappers integrating access to multiple information from different web sources.

The wrappers make the web sources function as databases that can be queried through the mediator's query language. It is impractical to construct wrappers for Web sources by hand for a number of reasons:

- Because the number of information sources of interest is very large.
- Because new sources of interest are added on the Web very frequently
- Because the format of the existing sources changes.

5.3 Kind of Web Information Sources

There are three kinds of Web information sources: multiple instance pages, single instance pages and free style pages. The three of them have challenges of their own in terms of how to access them or how their information has been structured.

The multiple-instance category has information provided from databases. These databases are queried directly. The disadvantage of this category is that the databases can only be accessed by means of a login or a license fee. On the other hand, once this hurdle is jumped over, the information can be accessed readily, and the building of the wrapper to query sources is feasible.

The single instance pages have semi-structured information on a single page. The free style pages, such as personal homepages, lack identifiable sections like headings. The use of graphs and images make it difficult to build a wrapper. Also, the extraction of information is difficult.

5.4 Automatic Wrapper Generation

The interest of this thesis is to attempt to use the automated processes of building wrappers for new web sources and to improve the business intelligence processes with this new kind of technologies for data extraction. The next steps describe the generation of a wrapper:

- Determine sections of interest on a webpage (See figure 5.4).
- Develop a parser that can extract selected regions from a webpage.
- Develop capabilities that communicate the wrapper, the mediator and the web sources in order to answer queries.

5.4.1 Sectioning the Source

To define the sections or structure of a Webpage, the following steps should be considered:

1. Identification of the sections of interest on a page. The sections will indicate the heading of a section. A heading is the beginning of a new section; thus through the identification of headings, it is possible to identify the section on a page.

2. Identification of the hierarchy with its sections. After a page has been dismantled in its sections, it is necessary to identify their nesting structure.

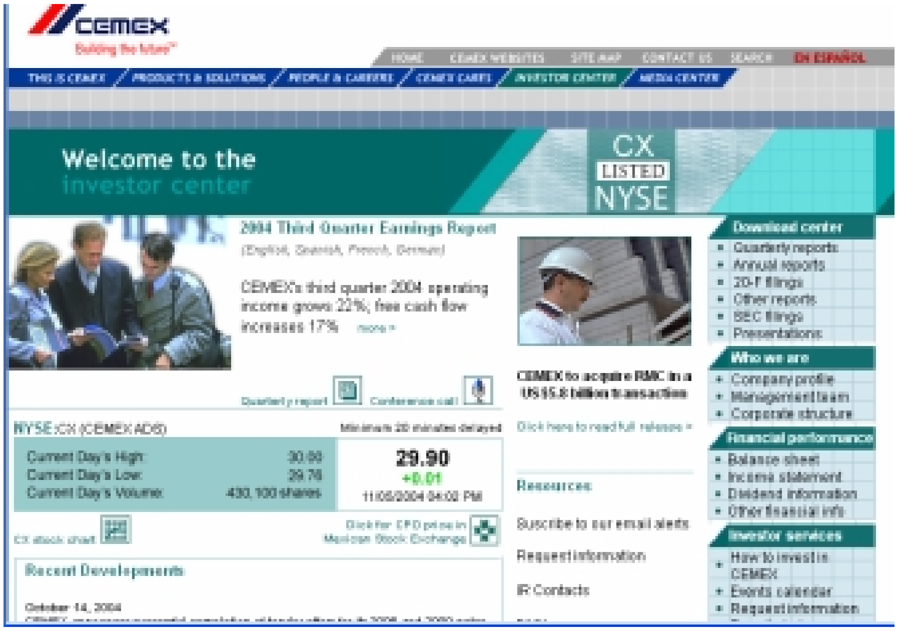
The main objective is that a computer program decomposes the HTML from the source and then it will be possible to identify the key sections on that page. The heuristics used for identifying important tokens on a page and the algorithms used to organize sections into a nested hierarchy are important for this work.

Once the section identification has ended, it is necessary to acquire the nesting hierarchy for the different sections of a page. The hierarchy will be obtained in an automatic form for a set of pages.

The process for querying the data on the Webpage requires that the answers be extracted from the semi-structured data stored in documents with an HTML format. It will be seen that these documents may contain data that will be not important to the query.

For the purpose of this section, the CEMEX (Cementos Mexicanos) webpage was considered. It offers investors-related information. An example of querying the webpage with an Investor Center binding is displayed in figure 5.4. Additional information about the recent developments and the NYSE is also showed in this webpage. This webpage supports a number of query bindings, e.g., country, customers.

For each set of allowed bindings on the CEMEX webpage, the translated query (URL) is different. It is required to specify the particular URL constructor expression for each binding on the subsequent web pages . For a query with an Investors binding, the URL is http://www.cemex.com/ic/ic_lp.asp. The URL for each of the bindings is different.



2004 Third Quarter Earnings Report

(English, Spanish, French, German)

CEMEX's third quarter 2004 operating income grows 22%; free cash flow increases 17% more »

NYSE:CX (CEMEX ADS) Minimum 20 minutes delayed

Current Day's High	30.00	29.90
Current Day's Low	29.76	+0.01
Current Day's Volume	430,100 shares	11/05/2004 04:02 PM

Click for CPO price in Mexican Stock Exchange

CEMEX to acquire RMC in a US $5.8 billion transaction

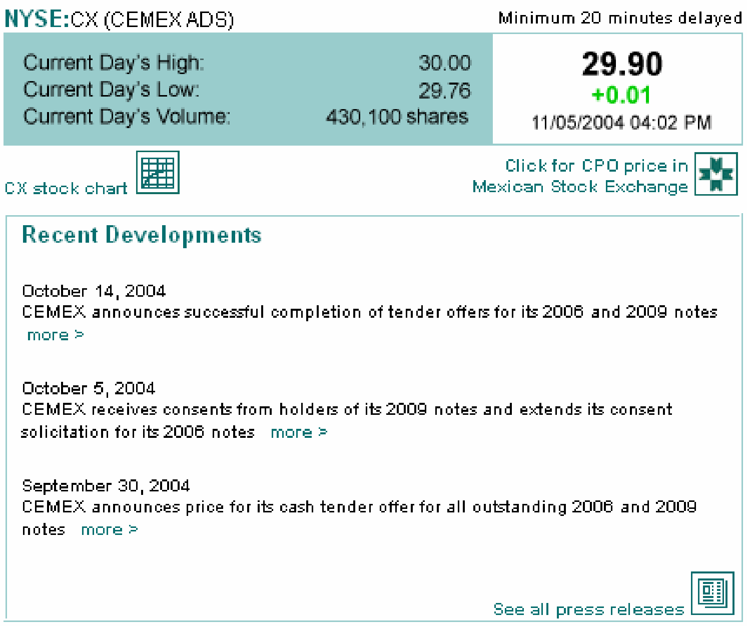
Click here to read full release »

Recent Developments

October 14, 2004

Figure 5.4 An example of querying the web source for the investor center with a set of bindings.

Finally, structure data corresponding to the output must be extracted from the replied HTML document. This webpage will most probably contain irrelevant data as well. Figure 5.5 shows a portion of an HTML document with the answer to the query to the Investors' Center.



NYSE:CX (CEMEX ADS) Minimum 20 minutes delayed

Current Day's High:	30.00	29.90
Current Day's Low:	29.76	+0.01
Current Day's Volume:	430,100 shares	11/05/2004 04:02 PM

CX stock chart

Click for CPO price in Mexican Stock Exchange

Recent Developments

October 14, 2004
CEMEX announces successful completion of tender offers for its 2006 and 2009 notes more >

October 5, 2004
CEMEX receives consents from holders of its 2009 notes and extends its consent solicitation for its 2006 notes more >

September 30, 2004
CEMEX announces price for its cash tender offer for all outstanding 2006 and 2009 notes more >

See all press releases

Figure 5.5 The result of the query for investor center binding

5.4.2 Protocol for wrappers

The protocol for wrappers is used by mediator systems. A protocol exports a schema and facilitates that the mediator can query an external data source.

It is necessary that the protocol has been developed to be integrated as part of the mediator query processing system. The protocol describes the meta information and the schema for the web source. One of the main objectives of the protocol is to provide the link between the query bindings and the output data. For this reason, the wrapper will provide the following information about the web source:

- *Input attributes*. They are the bindings accepted by the web source

- *Yield types*. They generate the model for the type of data produced by the wrapper.

- *Relationship between input and yield*. The input is a predefined subset of attributes and the output is the reciprocal process of the defined attributes. However, it can bring along more information since the query has analyzed all the possible pages. Thus, the relationship between input and yield defines the existing capability of the web sources.

5.5 Functionality of a Simple Extractor

The objective of an extractor is to analyze an HTML document and generate some objects of output specified in the corresponding row of the Table of Wrapper Capability. The information contained in an HTML document will be represented as a tree (See figure 5.6). With this ordered tree, the declarative queries may be written to get the data that will be extracted from the document. The queries will use expressions to identify nodes based on their contents; transverse the ordered tree; and extract data from the elements (nodes).

5.5.1 Data Model for HTML Documents

The information contained in an HTML document and its structure are described as a structured tree, in which the document is the root. The tree leaves are the HTML components such as list, table rows, table images, lists, frames, objects, options, scripts, etc. The tree branches represent the position of the HTML components within the documents. The document has its components enumerated orderly top-down in accordance with the HTML components. The representation of an HTML document allows to simplify the query language of extractors. [8,18,34,37,15,44,16,3]

Figure 5.6 is an example of an HTML document represented with a tree. It contains a list followed by two tables, each of which is a node. The components of these elements are also nodes such as lists (*li*), table rows (*tr*), and table data (*td*). One of these table rows (*tr2*), has table data (*td2),* element that has an *img1* element, which is also a node.

For this research project each of the web pages has a data model and each HTML document is described by a node in the ordered tree that will include the following attributes to conform the full HTML document:

- *Name*: This represents the type of HTML element such as: *table, tr, td, img, src*, etc.

- *Occurrence*: This is an integer representing the position of this object in the HTML document. In Figure 5.6, the second table object has an occurrence of 2, indicating that it is the second table. Also, it is the third child of the parent (root).

- *Data*: It represents the optional data associated with an HTML element. The data are the type of element such as: *img* or *scr* that have a URL as a parameter.

- *Child:* This relates the node to the list of objects embedded in it. It represents the branches between a node and its children in the tree. For example, the second occurrence of the HTML element table in Figure 5.6, has two children table rows *tr1* and *tr2*. The table row *tr2* has one child table data *td2*, which has one child of the type *img* and the image has a parameter .

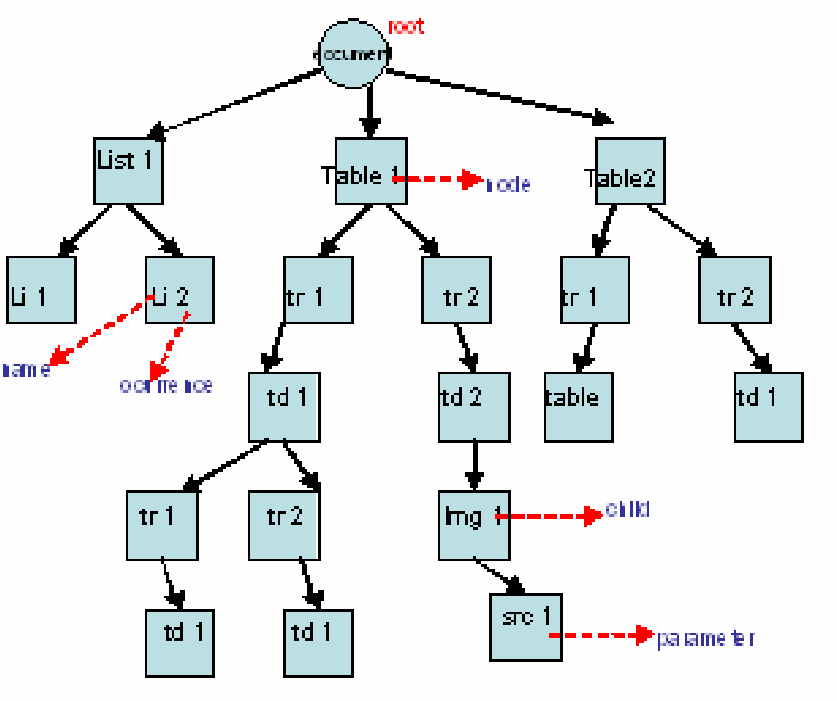


Figure 5.6 An ordered tree representation of an HTML document

5.6 Capabilities between the Wrapper and the Mediator

When a query is executed, a wrapper should bring the pages that contain the requested information from the web source. In order for there to exist communication between the wrapper and the mediator, some programs are required. It is important to mention that the wrapper and the mediator should be separated processes running in different places and the wrapper needs communication functionalities such as:

1. Identifying the URL of page or pages to answer queries. For single pages, the URL is always known by the wrapper. For web sources with more than one page, it is required to have a network between the query and the URL's.

2. Capability to make the HTTP connections with the web sources to retrieve data.

3. Communication between the mediator and the wrapper.

The previous functionalities are the last step in the wrapper generation process for a new web source. (See figure 5.7). The parser for pages from a web source and the above communication functionality results in a complete wrapper for that Web source [88,72,29,101].

5.7 Wrapper Tools.

The wrappers can be manually designed and for its development, some of the advanced programming languages such as java, python and c++ can be used; additionally, regular expressions should be used. This approach is possible for smaller applications. However, if it is needed to use a set of large numbers of wrappers, it is necessary to develop wrappers with a professional tool that facilitates the parameters definition for a given web source.

An important aspect to bear in mind for the wrapper generation is the format in which the extracted data can be exported. As a standard rule, the extracted data will be converted into an XML format, because XML facilitates the processing of a large number of software applications. In the case of this research work, an XML is used to sharing the information with the information systems inside Cemex through the Corporate Information System.

The tools for the generation of a wrapper can be categorized by their output methods, interface type, graphical user interface and other characteristics [59] and methods. Some tools are based on command lines and require some procedures generated by some scripting language to generate wrappers for specific web sources.

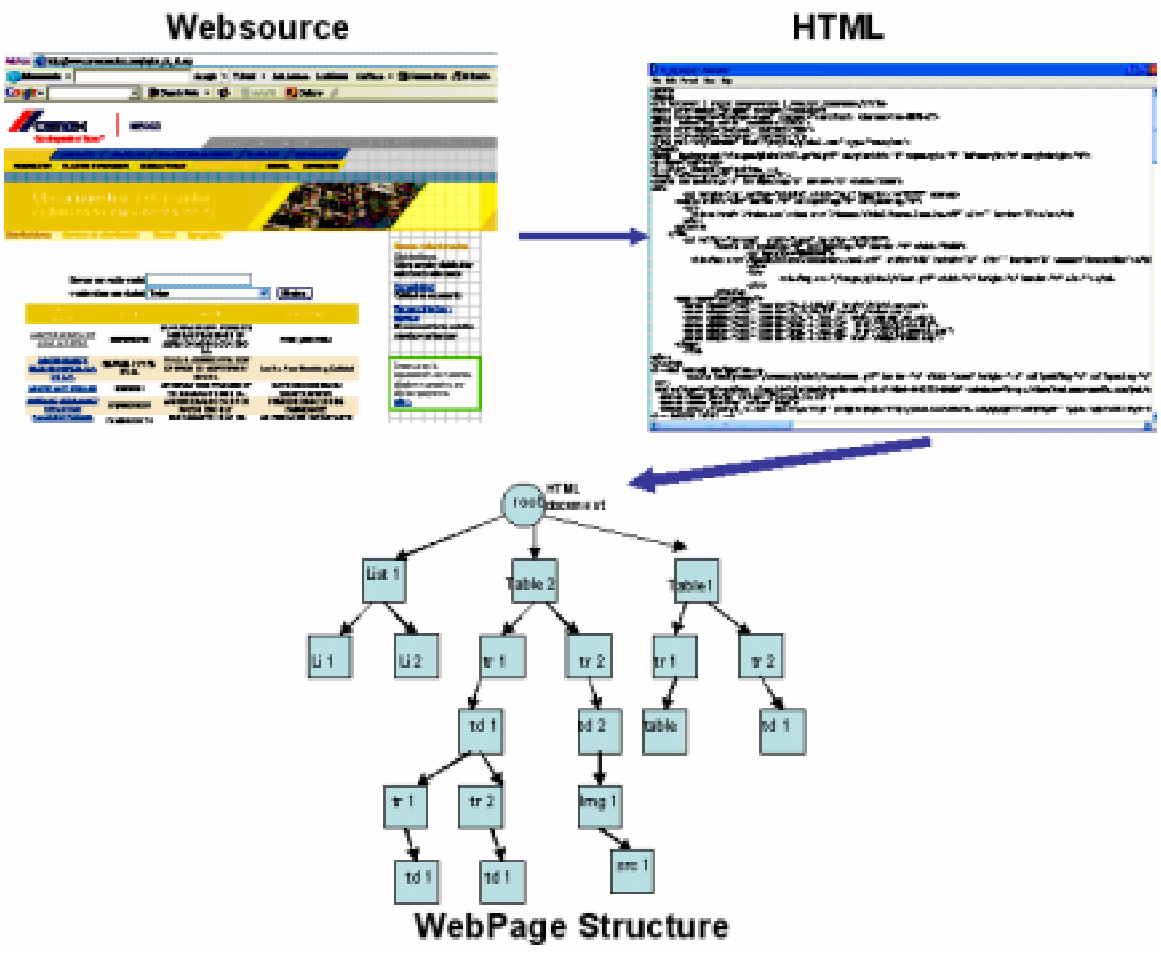


Figure 5.7 Workflow for the wrapper generation from a CEMEX web source

The script languages for wrapper development are used with text editors and are generated by specific general purpose programming languages such as: perl, java and python. Most of these tools offer a graphical user interface that allows to select relevant data, highlighted with a mouse click. Afterwards, the program will generate the wrapper based on the information previously selected [62]. In the case that the automatically generated results pop a NOT FOUND message for the specified requirements, the user has the option of implementing changes via an editor within the tool. Whether frequent corrections are necessary or not depends largely on the algorithms and the functionality of the tool.

The majority of the academic tools are common-line oriented and include some user support and help. Most of this help is given mainly to establish a network connection and submission of the HTML document. The tools provide simple scripting languages to develop the rules for the data extraction. Also, the evolution of the modern programming languages and the free software libraries with predefined classes simplify the generation of wrappers.

Most of the free software and academic tools do not provide the standard output format. They are limited in their use or area of application and require special interfaces for their communication with external applications.

In general, most of the commercial tools provide XML as the output format from the data extraction. They also include a graphic user interface and are click driven. In this classification, there are some professional tools as was mentioned in Chapter 4, such as: Lixto [47,9], Caesius Software, Crystal Software, Fetch Technologies, Wisoft and Kapow Technologies that do not need to specify all the data that are required to be extracted. Tools such as Lixto [47,9] only need a couple of examples to generate the rules of extraction, with Lixto the user does not need a previous HTML knowledge. Thus, the user can obtain a fully functional wrapper efficiently [10].

Many of the data extraction tools can generate wrappers that have the ability to retrieve single pages and some of them have the capabilities to retrieve multiple linked pages and in some cases enables the wrappers to retrieve a set of pages in a search request without mattering if the required data are contained in different HTML documents. The ability to trace another URLs links is used to retrieve the data contained in web pages or websites under study.

5.8 Wrapper Upkeep

The main objective of the wrapper maintenance is to maintain a wrapper effective and updated to have the capabilities and control of the source from which the data is retrieved. It was previously mentioned that a minimal change on the Web source will turn the wrapper unusable. Some websites are very dynamic and change often their structure. There are two activities to check if a wrapper is effective: first, determine if a wrapper is still working and opening correctly; second, determine if the extraction rules are effective and the web structure is the same. The process of verifying if a wrapper is still valid is not an easy task since the web source(s) may change in content or configuration. The wrapper verification should include both. [57].

Chapter 6

From Data Extraction to Information Modeling

This chapter is related to the topic of information modeling. Information modeling is the step that follows after the data from web pages or websites have been extracted. The process of information modeling is a key factor for Business Intelligence.This chapter will define the model driven development; it will also point out the definition considered for model, and the differences between model and replica. It will present two types of models, the token model and the type model; and finally it will discuss the issue if language definition should be considered a metamodel.

6.1 Model driven development

The last tendecies to model driven development bring up the important matter of defining the concept. The word “model” could have different meanings depending on who is defining it; thus the concept acquires different dimensions if used by pychologists, architects, engineers, researchers or mathematicians. This thesis focuses on the models that can be generated after extracting data from different web sources.

A “model” in software engineering is referred to as a modeling language as the one used in the UML [100]. This modeling language can describe a system through a set of diagram types. The model description is commonly graph-based.

Other software such as Java and Python are considered modeling languages as well. One of the steps in this work is to apply the principles of unification and modeling to provide benefits, for example: to interpret codes generated from such models as Unified Modeling Language (UML) or Logic Data Structure (LDS) to go to a “model to model” transformation.

Generally speaking, anything can be characterized as an “object”; however, the paradox is that technically speaking an object only provides value if everything else is not considered an object. For modeling purposes, it is necessary to define the boundaries of a “model”, so as it be large enough to unify but certainly small enough to exclude rendition. If that is made possible, then it can be indiscriminately applicable.

6.2 What is a Model?

One of the simplest definitions for model is provided in the Webster´s encyclopedic dictionary [68]:

a) A minuature representation of something or a pattern of something to be made.

b) A representation for imitation or emulation.

Stachowiak determined that a model must consider three key characteristics [45]:

***mapping* .-** It means that a model should be created considering the original one.

***reduction* .-** It means that a model only shows a relevant choice of the original´s attributes.

***pragmatic* .-** It means that a model needs to be functional without being the original for a specific function or role.

The characteristics of mapping and reduction merge when designing a model based on the original description and information. However, part of the information which is not relevant to the model is lost. To illustrate this, it can be considered the map of Mexico. From this map, the relevant information or abstracted information could be to locate states, cities, towns, neighborhoods, streets, numbers and zip-codes. Irrelevant information, that is, information that will be lost, could be oceans, mountains, rivers, highways and map scale among others.

Steinmüller sustains that a model is information about something (content, meaning), created by someone (sender), used by somebody (receiver) and utilized for some purpose (usage context). See figure 6.1.

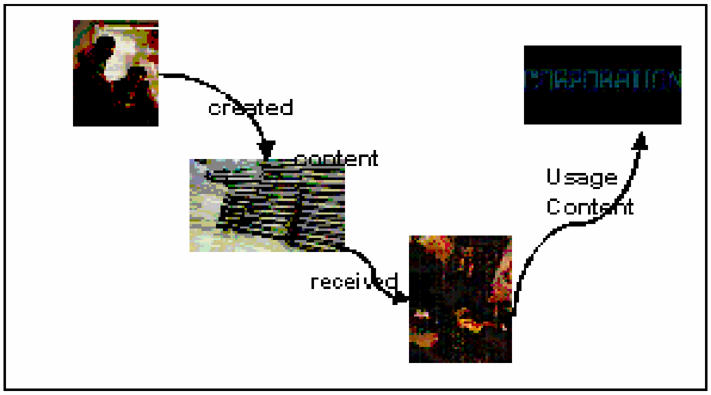


Figure 6.1 A model is information

To illustrate this kind of model, it is suggested to think in a scale airplane model. Using this model, some of the mechanical and physical attributes of the plane can be analyzed. In regards to abstraction, it can be said that if the model is "exact" it means that the model contains the attributes to be retained, but it doesn´t mean that the model is "complete", that is, it does not have all of the attributes.

6.2.1 A replica is not a Model

Suppose that it has been decided to build an airplaine exactly like the original one. In order to do that, it is necessary to be precise in all the details, and at the end, a model has not been developed, but a replica has. If we use the replica in an impact test, it will be seen that replicas never give the benefits of models in terms of cost reduction, nor their disvantages as imprecision in regards to the original.

6.2.2 Incentive for Modeling

In the process of model driven development, it is not necessary to use a physical model since the models are mainly linguistic. They are described in linguistic communication and use languages such as UML or LDS. Thus, it can be inferred that a model is a description of something.

It is possible to include some attributes to the definition previouly given, such as:

a) A model is a description that aids creating a mental image of something which is not feasible of being watched directly.

b) A model is a pattern of something to be built

Considering these new attributes, it will be possible to make simulations and construction plans.

Under the concept of software engineering, the models are created in the analysis stage and are used to sketch out the problem in a simple form using a spectrum of modeling approaches (See Figure 6.2), but more detailed forms are used in the construction plans in the design stage.

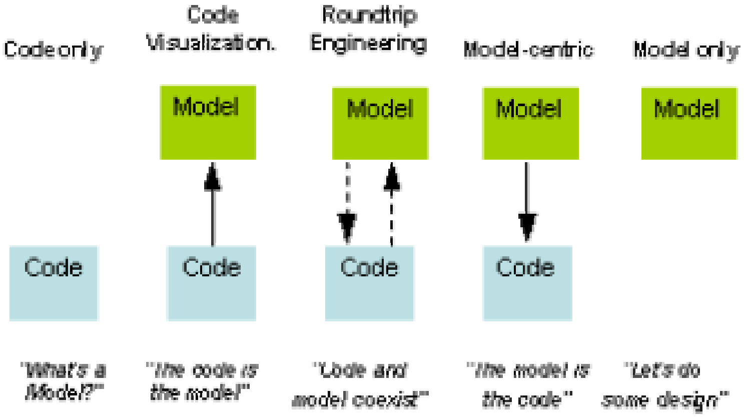


Figure 6.2 Spectrum of modeling approaches

It must be bore in mind that there are cases in which the original model does not exist because it is either under construction or imaginary.

6.2.3 So, What exactly is a Model?

Models can be transformed, but it doesn´t mean that everything transformable is a model. The characteristics of technical models must be adapted to also be used in “model driven development”. If it doesn´t have one of these characteristics, it cannot be a model.

It is also important to consider having an answer to such questions as:
What is the origin of modeling?
What kind of abstraction should take place?

The answer to these questions guarantees that the focus is a model and that it will be possible to unify the concepts of class, object, and method into one generalization called “pattern”.

The pattern could be nested within a parent pattern, within a nested class, within a method or even within a local object.

Thus, to conclude this section, a model is the representation of something that preserves some of the attributes of an original object or perception.

6.3 Type of Models

When two or more experts need to decide what type of model to use, they might have some differences in opinion. Prior to making their decision, they must first define what a model is, what transitivity is, and what is a metamodel. Once they have consolidated their points of view they can decide what model to use since there should not be two different models for the same purpose.

For the purpose of this thesis, two kinds of models will be analyzed and used for the data model driven development: the type and the token models.
In this work, a type model is a general representation of an object or perception which considers general parameters. A token model derives from a type model since it has more defined parameters of interest. However, a token model could be called a type model if it will go into further definition of parameters. Here, it is relevant to present to consideration the term transitivity. Transitivity refers to the possibility of representing attributes within the same type of model. It is worthwhile noting that it cannot be called transitivity if there is further definition of attributes which causes the creation of a token model, coming from a type model or vice versa. In such case it is called transformation.

6.3.1 Token Models

A token model is a map. (See Figure 6.3) The components of a map in a token model can show the most relevant elements of the original map. These elements will be arranged in object diagrams by using UML or LDS.

With the UML, the token models are referred to as snapshot models or representational models. This kind of models could capture a simple configuration of a dynamic system.

In the interest of modeling the data extracted from some of the competitors' web pages (warehouses and distribution center lists), a map will be created. Such map will stand for an abstraction of reality and will have some attributes such as used scale, geographic information, and oceans among others.

Most of the times, the models are an abstraction of reality or an abstraction of some other models, causing that in some cases all the information of the original model is not provided and others is irrelevant. The map of Figure 6.3 is a model of a model, so we can call it a "**metamodel**". "**Meta**" means that an activity has been executed twice.

The use of the prefix "meta" suggests that the repetition of operations is not transitive. To illustrate this concept, it can be considered that when generalizing twice, the result is called "superclass", "super superclass" or "metaclass". Any operation of generalization that creates a transitivity relationship between its elements does not need the repeated applications represented with the prefix "meta." Transitivity is something expressed as a gramatical category by means of verb morphology.

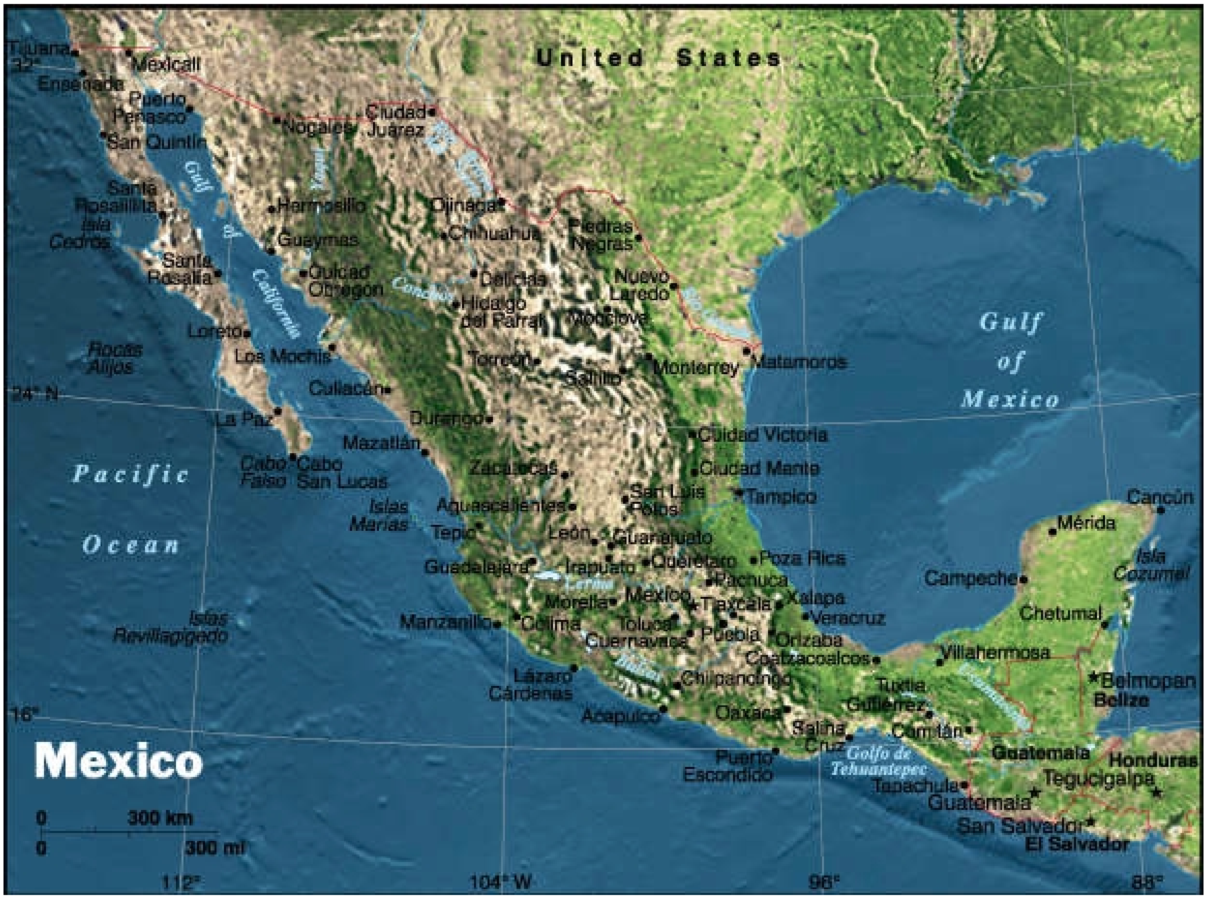


Figure 6.3 Transitivity Model and Token Model

The relationship “representedBy” means that it is transitive for token models. In this study case, the map used (See figure 6.3) derived from a finer map which is a legal model from the original. With these arguments, it can be said that a token model derived from another token model is not a metamodel. The process of transforming a model to another form is known as “model transformation”.

The token models are functional in regards to getting the configuration of a system for a simulation process. The best way to describe a token model is by considering the picture that people have in their minds. So, people will utilize examples such as diagrams or designs of an airplane, both are token models. For this reason, it is important to be explicit about this definition and discuss the differences between a token model and a “type model”.

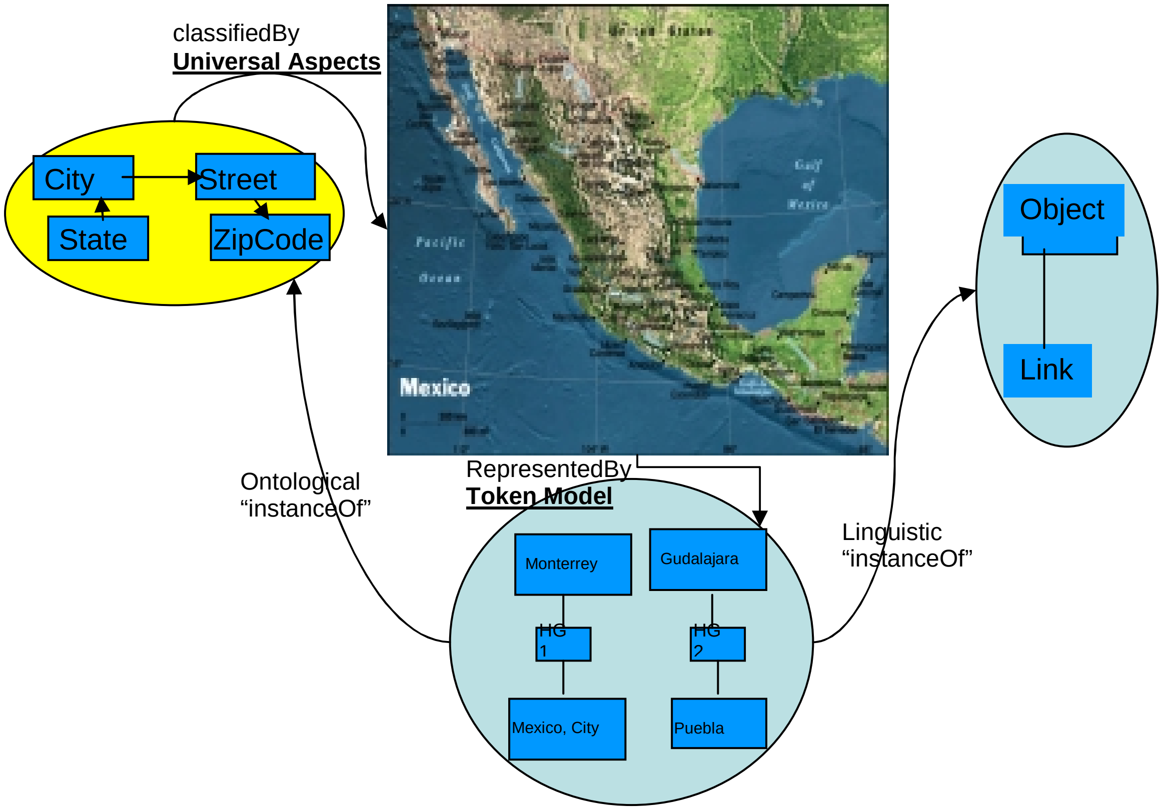


Figure 6.4 Kind of model (classifiedBy & RepresentedBy)

6.3.2. Type Models

Nowadays, it is possible to extrapolate the power of a “type models” through its object properties for its classification and then generate some conclusions. With this approach, it is not necessary to memorize all the facts although it is necessary to collect concepts and general properties [54].

A “type model” grasps the general aspects of the original object or perception. In general, “type models” are the most used models in the model driven development. Type models are not oriented to catch the singular aspects as a “token model” is. Figure 6.4 shows a “type model” of Mexico, using UML for the generation of the natural diagrams for “type models.

”The “type model” of Figure 6.4 is not only used as a “type model” of the original, but it can also be used as a “token model” produced from the original. In this case, the “type model” is a model of a model but it can not be named a “metamodel” because both models come from the original model. If we generate a “type model” of a “type model” (See figure 6.4) that contains elements such as “ConnectorType” with instances such as “street”, “city”, etc; then it can be called a metamodel.

When talking about a metamodel created by applying the classification twice, the issue of “deep characterization” arises[23]. Deep characterization is a transformation of models. In the case of token models, each individual model in a chain of models will always characterize the original since the relationship “modelOf” is transitive.

A different way of generating a model of the “token model” that is not a model of the original in Figure 6.4, is by modeling the properties of the token model itself, and not of its content. The bottom part of the (representedBY) model in Figure 6.4 represents the language used to create the token model. As it was mentioned previously, Figure 6.4 is a metamodel since it is the model of a model which did not consider the original one. In this case, to describe the model, the UML language definition is designated by the UML metamodel [75].

The bottom model of Figure 6.4 is a classification model that represents a metamodel under the language used in the token model, and the right side model is a classification model (classifiedBy) in terms of the token model´s content. Mathematical modeling authors Atkinson and Kühne recommended to make a differentiation between the linguistic and the ontological classifications [23].

The bottom model describes the format of the token model. The right hand side model defines the representation of the token model. This is a way to verify if the “token model” has valid “sentences” of the language in the “ontological” model type. For modeling purposes, it is necessary to bear in mind that there are two kinds of model languages: “content language” and “representational language” This classification shows the relationship between “language definition” and “models.”

It is important to mention that a type model can be reused as a “token model.” Figure 6.5 shows that a UML diagram can be reused as a “type model” for the implementation of a programming language such as Java or Python. The class diagrams represent the relevant aspects of Java or Python classes on a one-to-one mapping.

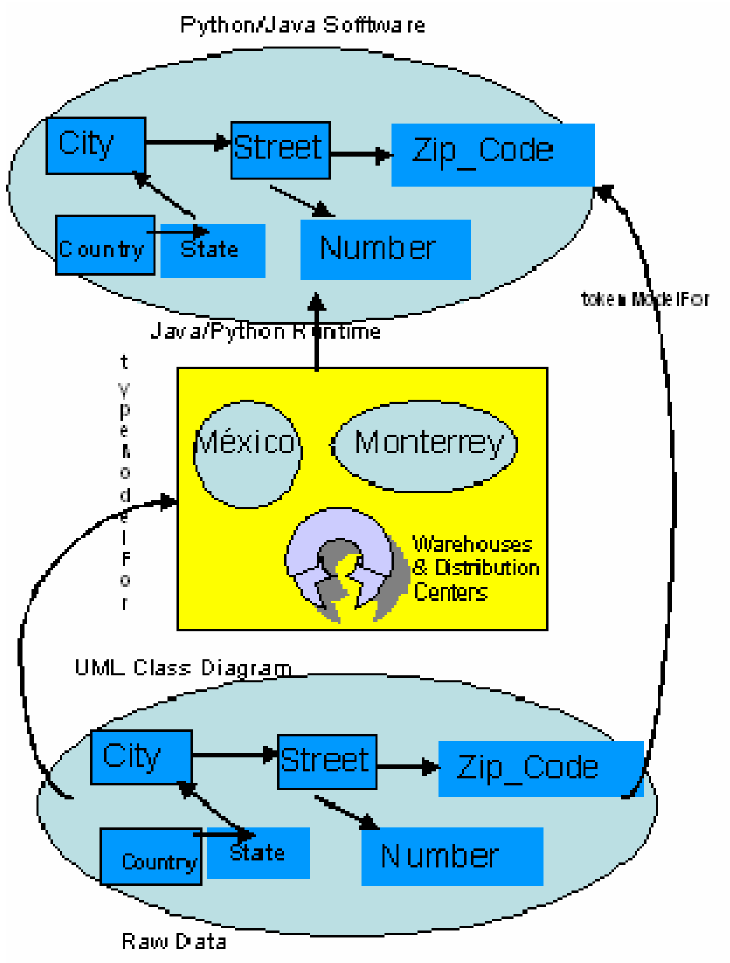


Figure 6.5 Software Models

6.4 Is Language Definition a Metamodel?

Metamodels can be represented in both a linguistic and an ontological way as language definitions. Figure 6.6 shows the connectivity from a system, to its model, to the related language utilized to compose the model. Also, it is possible to characterize the connectivity between a model and its metamodels as "intancesOf" instead of "conformsTo".

The utilization of a "metamodel" is justified with the complex syntax definition of the language. The aspects of semantic and syntax are left out. There are some aspects of the abstract syntax that are not needed to be specified.

It can be concluded that a "metamodel" is a model with simplified features compared to the language used in a token model. In a nutshell, a "metamodel" preserves the essence of any model described in it.

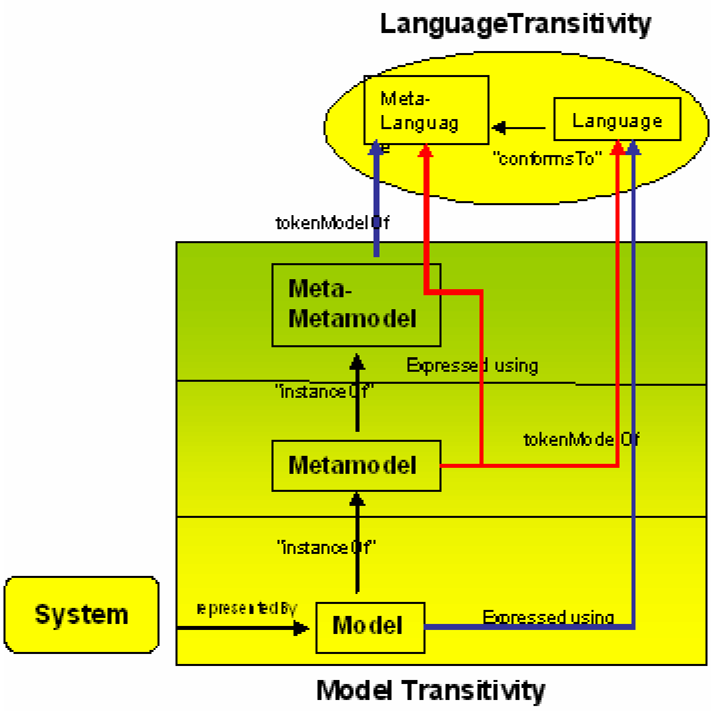
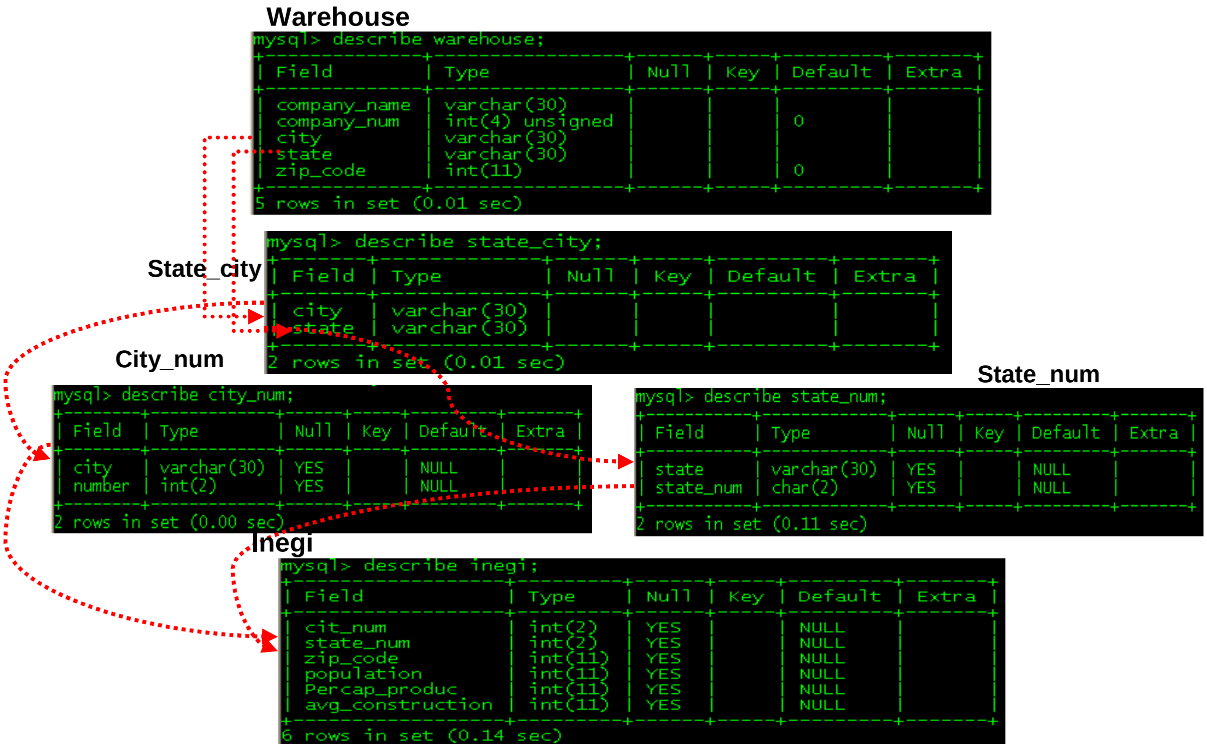


Figure 6.6 Metamodels & Metalanguage definitions

With the previous Metamodel and language definitions, it was possible to determine the table definition for the raw data and its interaction with other tables coming from the same database, as seen in Figure 6,7.

With this language definition, it is also possible to start the configuration of the enterprise modeling considering the different entities that are part of the whole enterprise model as it will be discussed in the following chapter.



```
mysql> describe warehouse;
+--------------+------------------+------+-----+---------+-------+
| Field        | Type             | Null | Key | Default | Extra |
+--------------+------------------+------+-----+---------+-------+
| company_name | varchar(30)      |      |     |         |       |
| company_num  | int(4) unsigned  |      |     | 0       |       |
| city         | varchar(30)      |      |     |         |       |
| state        | varchar(30)      |      |     |         |       |
| zip_code     | int(11)          |      |     | 0       |       |
+--------------+------------------+------+-----+---------+-------+
5 rows in set (0.01 sec)
```

```
mysql> describe state_city;
+-------+-------------+------+-----+---------+-------+
| Field | Type        | Null | Key | Default | Extra |
+-------+-------------+------+-----+---------+-------+
| city  | varchar(30) |      |     |         |       |
| state | varchar(30) |      |     |         |       |
+-------+-------------+------+-----+---------+-------+
2 rows in set (0.01 sec)
```

```
mysql> describe city_num;
+--------+-------------+------+-----+---------+-------+
| Field  | Type        | Null | Key | Default | Extra |
+--------+-------------+------+-----+---------+-------+
| city   | varchar(30) | YES  |     | NULL    |       |
| number | int(2)      | YES  |     | NULL    |       |
+--------+-------------+------+-----+---------+-------+
2 rows in set (0.00 sec)
```

```
mysql> describe state_num;
+-----------+-------------+------+-----+---------+-------+
| Field     | Type        | Null | Key | Default | Extra |
+-----------+-------------+------+-----+---------+-------+
| state     | varchar(30) | YES  |     | NULL    |       |
| state_num | char(2)     | YES  |     | NULL    |       |
+-----------+-------------+------+-----+---------+-------+
2 rows in set (0.11 sec)
```

```
mysql> describe inegi;
+------------------+---------+------+-----+---------+-------+
| Field            | Type    | Null | Key | Default | Extra |
+------------------+---------+------+-----+---------+-------+
| cit_num          | int(2)  | YES  |     | NULL    |       |
| state_num        | int(2)  | YES  |     | NULL    |       |
| zip_code         | int(11) | YES  |     | NULL    |       |
| population       | int(11) | YES  |     | NULL    |       |
| Percap_produc    | int(11) | YES  |     | NULL    |       |
| avg_construction | int(11) | YES  |     | NULL    |       |
+------------------+---------+------+-----+---------+-------+
6 rows in set (0.14 sec)
```

Figure 6.7 Warehouse table definitions and its relations

Chapter 7

Visual Business Intelligence Modeling using Lixto

Today, there are different techniques and tools to guide the process for wrapper generation and automated web information extraction. For this research work, it was decided to use the innovative Lixto tool [62].

Lixto provides the possibilities of generating wrappers by means of using the extraction of chunks of information from HTML pages and placing them in an XML format. Lixto is a user-friendly tool that assists in the wrapper generation process in a semi-automatic way providing a fully visual and interactive user interface.

7.1 Lixto Visual Wrapper Generation

As it is known, the information in the World Wide Web is mainly formated in HTML. In the future, the new web pages will have an XHTML format. This fashion will be seen in the medium and long ranges. This new trend will be of the benefit of all internet users (programmers, corporations and final users). The two languages HTML and XML are both oriented to the representation of semistructured data; HTML is oriented towards representation and visualization, whereas XML is oriented to manipulation and database applications [102].

The principal activity to get data from a webpage or a group of webpages is to use wrapper technology to extract relevant information from HTML documents and export it into XML. With the XML, the information will be queried and processed. Lixto[62] uses a new method of extracting relevant information from the HTML documents and then translates it to XML format. Lixto is well-suited for building HTML/XML wrappers.

After a wrapper has been built, the process of extracting information is continuous. It is applied automatically and the same wrapper will be applied to pages that are changing very often in content but not in structure. Only when the webpages have big changes in structure is it necessary to redesign the wrapper.

Nowadays, the Lixto tool interface is considered friendly. Lixto is very interactive, so it is possible to track the data needed to be extracted during the whole process of wrapper generation [82].

The most importat characteristics of Lixto are:
1.- Lixto is easy to learn and use; this tool is user oriented.
2.- Lixto uses easy marking to facilitate its use by those without experience
  with HTML to work in generating wrappers.
3.- Lixto lets a wrapper designer work directly on the browser starting with some
  sample pages.

With Lixto, it is relatively easy to do a wrapper through the Lixto Visual Wrapper tool. It allows extraction focusing on target patterns based around landmarks, on the contents itself, on HTML attributes, and / or on syntactic concepts. Lixto even allows for more advanced features such as looking for disjunctive pattern definition, crawling to other pages during the extraction, and recursive wrapping. This application uses datalog-like semantics and logical semantics.

An Elog application is a set of datalog rules containing special extraction conditions. Elog is not only transparent to the Lixto user, but it is also is easily deductible and flexible.

Figure 7.1 The fist screen of Lixto Visual Wrapper

7.2 The Visual Interface

Lixto is friendly in the generation of wrappers. The system guides all the process step by step. The first screen on the system consists of a set of menus (See Figure 7.1), thus, the first step is to load the example pages on the Lixto local browser.

To load the example webpage to the Lixto browser, it is required to know the URL of the webpage to be loaded. Then, the Lixto workspace will load this sample page and show this (see figure 7.2).

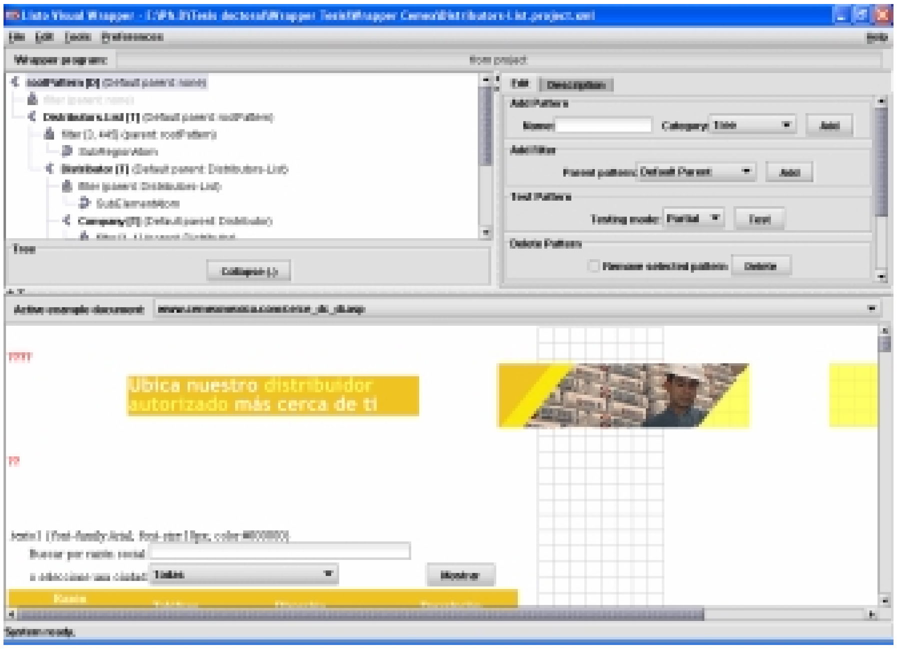


Figure 7.2 Competitor webpage loaded in the Lixto Visual Wrapper

7.3 Wrapper Generation with Lixto

Lixto generates wrappers using sequential hierarchical patterns out of the natural information structure. In this thesis for example, the list of warehouses of competitors' web pages will be extracted from the following URL.

http://www.holcim.com/MX/MEX/module/gnm0/jsp/templates/location/location_main_for_mexico.html (See Figure 7.3)

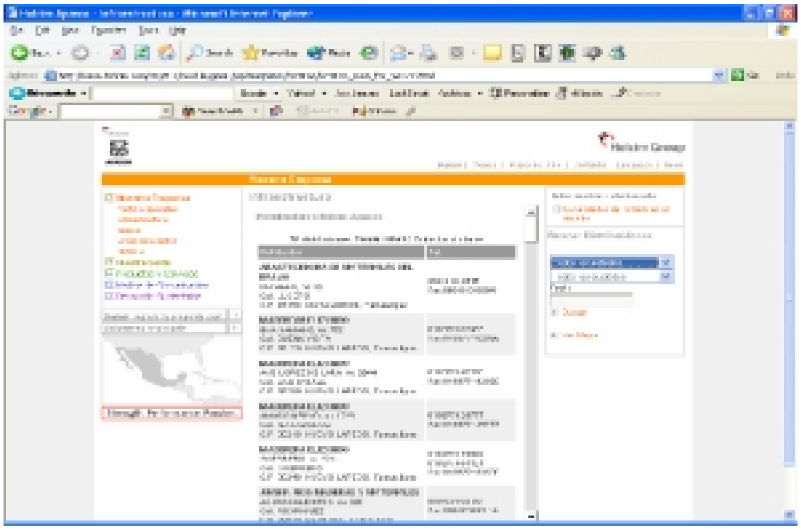
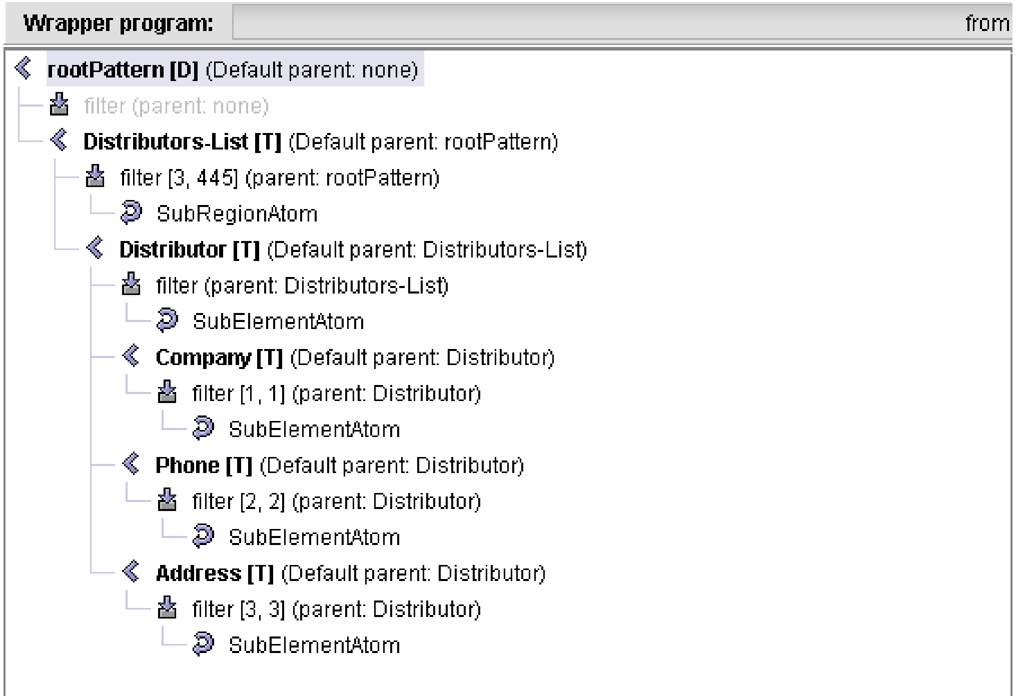


Figure 7.3 Competitor warehouses list in the webpage

With the structure of the Competitor's Webpage, different patterns to identify the information to be extracted can be defined. So the following have been considered: *pattern* **<Distributors-List>; then** a *subpattern1* **<Distributor>** and finally a *subpattern2* **<Company>.** The *subpattern1* is generated from a list of distributors and the *subpattern2* means that for each distributor there are a company, phone, address, ect (See Figure 7.4).



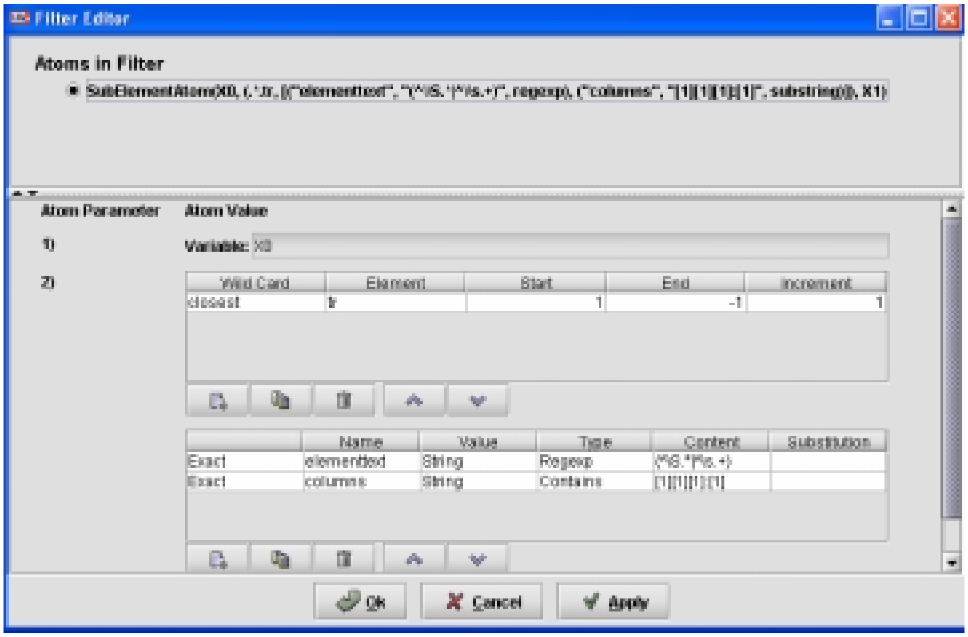
```
Wrapper program:                                          from
rootPattern [D] (Default parent: none)
  filter (parent: none)
  Distributors-List [T] (Default parent: rootPattern)
    filter [3, 445] (parent: rootPattern)
      SubRegionAtom
    Distributor [T] (Default parent: Distributors-List)
      filter (parent: Distributors-List)
        SubElementAtom
      Company [T] (Default parent: Distributor)
        filter [1, 1] (parent: Distributor)
          SubElementAtom
      Phone [T] (Default parent: Distributor)
        filter [2, 2] (parent: Distributor)
          SubElementAtom
      Address [T] (Default parent: Distributor)
        filter [3, 3] (parent: Distributor)
          SubElementAtom
```

Figure 7.4 Wrapper structure generated in lixto

The Pattern names in the wrapper will be the same as in the XML pattern names. Each pattern is defined by one or more filters. A filter allows the system to identify a set of similar nodes of the HTML parse tree (see Figure 7.5).

To create a filter it is necessary to follow the next two steps [82]:

1.- The user can use the mouse to select the relevant information of the target pattern. This process occurs within the example page.
2.- Then, it is necessary to add some conditions to the filter in order to set the boundaries of instances of the desired target.



Atoms in Filter

● SubElementAtom(X0, (,'.tr., [("elementtext", "\(^\S.*)\s.+)", regexp), ("columns", "[1][1][1][1]", substring)]), X1)

Atom Parameter	Atom Value
1)	Variable: X0
2)	

Wild Card	Element	Start	End	Increment
closest	tr	1	-1	1

	Name	Value	Type	Content	Substitution
Exact	elementtext	String	Regexp	(^\S.*)\s.+)	
Exact	columns	String	Contains	[1][1][1][1]	

Figure 7.5 Filters editor generation

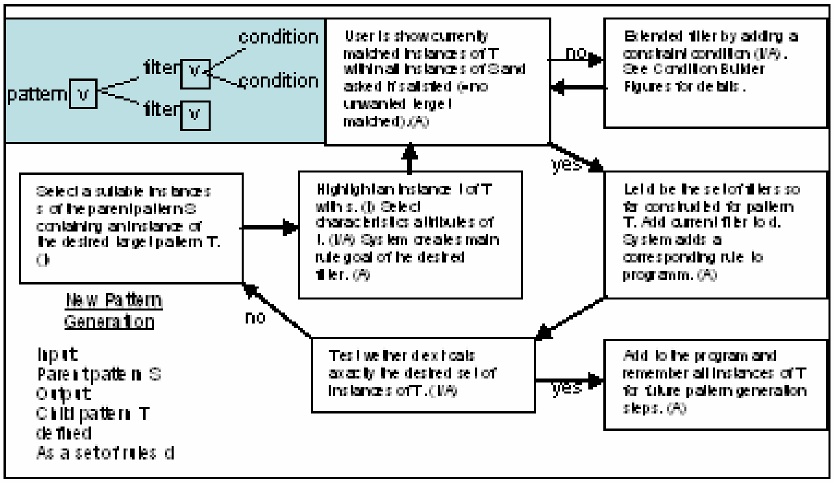
The conditions that are allowed for filtering are the following :

a. Before/ after condition
b. Not before/ not after conditions
c. Internal conditions
d. Range conditions

Additionally, by providing more conditions and new filters it can (also) extract the desired information from the sample page. The workflow for a pattern generation[1] is described in Figure.7.6.

7.4 Pattern Building

As previously mentioned, the user can define and refine a pattern hierarchically. At the beginning, the user can enter a pattern name, but s/he then needs to specify the parent pattern. All of this is done by selecting the information with the mouse and by clicking on the sample instances of the parent pattern and by marking them. When the user starts a new process to generate a pattern, the first document is the HTML document, and the first pattern is the <document>. This pattern has an only one instance that is the actual example document. A pattern may consist of several filters. Each filter has a number of conditions.


condition
filter
v
condition
pattern
v
filter
v
User is shown currently matched instances of T within all instances of S and asked if satisfied (=no unwanted target matched).(A)
no
Extended filter by adding a constraint condition (I/A). See Condition Builder Figures for details.
yes
Select a suitable instances s of the parent pattern S containing an instance of the desired target pattern T. (I)
Highlight an instance I of T with s. (I) Select characteristic attributes of I. (I/A) System creates main rule goal of the desired filter. (A)
Let d be the set of filters so far constructed for pattern T. Add current filter to d. System adds a corresponding rule to programm. (A)
New Pattern Generation
no
Input
Parent pattern S
Output
Child pattern T
defined
As a set of rules d
Test whether d extracts exactly the desired set of instances of T. (I/A)
yes
Add to the program and remember all instances of T for future pattern generation steps. (A)

Source: Visual Web Information Extraction with Lixto, Proceedings of the 27th VLDB Conference, Roma , Italy, 2001.

Figure 7.6 Diagram for a new pattern generation

Through the system, the user can generate one of two basic filters for data extraction. The first is through the selection of the data pattern directly from the sample web page. If the wrapper does not comply with the condition of extracting only the information asked for, thus bringing along additional, non-required information, it will be necessary to generate additional filters to restrict the extraction of the data required. The second kind of filter is by directly generating extracting patterns using advanced techniques of data extraction which applies regular expressions. Once the pattern has been created, the user can use this pattern as a parent for a new pattern. The final extraction is built over a pattern of a pattern, until the system reaches the information that the user wants to extract.

7.5 XML Raw Data

After the wrapper has extracted the information and has applied the respective filters, Lixto Visual Wrapper generates an output file in XML format (See Figure 7.7).

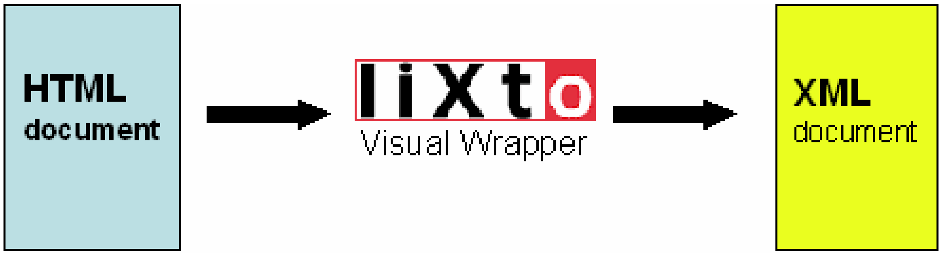
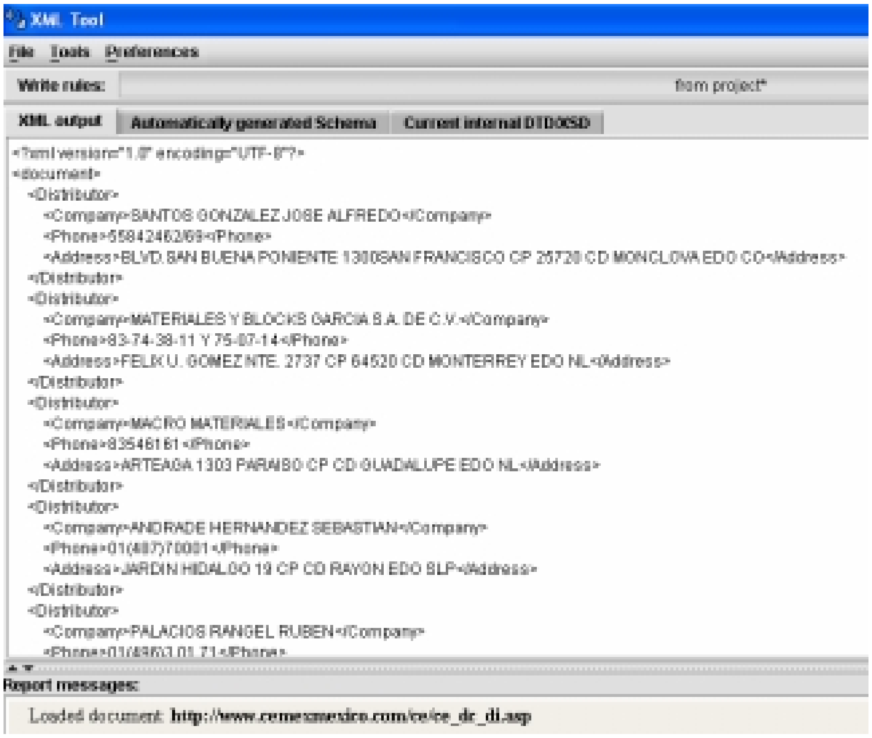


Figure 7.7 From HTML to XML through Lixto Visual Wrapper

It can be seen that the XML document generated (see Apendix A) uses the same tags and subtags previously defined in the configuration of the wrapper (See Figure 7.8.) With the new XML document, it is possible to translate it into The Business Intelligence Repository (BIR).



Write rules: from project*

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
  <Distributor>
    <Company>SANTOS GONZALEZ JOSE ALFREDO</Company>
    <Phone>55842462/69</Phone>
    <Address>BLVD. SAN BUENA PONIENTE 1300SAN FRANCISCO CP 25720 CD MONCLOVA EDO CO</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES Y BLOCKS GARCIA S.A. DE C.V.</Company>
    <Phone>83-74-38-11 Y 75-07-14</Phone>
    <Address>FELIX U. GOMEZ NTE. 2737 CP 64520 CD MONTERREY EDO NL</Address>
  </Distributor>
  <Distributor>
    <Company>MACRO MATERIALES</Company>
    <Phone>83546161</Phone>
    <Address>ARTEAGA 1303 PARAISO CP CD GUADALUPE EDO NL</Address>
  </Distributor>
  <Distributor>
    <Company>ANDRADE HERNANDEZ SEBASTIAN</Company>
    <Phone>01(487)70001</Phone>
    <Address>JARDIN HIDALGO 19 CP CD RAYON EDO SLP</Address>
  </Distributor>
  <Distributor>
    <Company>PALACIOS RANGEL RUBEN</Company>
    <Phone>01(486)3 01 71</Phone>
```

Report messages:

Loaded document: http://www.cemexmexico.com/ce/ce_dr_di.asp

Figure 7.8 XML document obtained after wrapper execution

To generate the BIR, another tool must be used, the Lixto Transformation Server. With this server, it is possible to define some pipes to integrate the different wrappers coming from different competitors' web pages and then to obtain a unique database or a set of databases for each competitor of the corporation. In this case, the information integrates all the warehouses and distribution centers of each competitor in Mexico. However, the business intelligence process can be extended to include other countries.

For the unification of information, different wrappers are used because the web pages of each competitor has a different structure (See Figure 5.6.) The purpose of this process is to unify the information in a single and handy structured set of data. The workflow to generate the unification of the information, from semi-structured data to structured data, is demonstrated in Figure 7.9. The unified information allows the user to work on structured data.

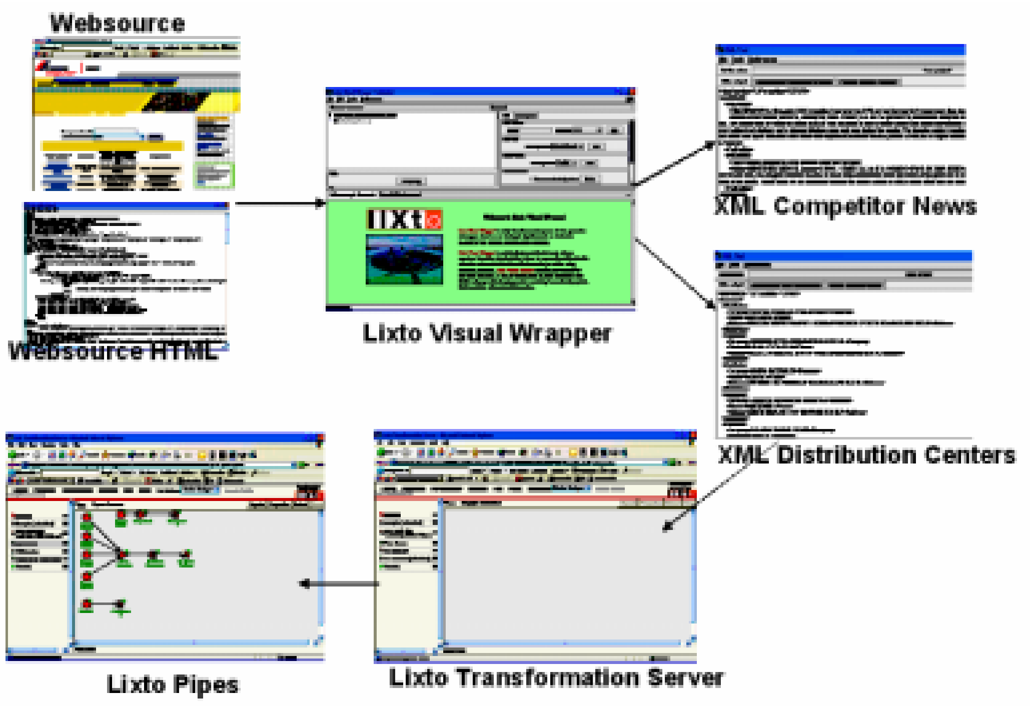


Figure 7.9 Workflow for unification of information using Transformation Server

The relevant role of pipes (See Figure 7.10) resides in the graphic interface of the Lixto Transformation Server. This server creates the workflow in charge of transforming the XML extracted information into a MySQL structured database.

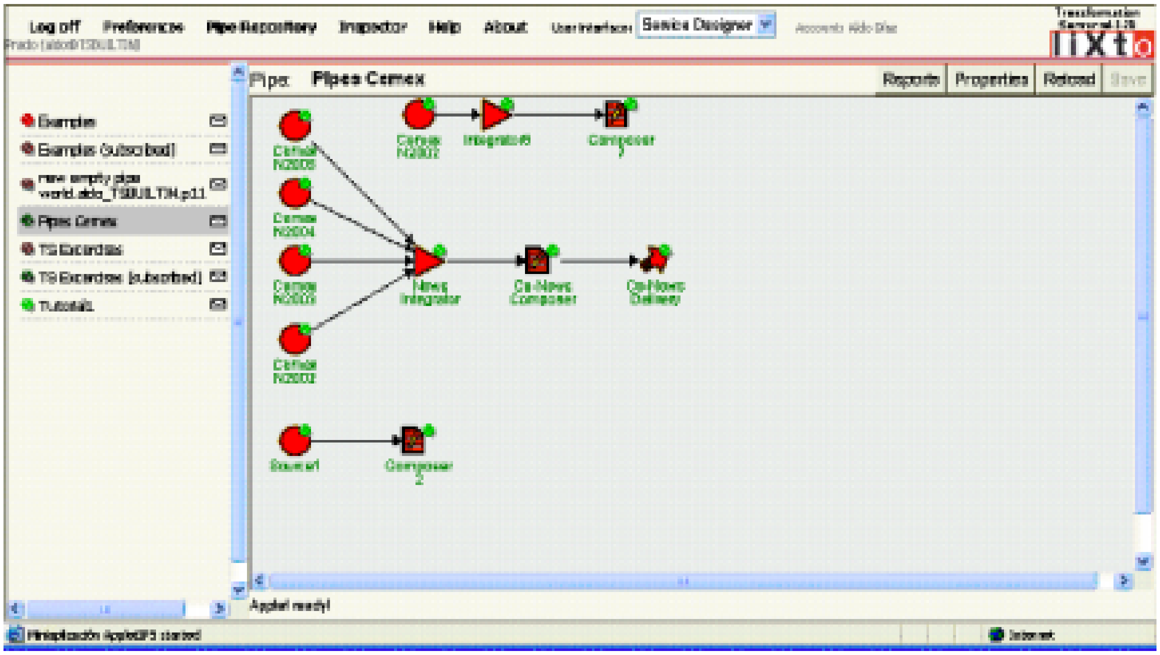


Figure 7.10 Pipes workflow to generate and update databases

7.6 Extrapolating an XML to Relational Databases.

XML is frequently used as an interface to work with relational databases. In this scenario, XML documents are exported from relational databases and published, or used as the internal representation in user applications. This fact has stimulated much research in exploring and querying data as XML views [35,98,97,20]. On the other hand, updating a relational database via an XML view has been somewhat overlooked by the industry and the academy. With the Lixto Transformation Server the information can be updated and delivered to the target platform through pipes (See Figure 7.10).

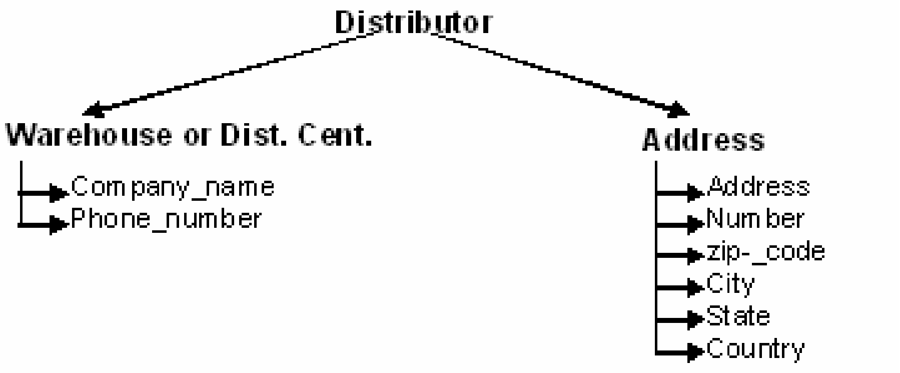


Figure 7.11 Distributor query tree

Business Intelligence applying Lixto can deliver information to a number of different clients (See Figure 7.12) ranging from mobile phones to PDA. Here, the application is delivered to a database in MySQL [71].

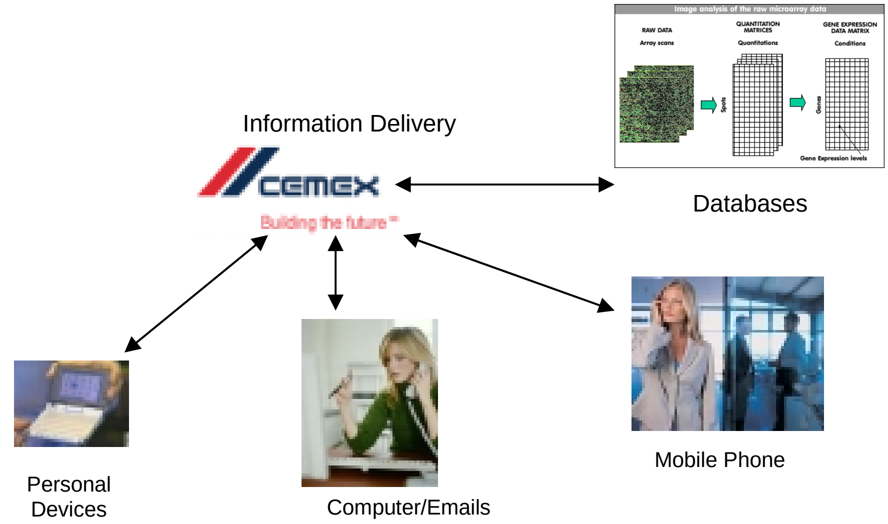


Figure 7.12 Information delivery with Lixto

Focusing on a common form of an XML view, it is shown that it allows nesting, composition of attributes, heterogeneous sets and repeated elements (See Figure 5.3.) The database model was generated from the query tree (See Figure 7.13). For additional information about the databases see Appendix B.

Distributor (Company , Address)

Company (Company name , Phone number)

Address (Address, Number, Zip code, City, State, Country)

Figure 7.13 Distributors' sample database & relational schema

XML view expressions are represented as query trees (See Figure 7.11). Query trees can be thought of as the intermediate representation of a query expressed by some high-level XML query language, and can provide a language independent enough to capture the XML views that have been encountered in practice, yet simple enough to understand and manipulate [21].

This strategy is similar to that adopted in [15] for XML views constructed, using nested relational algebra (NRA); however, the view and update language used here are far more general. Particularly, nested relations cannot handle heterogeneity.

7.7 Repositories.

Generally speaking, the industry prefers using XML to integrate application using XML messages and to describe metadata. Thus, a hard issue is on top of the table. How can information be stored in XML and then be reused?

7.7.1 XML Repositories

For this thesis, the option of using an XML repository for retrieving XML was analyzed because it is able to query the XML in contrast to an SQL which does not have the possibility to make queries directly to an XML repository.

So, why use a repository? At the end of the day, two reasons are present.

Reason one. If the user has really large XML files and s/he wants to retrieve small parts, then only an XML will impact with the costs. Is this common? Well, for large messages it appears to be increasingly frequent whereas for the metadata not so often. Most of the metadata seen in XML isn’t that big. However, for messages of considerable size, speed is also critical, which is why people avoid the parse cost. So, can it be afforded to put the XML message into the repository? It is feasible, however it is necessary to have several pieces retrieved numerous times and in somewhat unpredictable ways.

Reason two. For transactble information and speed of retrieval, the format is characterized by its unpredictability, hence, it will not turn into columns the elements and attributes queried. Thus, only an XML repository will work.

Ironically, SQL databases are not suitable to be used here. Since it cannot be predicted what elements or attributes will be queried, the user is forced to use text searching. Current text searching combined with the file system is better and faster. So, if the user needs transactions, there is already an argument for an XML repository. Also, researchers may be looking for deeply nested information, in which case there are two choices: Either mapping the XML into a set of related

SQL tables and turning the XML query into a SQL query; or putting the XML into an XML repository and using an Xpath to retrive the nested information.

In this work, it was considered that the best alternative is to transform the XML to structured data and generate a Business Intelligence Repository with MySQL. One of the most important reasons is that data manipulation will be easy and the transactional costs will be relatively low.

7.7.2 Business Intelligences Repository

This work proposes that the XML information, extracted through Lixto, be transferred to a relational database (MySQL). The information kept in the Business Intelligence Repository serves well to do the analysis on competitors and it can also be resutilized in some other applications (transactional or internal) within or outside the corporation boundaries.

As it can be seen in Figure 7.14, the Business Intelligence Repository is the First Level of the Business Intelligence Process. At the First Level of Business Intelligence, information related to competitors can be stored and updated continuously. Also, this information lends itself to further analysis by means of using specialized tools and information systems.

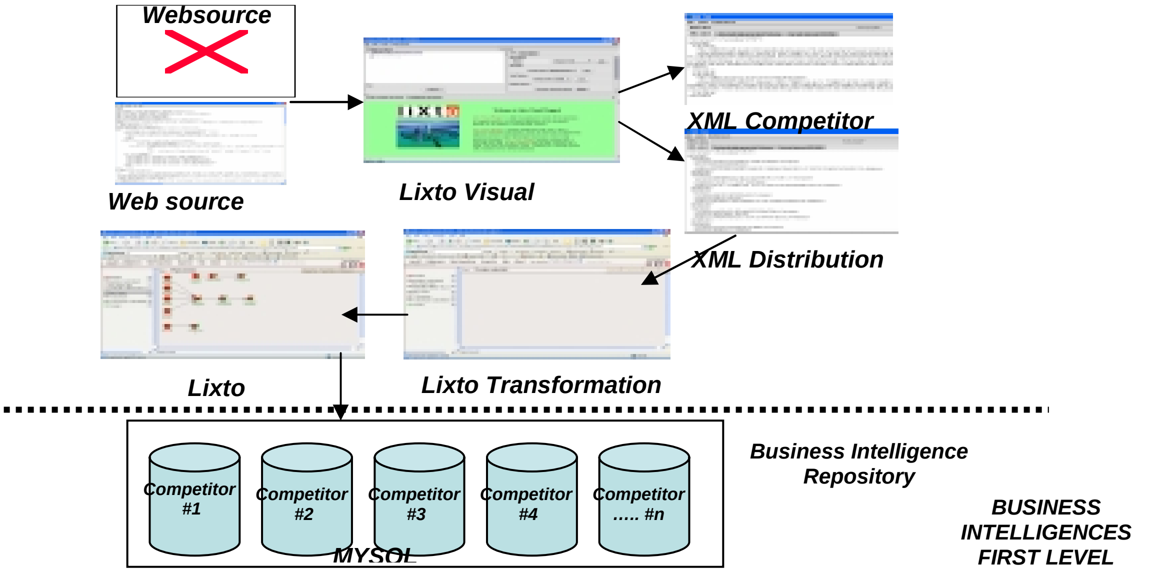


Figure 7.14 Fist Level: Business intelligence repository

7.7.2.1 Corporate Information Factory

As corporations embark on a Repository of Information, they are unearthing integrity and accuracy problems associated with the operational data. Another major challenge that has emerged is the integration of disparate operational data. This issue pertains to inconsistent data models, disparate data structures and dubious data which are key factors to an organization. (Business Intelligence Repository).

Organizations have foundation data regardless of whether they design and implement reference files/databases, and master databases for foundation data. The foundation data store is a mechanism to formalize this practice into an architectural construct like the operational data store and data warehouse.

The application or system that updates the foundation data store is primarily a data maintenance application. Therefore, it would not possess the complex processing logic of a transactional application such as an order processing system, a production scheduling system, or a trading system.

The framework shown in Figure 7.15 is a useful reference for administrators creating a new transactional application or reengineering a legacy application. Architectural separation of foundation and transactional applications along with their data stores is recommended. Work can be done mainly in the Foundation Section because the analysis of competitors will be generated once a month. Since this could affect some transactional applications, this kind of analysis will be discussed later.

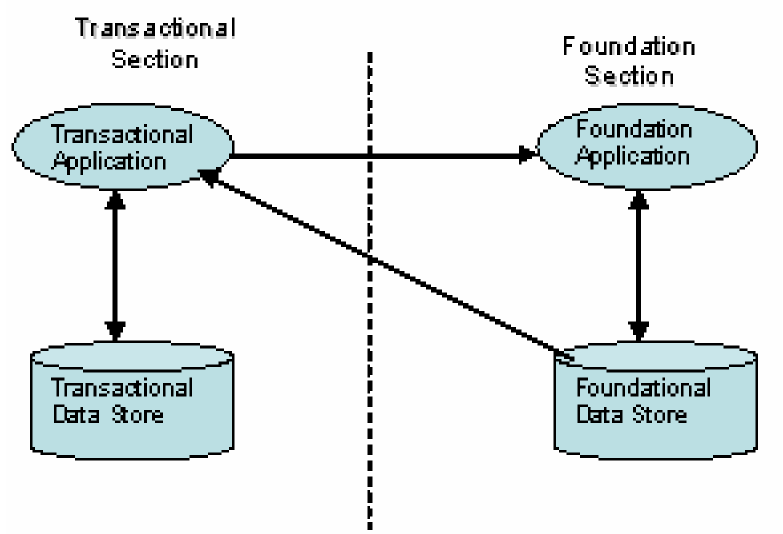


Figure 7.15 Framework for transactional and foundation applications

The foundation application should be a centrally developed application deployed in a distributed fashion. The purpose of the interaction between foundation and transactional applications (See Figure 7.15) is not only to give the transaction application user access to the foundation application, but also to make it appear seamless to the user.

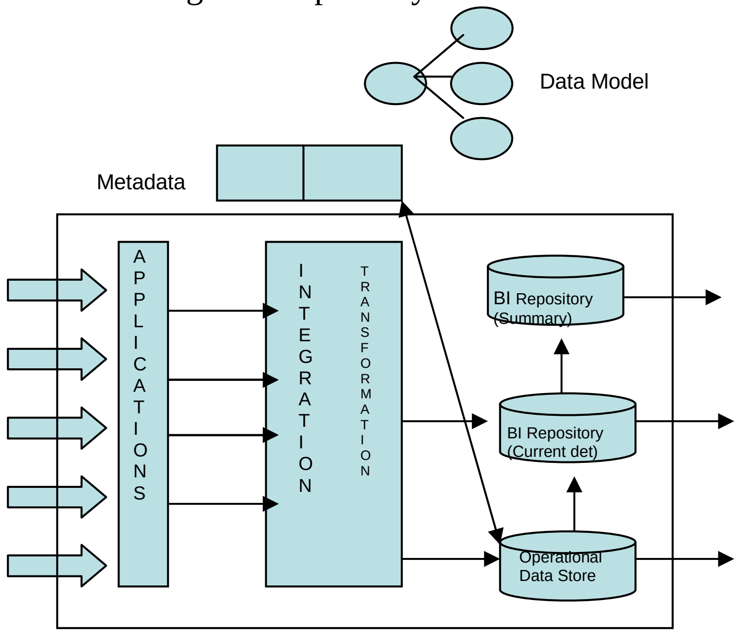
To start working at the first level of the Business Intelligent Repository, it is proposed to work with a common architecture called the corporate information factory. It includes the operational data store (ODS), the data repositories, the data warehouse and the legacy applications (See Figure 7.16). The data flow within the corporate information factory is described as follows:

1.- Raw, detailed data are put into the corporate information factory, by means of data capturing , entry, and transaction interaction with the legacy applications.

2.- The raw, detailed data are integrated and transformed, and then passed into the operational data store or the current detail level of the data warehouse applications.

3.- As the refined data leave the operational data store, these data go into the current level of the Business Intelligence Repository.

4.- Once the refined data are summarized, the data pass from the current detail level of the repository into the summarized level of data in the Business Intelligence Repository.



Source: Inmon, Imhoff, and Battas, Building the Operational Data Store (New York: Jhon Wiley and Sons, Inc., 1995)

Figure 7.16 Corporate information factory

The corporate information factory can be extracted to include the foundation data store as an integral architectural construct within the framework, as shown in Figure 7.16. The Foundational Data Store (FDS) functions as the official source (i.e., system of record) of an organization's foundation data. It maintains and supplies such data to transactional applications, the Operational Data Store (ODS),and the data can also be used by an organization. Following, it will be discussed how the FDS relates to the other components of the corporate information factory.

Transaction Applications. Ideally, the transactional applications should not have their own versions of the foundation data, but should access the centrally maintained data store. In other possible configurations, the transactional applications could make changes to local copy of the central store, and the change would be applied to the central store after authentication.

Integration and Transformation Layer. The implementation of the foundation data store makes the application component more integrated, which leads to a relatively simple and straightforward integration and transformation layer.

Operational Data Store. The ODS application usually uses the current version of the foundation data store. Therefore, ODS applications should be able to directly access the central foundation data store. An alternative way is to replicate a subset of the central foundation data store into the ODS environment.

Business Intelligence Repository. The major entities of the foundation data store become dimensions in a data warehouse. The data warehouse contains the historical snapshots of the foundation data store. The detail contained in the repository should refer to the appropriate snapshot of the foundation data.

Data Model. Design and implementation of the foundation data store should be based on a solid logical data model that is an integral part of an enterprise data model. In the case of foundation model data, translation from the logical model to the physical design is relatively straightforward, unlike the design of the repository or transactional databases that may require a lot of denormalization or summarization.

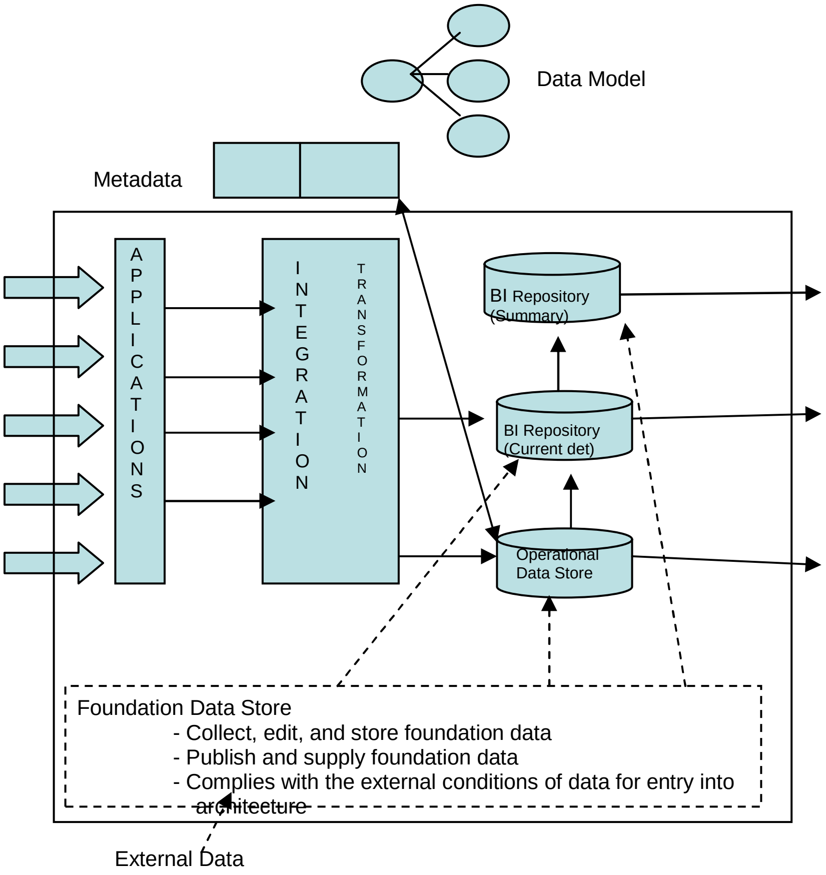

Data Model
Metadata
A P P L I C A T I O N S
I N T E G R A T I O N
T R A N S F O R M A T I O N
BI Repository (Summary)
BI Repository (Current det)
Operational Data Store
Foundation Data Store
- Collect, edit, and store foundation data
- Publish and supply foundation data
- Complies with the external conditions of data for entry into architecture
External Data

Figure 7.17 FDS as a component of the corporate information factory

Metadata. The metadata, coming from the data model, is employed to design the foundtion data store. One of the first steps in implementing a foundation data store is documenting the existing foundation data and developing the standards. Metadata is, therefore, a natural by-product of this process. In an organization that has an efficient FDS implementation, the data dictionary of the foundation database supplies a significant portion of the metada.

External Data. There is a potential need for external data at all levels of the corporate information factory. For example, the repository may contain external information to compare the historical performance of the organization with an external benchmark. Similarly, an order processing system may have electronic data interchange-related external data that are unique to that application or other information extracted from some webpages.

Most of the external data that an organization uses is foundation information. Typical examples include financial product information from market data providers and name and address lists purchased from competitors or vendors. It is essential that the external data are conditioned to adhere to standards before they are stored in the foundation data store.

External data should also be modeled and documented using the enterprise data model and metadata for them to be appropriately classified and integrated with the internal data in the business intelligence repository or other data stores.

7.8 Business Intelligence Integration

Once the Corporate Information Factory (CIF) has been defined, the process of crossing information from the Business Intelligence Repository with other information can start. From crossing information, it is expected that one or more of the three following activities happen: a) new kinds of analysis, b) creation of new information from the BIR and/or c) generation of the Information Systems inside the Corporation.

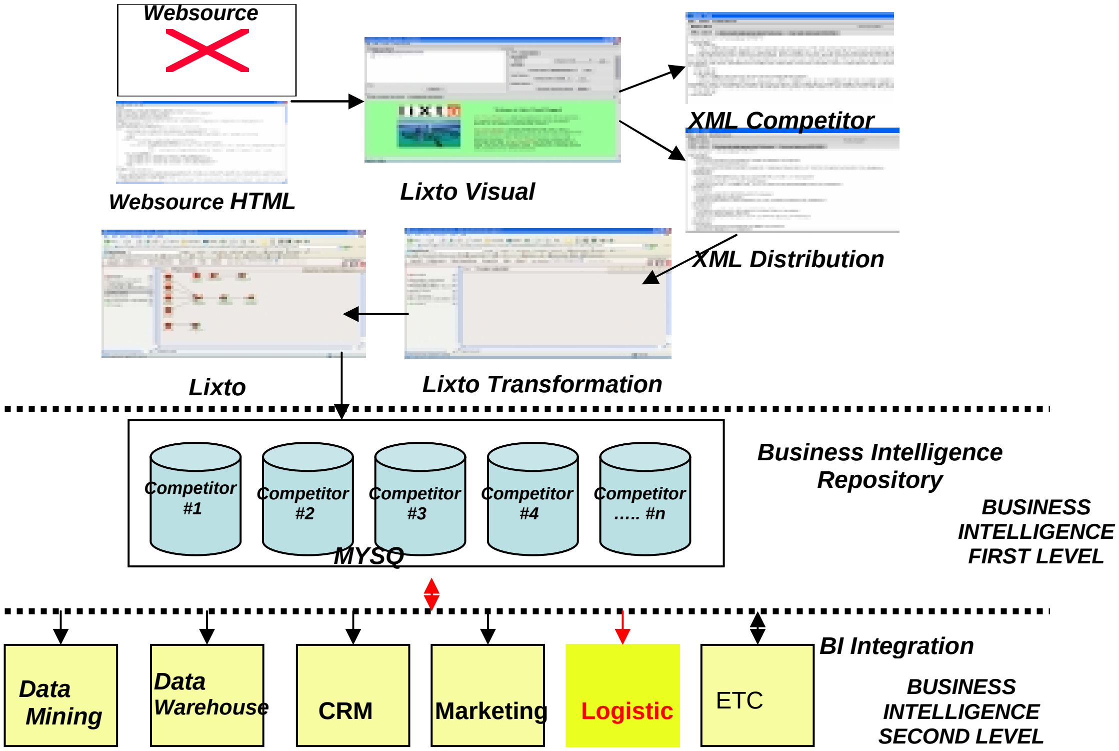


Figure 7.18 Second Level: Business intelligence repository

As it can be seen in Figure 7.18, work can start at the Second Level of Business Intelligence. At this level it is possible to recycle the information that was previously extracted from the webpages (semi-structured data) This information is now in a Business Intelligence Repository (structured data). This can be pulled through querying and then be transfomed into the format needed by the particular application the user is willing to handle.

At the Second Level of Business Intelligence, many of the software applications inside the corporation can be used, such as datamining, datawarehouse, marketing, financial, logistic, CRM, etc. For the last phase of this work, it was given consideration working with a logistic application as stated before.

For the logistic application, it was chosen to work with one of the tools that the corporation uses to analyze data related to Geographical Information Systems (GIS). This tool is MapInfo [14] (see figure 7.19).

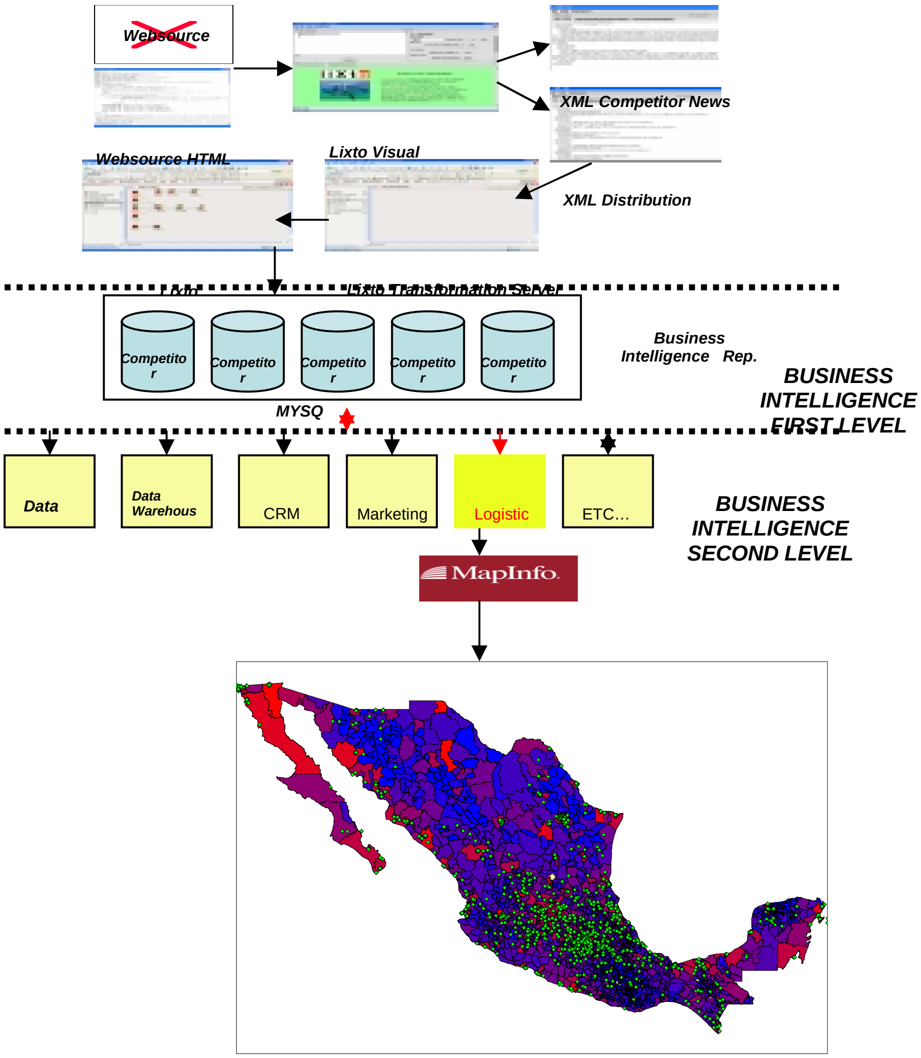


Figure 7.19 Business intelligence integration with MapInfo

To integrate the information of the Business Intelligence Repository to MapInfo, it is needed to do some queries to the MySQL databases and to get all the databases related to the competitors warehouses. As seen in Figure 7.20, the most important pieces of information for mapping the information in MapInfo are: Country, City, Street, Number and Zip-Code. Also, the database provides other bits of relevant information such as Company and Phone-number. However, this information is used as a reference for other kinds of analysis, not for the mapping.

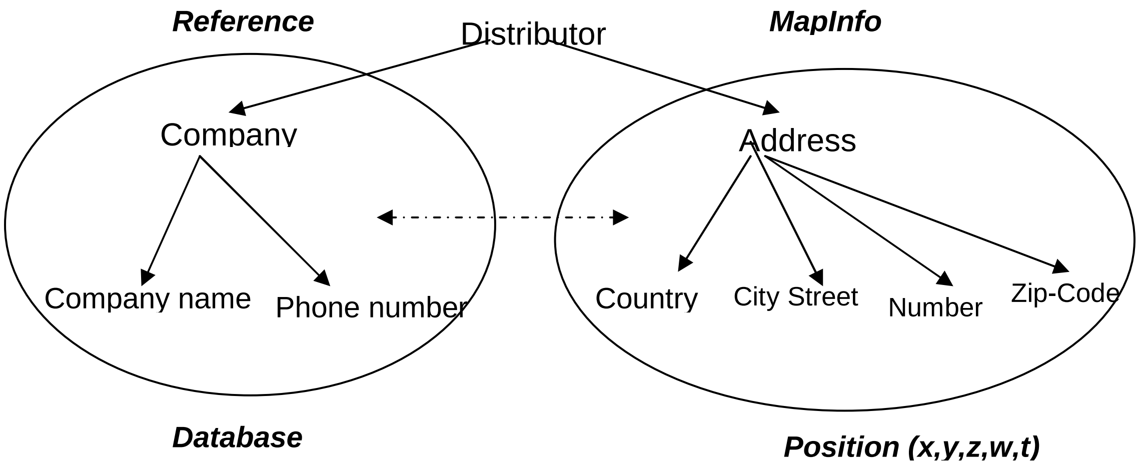


Figure 7.20 The five keys for mapping in MapInfo

Once this information has been obtained through querying the Business Intelligence Repository, it is necessary to transform it to another format. MapInfo accesses raw data as a Plain Text Format. The transformation can be made by using the new information in Plain text Format as a raw data (set of parameters) to MapInfo (See Figure 7.21).

MapInfo accepts external information as a set of parameters (raw data) in different formats such as Excel document files or as PTF . In this case, the information obtained from the databases is directly transformed to a plain text after the execution of a Python transformation program as seen in Figure 7.21.

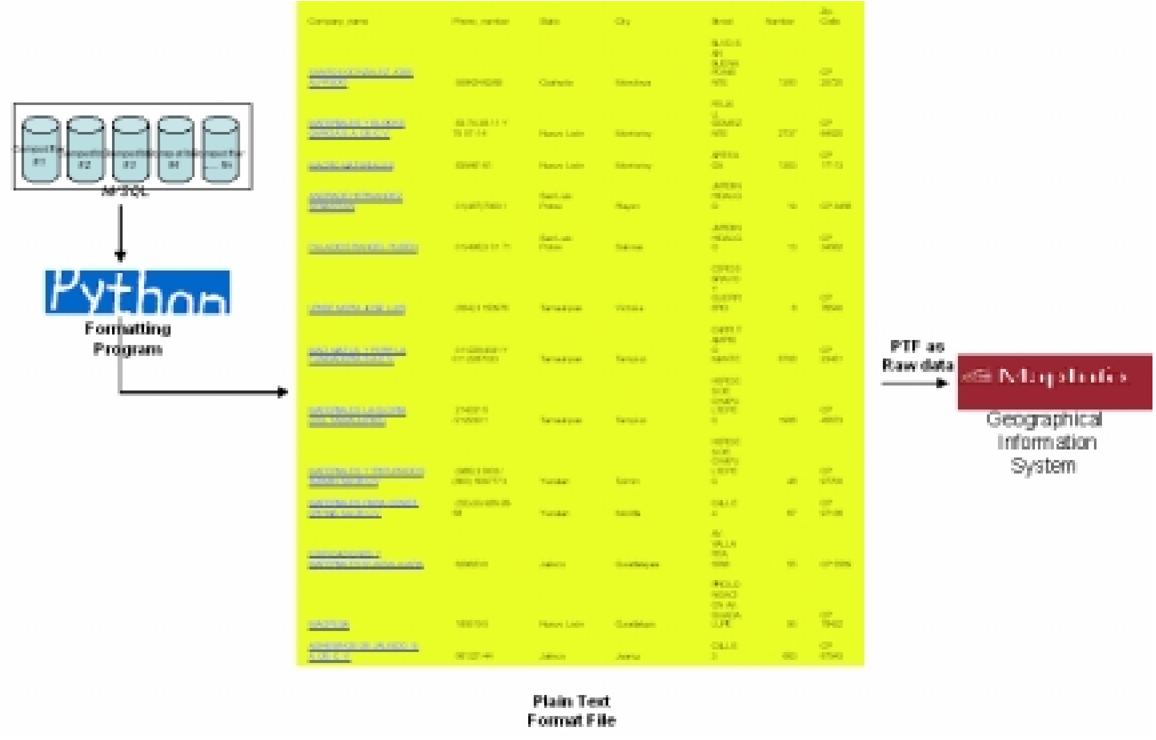
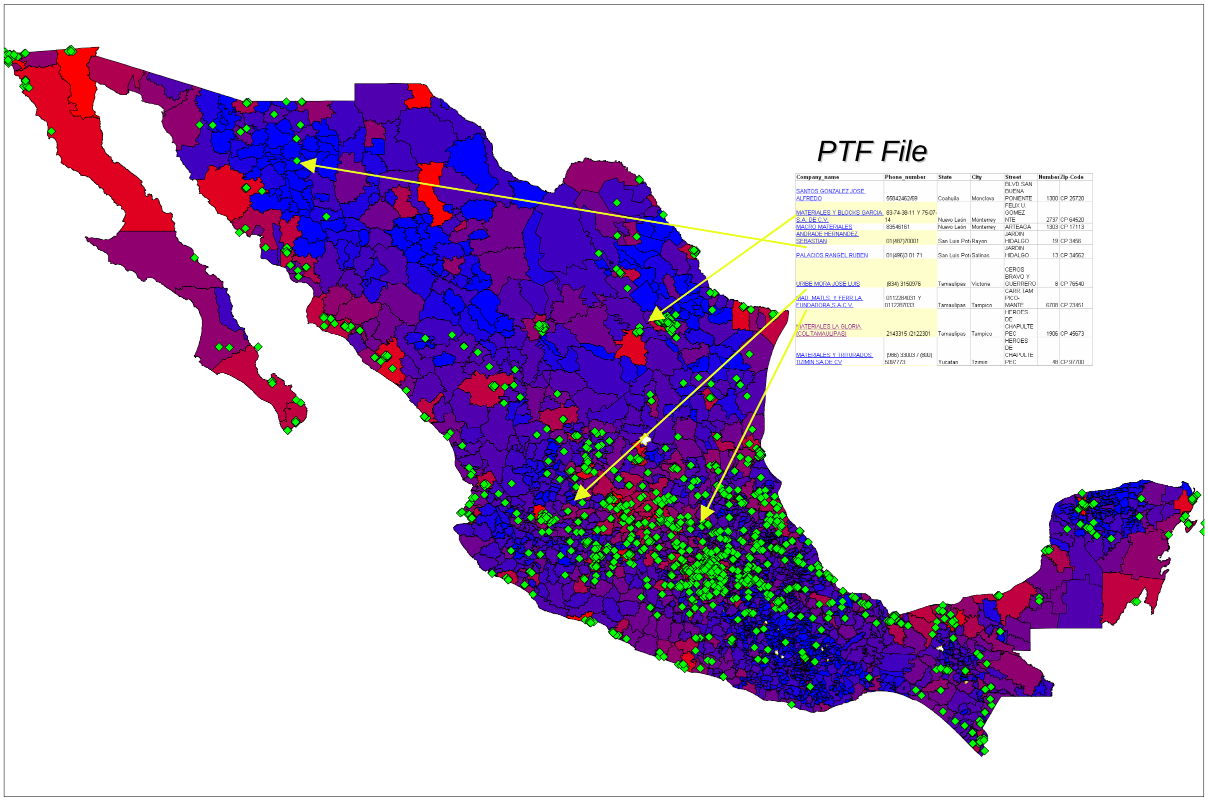


Figure 7.21 Plain text as a raw data to MapInfo

7.9 Generation of Business Intelligence Maps using GIS

After the information has been transformed into a PTF, it becomes raw data to be loaded in the Geographical Information Systems (MapInfo) to start the process of generating maps. MapInfo mainly uses the five key variables mentioned before (country, city, street, number and zip-code). By using this quintet of parameters, it turns somewhat easy to spot a site in a map for each of the competitor's warehouses. It is worthwhile recalling that the information had been previously extracted form the competitors' web pages using the Lixto wrapper technology.

Finally, a visual pattern map, showing the location of each warehouse can be seen. As a Business Intelligence process, it has been proven that recycled information extracted from web pages, then transformed and saved as visual information in a database, can produce a simple map or a set of maps (See Figure 7.22).



Company_name	Phone_number	State	City	Street	Number	Zip-Code
SANTOS GONZALEZ JOSE ALFREDO	55842462/69	Coahuila	Monclova	BLVD.SAN BUENA PONIENTE	1300	CP 25720
MATERIALES Y BLOCKS GARCIA S.A. DE C.V.	83-74-38-11 Y 75-07-14	Nuevo León	Monterrey	FELIX U. GOMEZ NTE	2737	CP 64520
MACRO MATERIALES	83546161	Nuevo León	Monterrey	ARTEAGA	1303	CP 17113
ANDRADE HERNANDEZ SEBASTIAN	01(487)70001	San Luis Pot	Rayon	JARDIN HIDALGO	19	CP 3456
PALACIOS RANGEL RUBEN	01(496)3 01 71	San Luis Pot	Salinas	JARDIN HIDALGO	13	CP 34562
URIBE MORA JOSE LUIS	(834) 3150976	Tamaulipas	Victoria	CEROS BRAVO Y GUERRERO	8	CP 76540
MAD.,MATLS. Y FERR.LA FUNDADORA,S.A.C.V.	0112284031 Y 0112287033	Tamaulipas	Tampico	CARR.TAM PICO-MANTE	6708	CP 23451
MATERIALES LA GLORIA (COL.TAMAULIPAS)	2143315 /2122301	Tamaulipas	Tampico	HEROES DE CHAPULTEPEC	1906	CP 45673
MATERIALES Y TRITURADOS TIZIMIN SA DE CV	(986) 33003 / (800) 5097773	Yucatan	Tzimin	HEROES DE CHAPULTEPEC	48	CP 97700

Figure 7.22 Visual information as a part of the BIP (competitors distribution centers)

7.10 Visual Business Intelligence Analysis

Part of the work of this project is to prove not only the feasibility of recycling and transforming information from the competitors' web sources into the corporation's information systems, but also that the cost of this information is relatively inexpensive by applying the Lixto Wrapper Technology compared with the common practice of Business Intelligence.

As mentioned in Chapter 2, there are some economic factors to be considered when using the Visual Business Intelligence Process with Lixto. This technological application yields the corporation´s wishes of generating economic value and competitive advantages.

With this kind of analysis, a corporation can design strategies if it knows where its competitors are geographically located, which kind of products and services they are offering, in which markets they are competing and what their value chains are. With this new concept of Visual Business Intelligences Approach, the managers of a corporation can develop strategies by analyzing the charts, images, or maps, that are generated with the concept of Visual Business Intelligence (for more detailes See Table 2.1)

Moreover, with this kind of technology (Visual Business Intelligence through Lixto), the corporation can do a comparative analysis using multiple competitors' databases from the Business Intelligence Repository. The corporation can compare where all the facilities of its competitors are located and then define some strategic information directions. Also, the managers can develop strategies, contingency plans and implement policy/strategy programs to generate competitive advantages over the competitors (See Figure 2.1).

This kind of visual information modeling is a good way of analysis for managers who otherwise would receive a lot of reports with large lists of information related to competitors, but , in a way, senseless and time consuming if tried to be read to produce important results.

The comparative analysis provides very rich strategic information. In a short time, it provides managers and the corporation with a picture of the competitors' strengths in a specific country, state, city or sector. This is one of the advantages of integrating information with a Geographical Information System (GIS). Figure 7.23 shows a comparative analysis of the locations of warehouses of the two main competitors of the corporation.

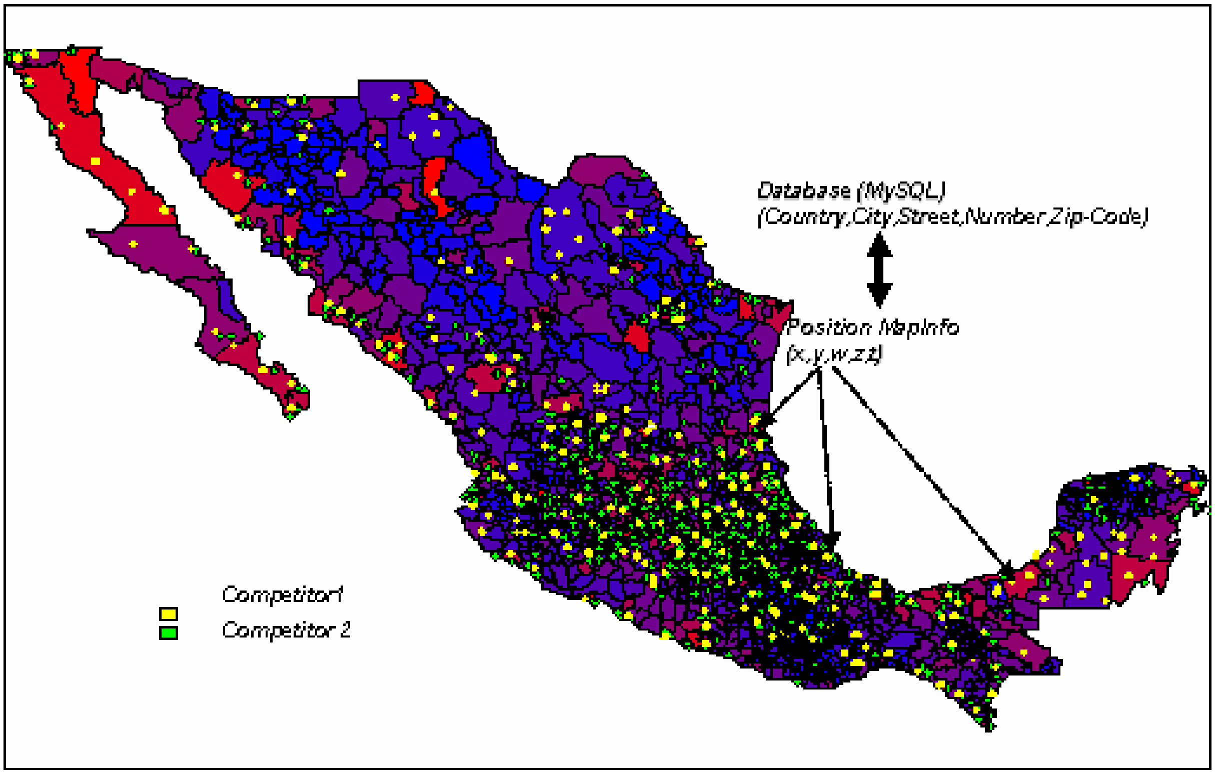


Figure 7.23 Visual comparative analysis (main competitor warehouses)

7.11 Integrating the Corporate Information Factory

As mentioned in Section 7.7.2.1, after the first approach of the Business Intelligence Analysis, it is possible to integrate new sources of information or databases as part of the Corporate Information Factory (See Figure 7.24).

To complement this analysis, other geographical databases related to the socio-economic attributes of the different society sectors of each of the cities in Mexico were integrated. It is the National Institute of Statistics, Geography and Informatics (INEGI) the office in charge of generating such databases. [51]

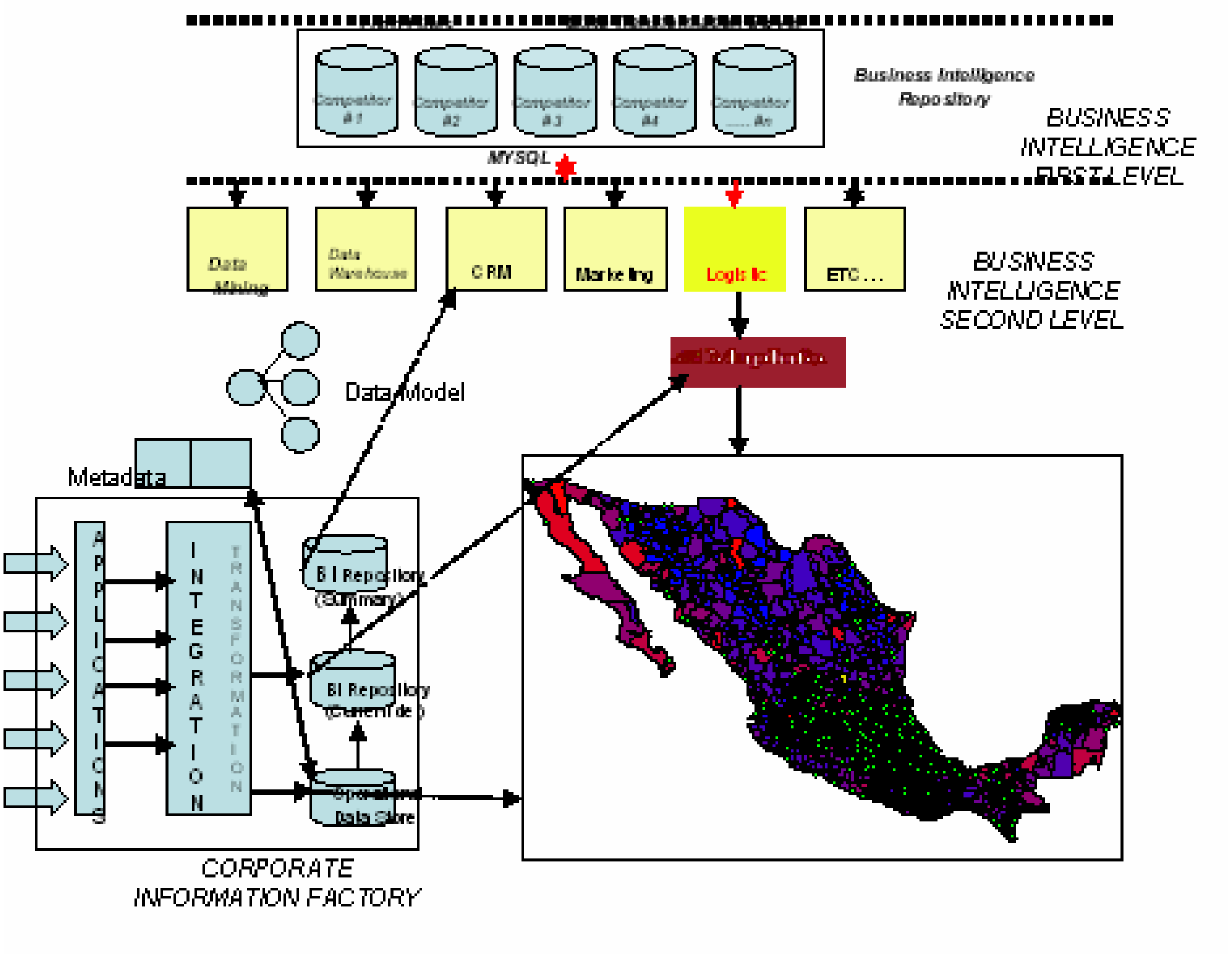


Figure 7.24 Integration with the corporate information factory

The INEGI databases complement the Visual Business Intelligence Analysis, adding new information sources to the MapInfo visual model. Through the location of the warehouses, the information related to population statistics in each of this cities can be integratee and then interrelated with the geographical area of each of the warehouses.

The integration of the Corporate Information Factory (CIF) allows deeper analysis not only of the logistics department of the corporation, but also of areas such as marketing and finances Thus, the Visual Business Intelligence Analysis will be rich in information in a visual way as seen in Figure 7.25.

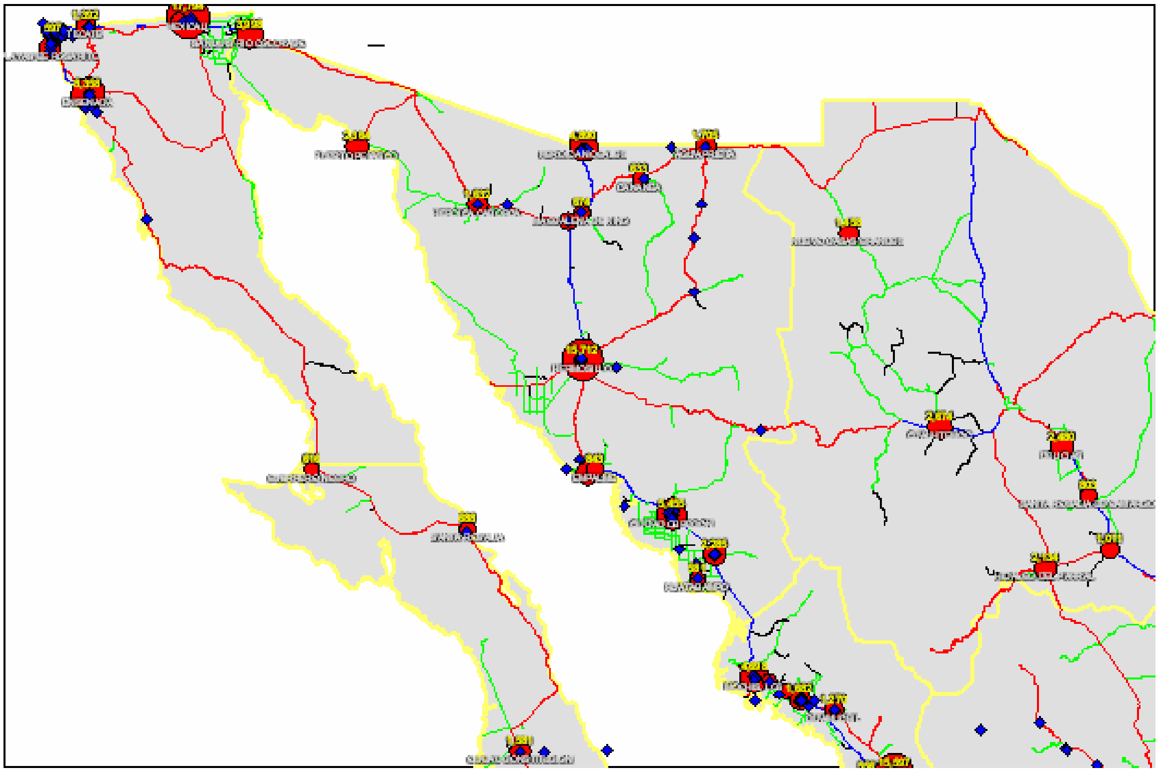


Figure 7.25 New visual business intelligence analysis integrating with the CIF

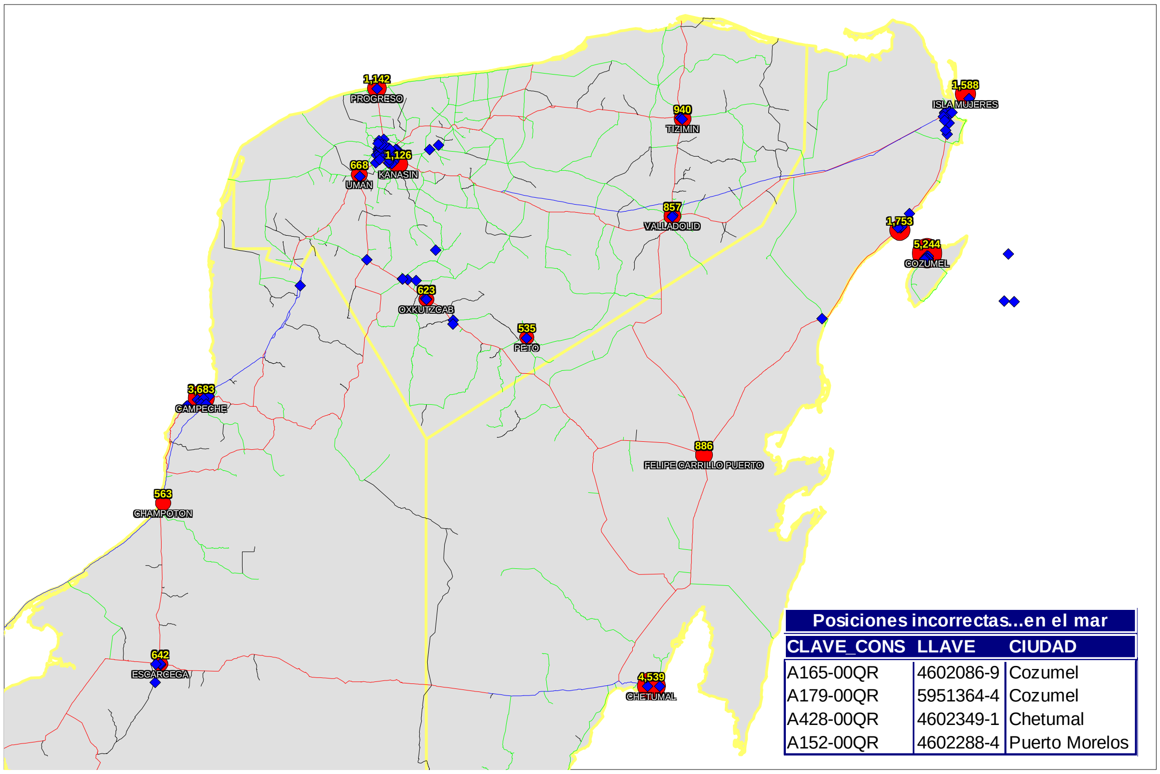
By adding the Corporate Information Factory, the corporation can do a more robust analysis integrating other data sources like INEGI with a unified data model, for example.

1.- The corporation can find the distance between warehourses in kilometers, including the roads connecting them.

2.- The corporation could know the population of a particular area in a specific city or state.

3.- The corporation can have information about the socio-economic profile of the population living in a specific sector, city or town.

The Visual Business Intelligence Approach can Recycle Information from Websources to other information systems in order to generate strategies of other sources of information or simply update existing sources.



Posiciones incorrectas...en el mar

CLAVE_CONS	LLAVE	CIUDAD
A165-00QR	4602086-9	Cozumel
A179-00QR	5951364-4	Cozumel
A428-00QR	4602349-1	Chetumal
A152-00QR	4602288-4	Puerto Morelos

Figure 7.26 Inconsistences of the information extracted

During this research work, we have found some inconsistencies in some data in the Visual Business Intelligence Map. As seen in Figure 7.26, there are some points in the sea. These inconsistencies are not due to the data extraction of the mapping. The problem arises from the information displayed on the website where the information was obtained. This kind of situations is not detected during the data extraction process, so there is the risk of extracting some wrong information like this.

The error factor in this kind of analysis is minimum, so it can be concluded that the technique for the Visual Business Analysis Approach is a good technique with a high degree of success for the strategy analysis and design.

Finally, the full schema used for the development of this thesis can be seen. After the Geographical Information Analysis with CIF, the corporation will generate or update new sources of information inside the same company as shown in Figure 7.27. With this schema, the concept of Web Recycling of Information for the Strategic Planning in the corporation that wants to use the Visual Business Analysis Approach. is shown.

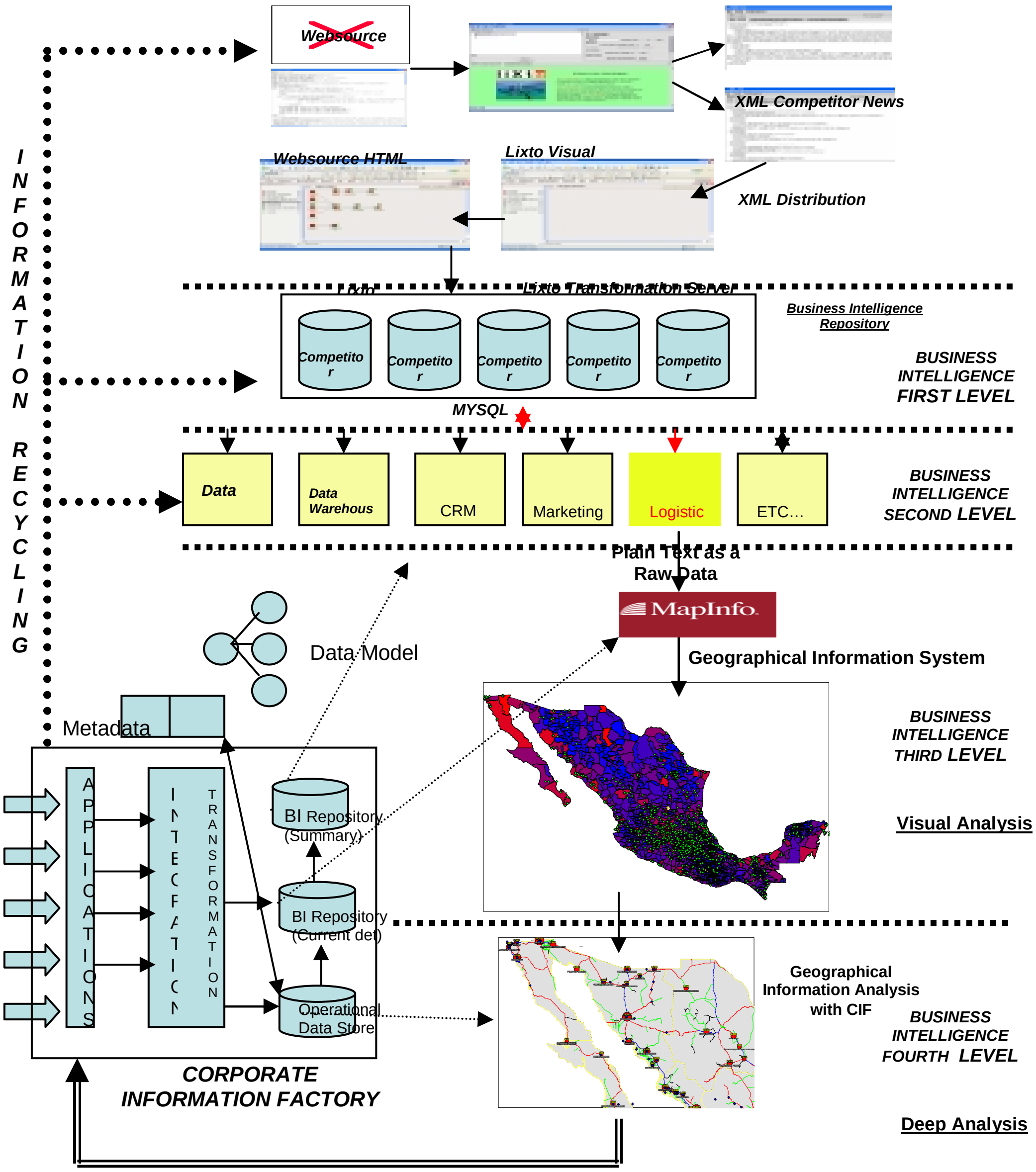


Figure 7.27 Full model for the visual business intelligence analysis and inf. modeling

Chapter 8

Epilogue

8.1 Related Works

Prior to developing this research work, different research works related to the subjects of Visual Information Extraction, Visual Business Intelligence, Data Extraction, Information Modelling, Web Ecology and Web Recycling were analized.

Literature related to some of these topics was found, however I did not find literature of research works that integrated those topics as is proposed in this thesis was not found. Thus, its title is Business Intelligence in the Logistics Domain using Visual Information Extraction. For the development of this thesis it was necessary to emphasize the use of technologies for data extraction, recyclying, integration of information and information modeling.

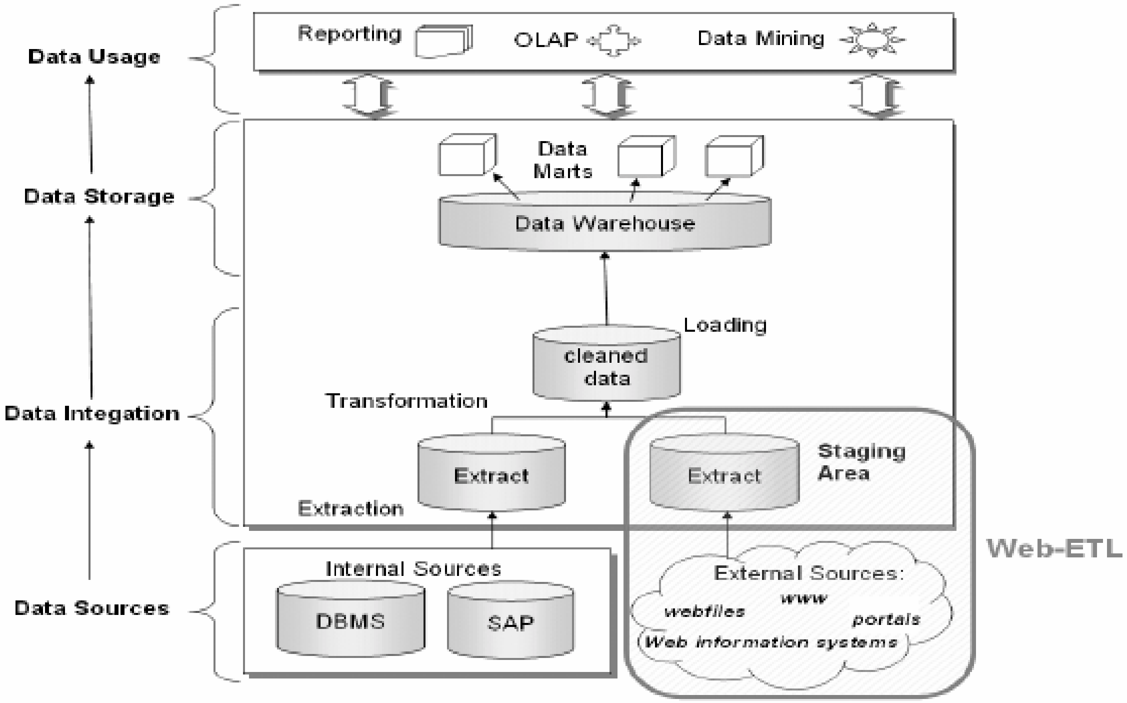
Even though it is not the purpose of this work, it is worthwhile mentioning that a lot of research has been done around web data extraction.. In contrast, in the case of Web Data Extraction for Business Intelligences, Baumgartner and Frölich's work has been appropriate for this research. [12]. The authors analyzed the process of business inteligences through data extraction from some web sources, and the integration and normalization of the information in an automatic way.

Baumgartner and Frölich [12] also made use of the Lixto technologies to extract, transform and deliver several semi-structured sources like web pages and various customer systems. In that work they integrated the data extracted into a SAP

Business Information Warehouse (SAP BW). The integration of data was done by loading XML archives through a SOAP interface (See Figure 8.1).
The main focus of this work was to improve the Business Intelligence Processes not only by extracting information and storing it in data bases or repositories for future analysis through queries or reports –which had been done before- but also to actually analyze it by means of generating the Visual Information Extraction. This method provides the alternative to show the information in a visual way through information modeling and to recycle the extracted information.

This concept of Visual Information Extraction enhances the significant value of extracting information by means of the Visual Lixto Wrapper and the Lixto Information Server. [62].

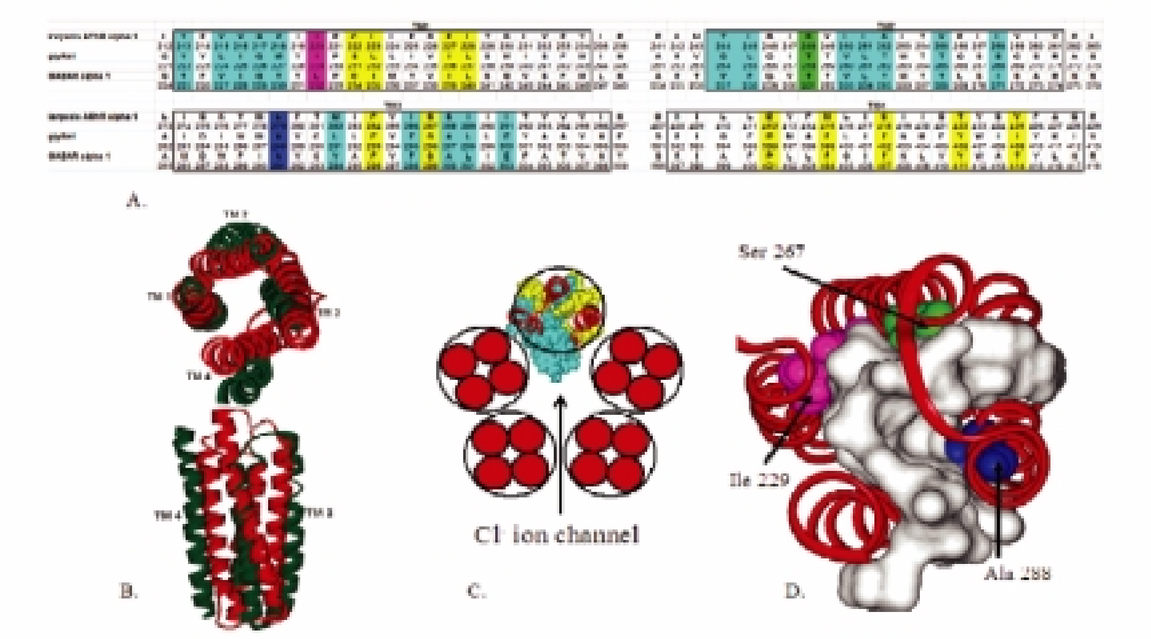
About information modeling, some interesting developments in areas such as chemistry and biotechnology were found . The main contribution is through the operation of great volumes of information contained in large databases or repositories of information. With this information, it is possible to generate computational models, patterns and tendencies such as modeling DNA chains [60].



Source: Web Data Extraction for Business Intelligence: the Lixto Approach. Proceedings of the BTW2005, Karlsruhe, Germany, 2005.
Figure 8.1. The data extraction and the BI integration process

Nevertheless, it has not yet been exploited the possibility of extracting data of DNA sequence data bases, geographically distributed around the world so as to consolidate more robust repositorios of bio-information.

Based on the information modeling in the biochemical and biological areas, it was looked for the extrapolation of the decision-making processes sustained on the information modeling. Among others, the research work "Homology Modeling of to the Human Glycine Receiving Alpha 1 Reveals a Plausible Anesthetic Blinding Site" [14] reports that through the information modeling and experimental data it is prosible to produce exact models. In this case several techniques of bioinformatics were used to predict and to generate the topology model of the transmebrane of a glycine alpha. (See figure 8.2).



Source: Homology Modeling of a Human Glycine Alpha 1 Receptor reveals a Plausible Anesthetic Binding Site. Edwar J. Bertaccini, American Chemical Society Published on Wevb

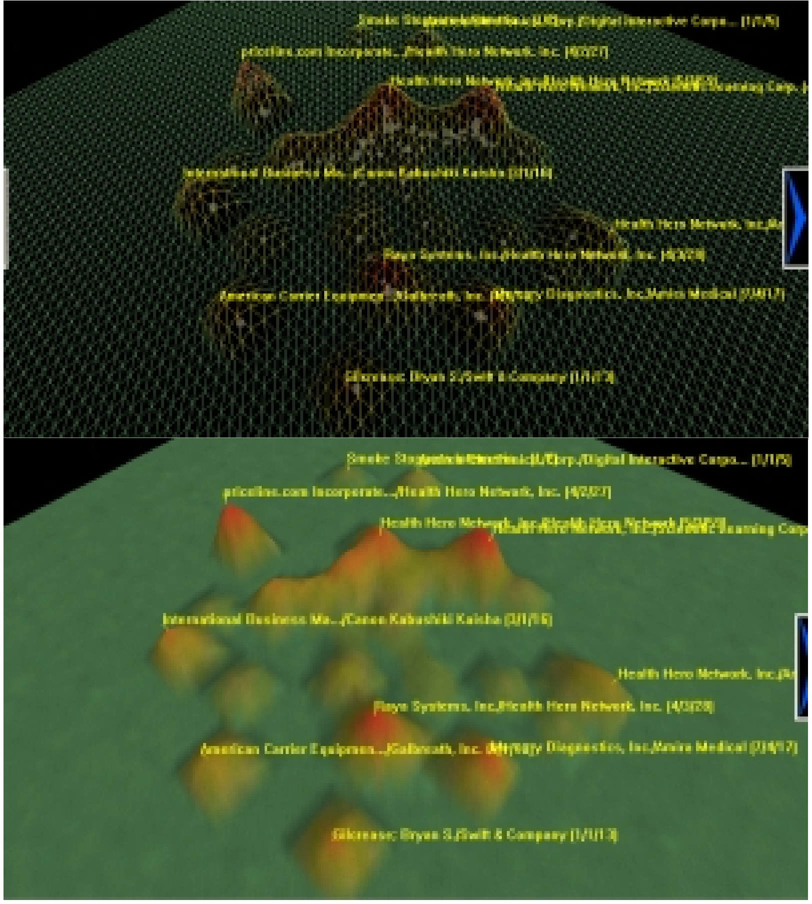
Figure 8.2. Visual modeling of a multiple sequence aligment of torpedo nAChR with human glyRal and GABARa1.

Another area that was analyzed and that has contributed to this research work was in the subject of Web Recycling and Web Ecology.

Awareness on web ecology was created when pointing out the importance of recycling existing information on Web pages. Web Ecology considers the application of different technologies such as HTML, XML, XLST, CSS, XSL-FO, Python, Java, Perl, etc. Under this focus, Web Recycling and Web Ecology it is demonstrated how information from different sources can be recycled to generate new and more robust sources of information as shown in the model of the Corporate Information Factory.

Moreover, another opportunity area for the Visual Data Extraction & Information Modeling Approach is in the process of Patent Analysis or Morphological Patent Analysis. At present, there are some patent databases around the world; unfortunately, this information is available in html and in some cases in pdf. Consequently, with this kind of format it is not easy to handle the massive access of information for its future analysis.

Other sources of information are available in some comercial databases. However, in this case the corporation needs to pay an expensive membership fee per year to access the full data base. Fortunately, with the technology of data extraction, it is possible to extract the information related with the patents and then store it in a repository (database). After the information is available in the repository, it is possible to do the modeling of this information using different kinds of software to do the Visual Business Intelligences so as to generate some landscape patent mappings as can be seen in Figure 8.3.



priceline.com Incorporate.../Health Hero Network, Inc. (4/2/27)
International Business Ma.../Canon Kabushiki Kaisha (2/1/16)
Health Hero Network, Inc./
Roya Systems, Inc./Health Hero Network, Inc. (4/3/28)
American Carrier Equipmen.../Galbreath, Inc.
Diagnostics, Inc./Amira Medical (2/4/17)
Gilcrease; Bryan S./Swift & Company (1/1/13)

Figure 8.3. Patent landscape analysis

As part of this analysis, it can be concluded that in an ever-evolving world, corporations and organizations demand not only timely but safe and reliable information coming from external sources. Consequently, the corporations are interested in the integration of information from some websources into the Business Intelligence and Warehouse systems that involve some critical factors for these trends, such as:

- The corporation makes available the Business Intelligence systems and data on the desks of large numbers of users; thus, the information is democratized and the user gets more detailed information on different devices and forms.

- The corporations require making decisions more accurately in shorter time periods and the response time between making a decision and the feedback is becoming shorter. Also the ability to make intelligence business decisions in shorter time is fundamental for competitiveness.

- The data must be customized on a massive scale and the data warehouses need to be flexible to supply different types of information to different kinds of users.

- The variety of data is tremendous. The corporations store data from different sources as Enterprise Resources Planning, Customer Relationship Management and Web Systems, and in different formats such as text, images, voice, video, unstructured data, and more.

- The last of these critical aspects is the increased need to improve the data access availability and the large number and types of users. The corporations around the world are looking for alternatives to decrease the adventure associated with managing growing and contradictory forms of data.

In addition to these critical factors, there are a number of industry-specific drives (See Figure 8.4) that could be of interest for future application in Business Intelligences Using Visual Information Extraction as it has been mentioned in this thesis.

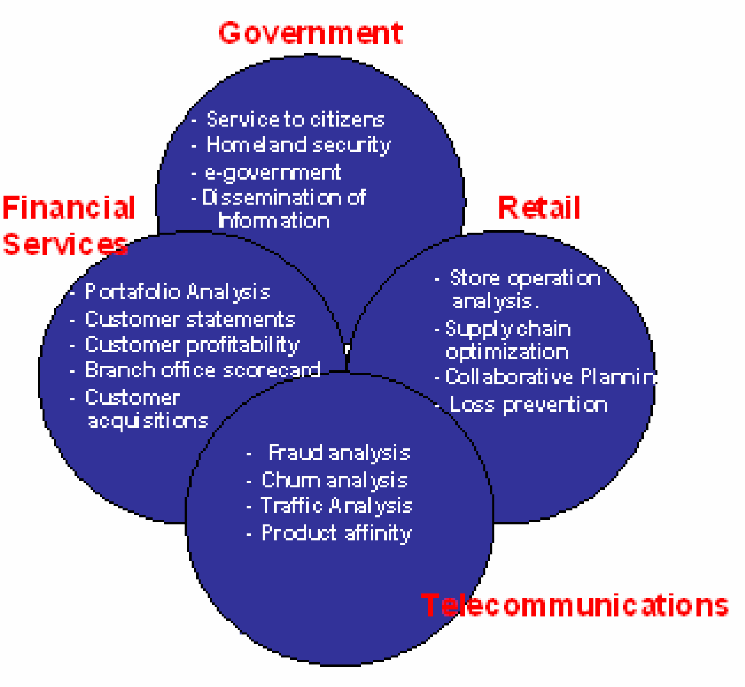


Figure 8.4 Industry needs driving business intelligence and datawarehouse

8.2 Conclusions and Future Work

After the development of this thesis, it can be concluded that the concept of Visual Information Extraction contributes in a satisfactory way to the process of Business Intelligence and Datawarehousing through the information extraction, recycling and integration of web sources.

The main contributions to be reported are as follows:

1. Data extraction is a good technique to support the processes of Business Intelligence and Datawarehousing.

2. Owning a repository of information coming from web page extracted data is a good strategy to analyze background information which will affect the organization’s future decisions.

3. The integration of the information to other information systems, such as CRM, Data Mining, etc, can generate strategic Visual Analysis to improve the decision making process.

4. The Visual Analysis through information mapping gave a new approach to find hidden information in a large amount of information stored in repositories or large databases.

5. It is possible to undertake more in depth analysis by integrating additional information sources from external data bases after the stage of Visual Analysis.

6. The information modeling allows to have a snapshot of the status of the extracted data and to integrate it to the datawarehouse.

7. By visualizing the information, it is possible to integrate all the elements for making effective decisions in real time.

8. It is possible to recycle and transform the information extracted from different web sources for its future integration in the different processes of Business Intelligence and in the Information Modeling Approach.

9. There are economic benefits generated in the decision making process through the simultaneous generation of visual maps and information modeling.

The future work to set out from this thesis is mainly to explore the applications of the Visual Information Extraction and Information Modeling concepts to other areas from which corporations demand to extract great volumes of information to be analyzed.

It is suggested the extrapolation of the Visual Information Extraction model in the biotechnology industry (See Appendix C) jointly with bioinformatics, since these areas demand the access to large volumes of data to be able to model chains of DNA and other chemical properties There could be access to a lot of biological sequence information in public and private databases as: http://www.agr.kuleuven.ac.be/vakken/i287/bioinformatica.htm#Primary%20DB. (See Appendix D).

It is also considered as a potential opportunity area to apply the concepts of Visual Information Extraction and Information Modelling through the application of Lixto.

Appendix A

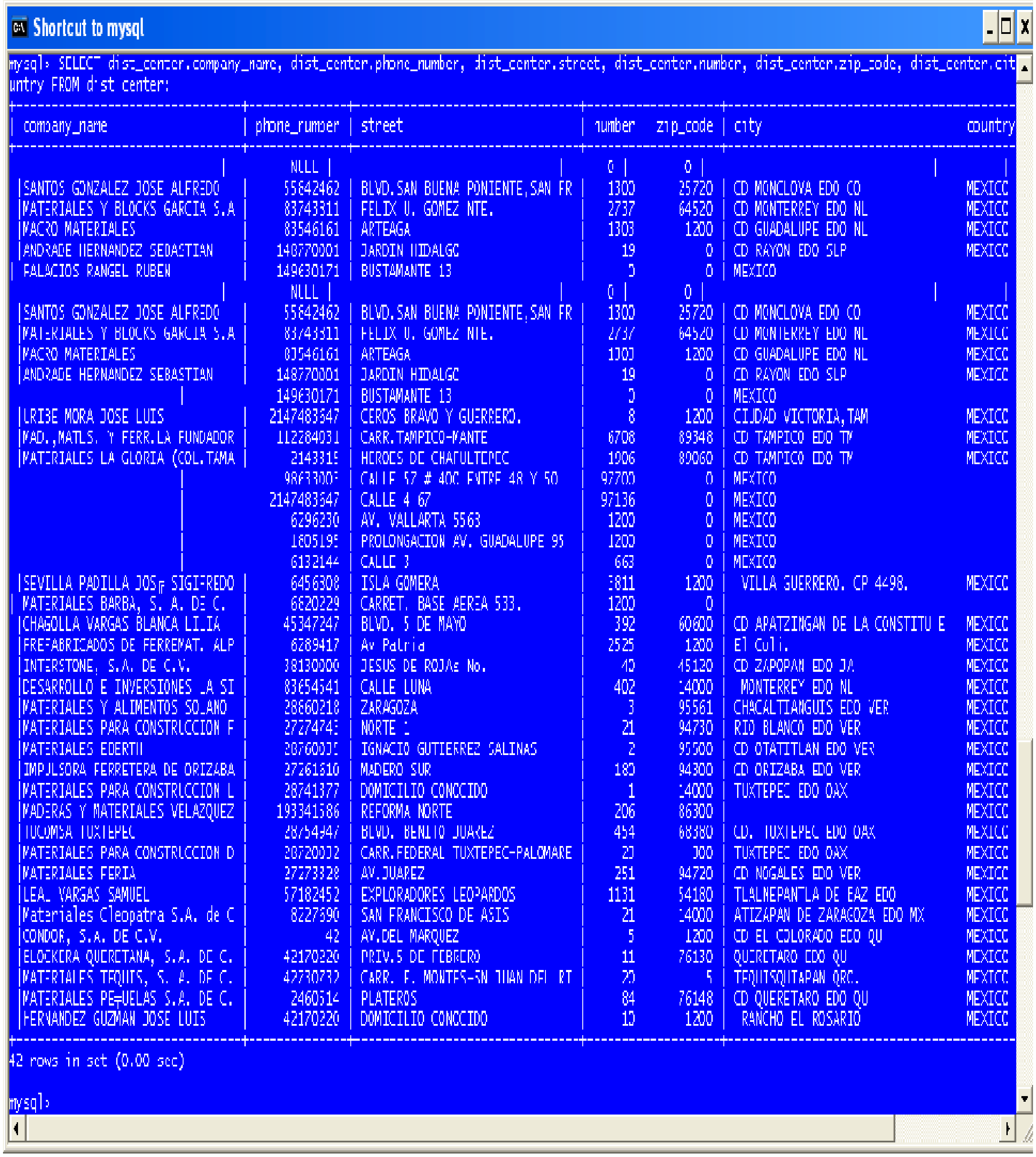
```
<?xml version="1.0" encoding="UTF-8"?>
<document>
    <Distributor>
        <Company>SANTOS GONZALEZ JOSE ALFREDO</Company>
        <Phone>55842462/69</Phone>
        <Address>BLVD.SAN BUENA PONIENTE 1300SAN FRANCISCO CP 25720 CD MONCLOVA EDO CO</Address>
    </Distributor>
    <Distributor>
        <Company>MATERIALES Y BLOCKS GARCIA S.A. DE C.V.</Company>
        <Phone>83-74-38-11 Y 75-07-14</Phone>
        <Address>FELIX U. GOMEZ NTE. 2737 CP 64520 CD MONTERREY EDO NL</Address>
    </Distributor>
    <Distributor>
        <Company>MACRO MATERIALES</Company>
        <Phone>83546161</Phone>
        <Address>ARTEAGA 1303 PARAISO CP CD GUADALUPE EDO NL</Address>
    </Distributor>
    <Distributor>
        <Company>ANDRADE HERNANDEZ SEBASTIAN</Company>
        <Phone>01(487)70001</Phone>
        <Address>JARDIN HIDALGO 19 CP CD RAYON EDO SLP</Address>
    </Distributor>
    <Distributor>
        <Company>PALACIOS RANGEL RUBEN</Company>
        <Phone>01(496)3 01 71</Phone>
        <Address>BUSTAMANTE 13 CP CD SALINAS EDO SL</Address>
    </Distributor>
    <Distributor>
        <Company>URIBE MORA JOSE LUIS</Company>
        <Phone>(834) 3150976</Phone>
        <Address>8 CEROS BRAVO Y GUERRERO. CIUDAD VICTORIA,TAM</Address>
    </Distributor>
    <Distributor>
        <Company>MAD.,MATLS. Y FERR.LA FUNDADORA,S.A.C.V.</Company>
        <Phone>0112284031 Y 0112287033</Phone>
        <Address>CARR.TAMPICO-MANTE 6708MEXICO CP 89348 CD TAMPICO EDO TM</Address>
    </Distributor>
    <Distributor>
        <Company>MATERIALES LA GLORIA (COL.TAMAULIPAS)</Company>
        <Phone>2143315 /2122301</Phone>
        <Address>HEROES DE CHAPULTEPEC 1906 EXTTAMAULIPAS CP 89060 CD TAMPICO EDO TM</Address>
    </Distributor>
    <Distributor>
        <Company>MATERIALES Y TRITURADOS TIZIMIN SA DE CV</Company>
        <Phone>(986) 33003 / (800) 5097773</Phone>
        <Address>CALLE 57 # 400 ENTRE 48 Y 50 CP 97700 CD TIZIMIN EDO YUC</Address>
    </Distributor>
    <Distributor>
        <Company>MATERIALES PARA CONST. ORRAB SA DE CV.</Company>
        <Phone>(52)-(9)-926-26-58</Phone>
        <Address>CALLE 4 67FELIPE CP 97136 CD MERIDA EDO YU</Address>
    </Distributor>
    <Distributor>
        <Company>EDIFICACIONES Y MATERIALES GUADALAJARA</Company>
        <Phone>6296230</Phone>
        <Address>AV. VALLARTA 5568 CP CD GUADALAJARA HI</Address>
```

```
</Distributor>
<Distributor>
   <Company>MAGPESA</Company>
   <Phone>1805195</Phone>
   <Address>PROLONGACION AV. GUADALUPE 95 MIRAMAR</Address>
</Distributor>
<Distributor>
   <Company>ADHESIVOS DE JALISCO, S. A. DE C. V.</Company>
   <Phone>06132144</Phone>
   <Address>CALLE 3 663 AZONA INDUSTRIAL CP CD GUADALAJARA JUAREZ EDO JA</Address>
</Distributor>
<Distributor>
   <Company>SEVILLA PADILLA JOSÉ SIGIFREDO</Company>
   <Phone>06456808</Phone>
   <Address>ISLA GOMERA 3811. VILLA GUERRERO. CP 4498.</Address>
</Distributor>
<Distributor>
   <Company>MATERIALES BARBA, S. A. DE C. V.</Company>
   <Phone>06820229</Phone>
   <Address>CARRET. BASE AEREA 533. SAN JUAN DE OCOTAN. CP CD</Address>
</Distributor>
<Distributor>
   <Company>CHAGOLLA VARGAS BLANCA LILIA</Company>
   <Phone>453-47247</Phone>
   <Address>BLVD. 5 DE MAYO 392 CP 60600 CD APATZINGAN DE LA CONSTITU EDO MI</Address>
</Distributor>
<Distributor>
   <Company>PREFABRICADOS DE FERREMAT. ALPE SA CV</Company>
   <Phone>06289417</Phone>
   <Address>Av Patria 2525. El Coli. CP 450</Address>
</Distributor>
<Distributor>
   <Company>INTERSTONE, S.A. DE C.V.</Company>
   <Phone>(3) 8130000</Phone>
   <Address>JESUS DE ROJAs No. 40 ALOS PINOS CP 45120 CD ZAPOPAN EDO JA</Address>
</Distributor>
<Distributor>
   <Company>DESARROLLO E INVERSIONES LA SILLA</Company>
   <Phone>83654541/42</Phone>
   <Address>CALLE LUNA 402 E COL. CONTRY MONTERREY EDO NL</Address>
</Distributor>
<Distributor>
   <Company>MATERIALES Y ALIMENTOS SOLANO</Company>
   <Phone>(288) 6-02-18</Phone>
   <Address>ZARAGOZA 3 CP 95561 CD CHACALTIANGUIS EDO VER</Address>
</Distributor>
<Distributor>
   <Company>MATERIALES PARA CONSTRUCCION FERRE GALI</Company>
   <Phone>272-74743 Y 272-76123</Phone>
   <Address>NORTE 1 21-A CENTRO CP 94730 CD RIO BLANCO EDO VER</Address>
</Distributor>
<Distributor>
   <Company>MATERIALES EBERTH</Company>
   <Phone>(287) 6-00-35</Phone>
   <Address>IGNACIO GUTIERREZ SALINAS #2 CENTRO CP 95500 CD OTATITLAN EDO VER</Address>
</Distributor>
<Distributor>
   <Company>IMPULSORA FERRETERA DE ORIZABA S.DE R.L</Company>
   <Phone>(272)6-16-10, 6-16-03, 5-12-57</Phone>
   <Address>MADERO SUR 180 CENTRO CP 94300 CD ORIZABA EDO VER</Address>
</Distributor>
<Distributor>
   <Company>MATERIALES PARA CONSTRUCCION LA CARLOTA</Company>
   <Phone>(287) 4-13-77, 4-15-89</Phone>
   <Address>DOMICILIO CONOCIDO EJIDO LA CARLOTA CD TUXTEPEC EDO OAX</Address>
</Distributor>
<Distributor>
   <Company>MADERAS Y MATERIALES VELAZQUEZ</Company>
   <Phone>01-933 4-15-86 o 4-38-76</Phone>
   <Address>REFORMA NORTE 206 CP 86300</Address>
</Distributor>
<Distributor>
   <Company>TUCOMSA TUXTEPEC</Company>
   <Phone>(2)8754947</Phone>
   <Address>BLVD. BENITO JUAREZ 454 COL. LA PIRAGUA CP 68380 CD. TUXTEPEC EDO OAX</Address>
</Distributor>
<Distributor>
   <Company>MATERIALES PARA CONSTRUCCION DE BETHANIA</Company>
   <Phone>(287) 2-00-32, 2-00-66</Phone>
   <Address>CARR.FEDERAL TUXTEPEC-PALOMARES KM 23+300 CD TUXTEPEC EDO OAX</Address>
```

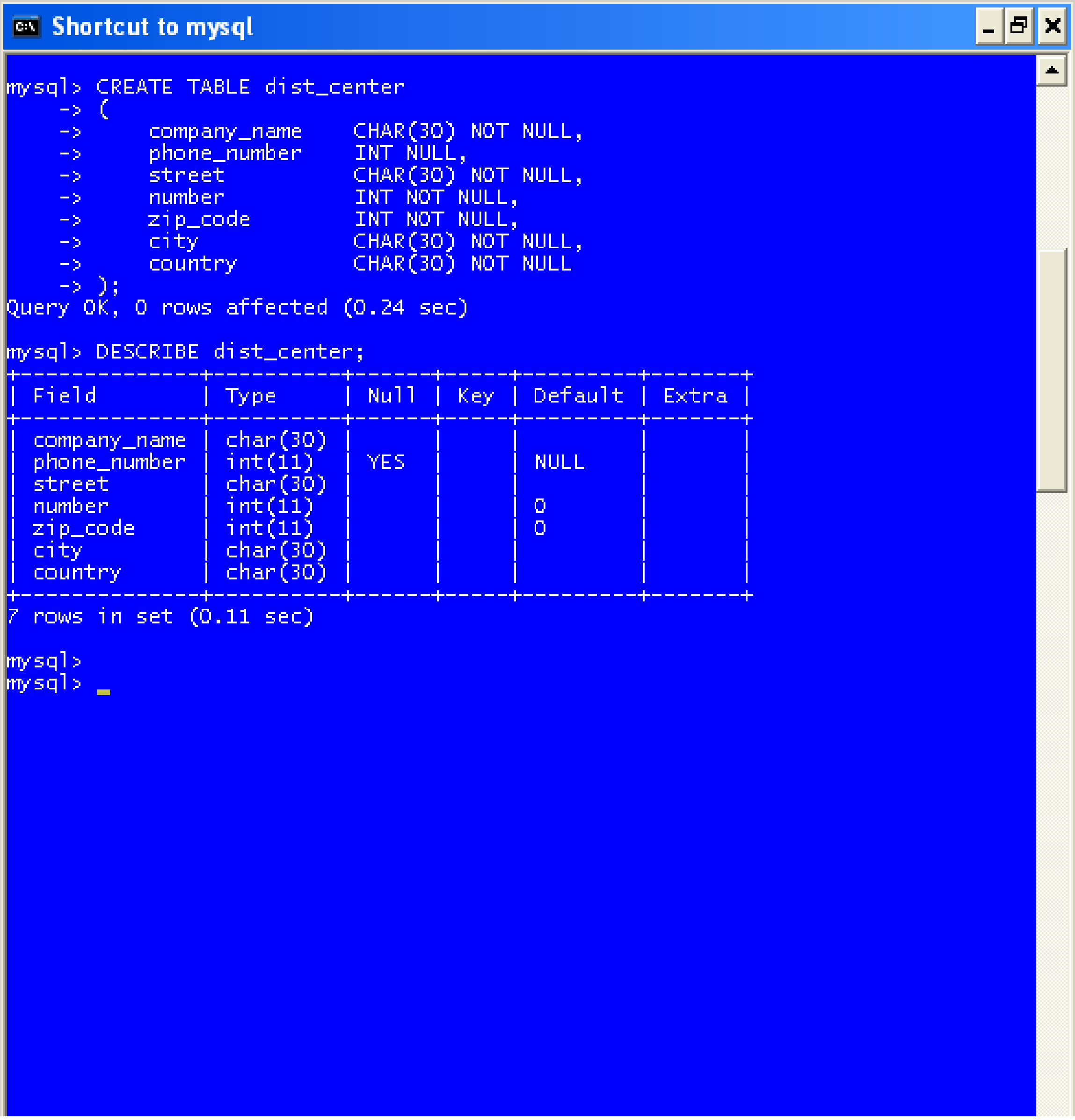
```
  </Distributor>
  <Distributor>
    <Company>MATERIALES FERIA</Company>
    <Phone>27273328</Phone>
    <Address>AV.JUAREZ 251-B CENTRO CP 94720 CD NOGALES EDO VER</Address>
  </Distributor>
  <Distributor>
    <Company>LEAL VARGAS SAMUEL</Company>
    <Phone>57182452 57181453</Phone>
    <Address>EXPLORADORES LEOPARDOS 1131LAZARO CARDENAS , LA PRESA CP 54180 CD TLALNEPANTLA DE BAZ EDO MX</Address>
  </Distributor>
  <Distributor>
    <Company>Materiales Cleopatra S.A. de C. V.</Company>
    <Phone>58227690 FAX 58250455</Phone>
    <Address>SAN FRANCISCO DE ASIS 21LA ERMITA ATIZAPAN CP CD ATIZAPAN DE ZARAGOZA EDO MX</Address>
  </Distributor>
  <Distributor>
    <Company>CONDOR, S.A. DE C.V.</Company>
    <Phone>4 221 51 78 / 79</Phone>
    <Address>AV.DEL MARQUEZ No. 5 PARQUE INDUSTRIA BERNARDO CP 76246 CD EL COLORADO EDO QU</Address>
  </Distributor>
  <Distributor>
    <Company>BLOCKERA QUERETANA, S.A. DE C.V.</Company>
    <Phone>(4) 2170220</Phone>
    <Address>PRIV.5 DE FEBRERO 11SAN PABLO CP 76130 CD QUERETARO EDO QU</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES TEQUIS, S. A. DE C. V.</Company>
    <Phone>(427)3-07-32 Fax: 3-19-78</Phone>
    <Address>CARR. E. MONTES-SN JUAN DEL RIO KM 20.5 TEQUISQUIAPAN QRO.</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES PEÑUELAS S.A. DE C.V.</Company>
    <Phone>2460514(15)2207361</Phone>
    <Address>PLATEROS 84SAN PEDRITO PE#UELAS CP 76148 CD QUERETARO EDO QU</Address>
  </Distributor>
  <Distributor>
    <Company>HERNANDEZ GUZMAN JOSE LUIS</Company>
    <Phone>42170220</Phone>
    <Address>DOMICILIO CONOCIDO S/N RANCHO EL ROSARIO (EL MARQUEZ)</Address>
  </Distributor>
  <Distributor>
    <Company>GRUPO NUEVA OXTOTITLAN, S.A. DE C.V.</Company>
    <Phone>01 72 (17 90 02) (17 89 62)</Phone>
    <Address>VENUSTIANO CARRANZA PTE. 2725COL. SEMINARIO 2A.SECCION CP 50170 CD TOLUCA EDO MX</Address>
  </Distributor>
  <Distributor>
    <Company>MAT DE CONST CIAL MICHOACANA SA DE CV</Company>
    <Phone>0043120006</Phone>
    <Address>GUADALUPE VICTORIA 585INDUSTRIAL CP 58130 CD MORELIA EDO MI</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES PERALDI DE LA COSTA</Company>
    <Phone>53 7 15 66 53 7 44 55</Phone>
    <Address>AV. MELCHOR OCAMPO #1079 CP 60950 CD CIUDAD LAZARO CARDENAS EDO MI</Address>
  </Distributor>
  <Distributor>
    <Company>NAVARRO LEDEZMA FEDERICO</Company>
    <Phone>55-15-56-55</Phone>
    <Address>AV.TOLTECA 83 SAN PEDRO DE LOS PINOS CP 01180 CD ALVARO OBREGON EDO DF</Address>
  </Distributor>
  <Distributor>
    <Company>DISTRIBUIDORA DE MATERIALES ARIES SA CV</Company>
    <Phone>54-28-34-64, 54-28-62-30</Phone>
    <Address>AV. B.JUAREZ 365 MZ.19,LT1 POLVORILLA CP 09750 CD CIUDAD DE MEXICO EDO DF</Address>
  </Distributor>
  <Distributor>
    <Company>DISTRIB.DE MATS. P/CONST. E B,SA DE CV</Company>
    <Phone>55-32-18-96</Phone>
    <Address>NORMANDIA 3PORTALES CP 03300 CD CIUDAD DE MEXICO EDO DF</Address>
  </Distributor>
  <Distributor>
    <Company>MATS. P/CONST. Y FERROTLAPALERIA CHICAGO</Company>
    <Phone>55639458 56150373</Phone>
    <Address>AV. CHICAGO LT.24 MZ.18-ALOMAS DE BECERRA CP CD ALVARO OBREGON EDO DF</Address>
  </Distributor>
  <Distributor>
    <Company>GOMEZ LUNA LUIS MANUEL</Company>
    <Phone>56944689/9345</Phone>
    <Address>CALLE 16 MZ.47 LT.480 COL. LEYES DE REFORMA 2DA. SECC. CP 09310 DEL. IZTAPALAPA CD. DE MEXICO,
D.F.</Address>
  </Distributor>
  <Distributor>
    <Company>EL IMPERIO CONSTRURAMA</Company>
    <Phone>0159770731</Phone>
    <Address>JUAN FLORES I CASAS COL CENTRO CP 56860 CD JUCHITEPEC EDO MX</Address>
  </Distributor>
  <Distributor>
    <Company>SURTIDORA DE LA VIVIENDA,S.A. DE C.V.</Company>
    <Phone>55987899 y 56433097</Phone>
    <Address>ANDREA DEL CASTAGNO 82NONOALCO CP 03700 CD CIUDAD DE MEXICO EDO DF</Address>
  </Distributor>
  <Distributor>
    <Company>RIVERA CONTRERAS MARIANO</Company>
    <Phone>54284712</Phone>
    <Address>AV.VILLABRAZ MZ 20 L21COL. DESARROLLO UBANO QUETZALCOATL CP 09700 CD CIUDAD DE MEXICO EDO
DF</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES Y FERRETERIA ELIZONDO GUADALUPE</Company>
    <Phone>01 81 8367 1110</Phone>
    <Address>AV. BENITO JUAREZ 401 OTE. CENTRO</Address>
  </Distributor>
  <Distributor>
    <Company>SANTOS GONZALEZ JOSE ALFREDO</Company>
    <Phone>01 866 640 8759</Phone>
    <Address>BLVD.SAN BUENA PONIENTE 1300 SAN FRANCISCO CP 25720</Address>
  </Distributor>
  <Distributor>
    <Company>CENTRAL FERRETERA Y MADERERA S.A. DE C.V</Company>
    <Phone>(877) 772-5176, 772-24471</Phone>
    <Address>GUERRERO SUR 990CENTRO</Address>
  </Distributor>
  <Distributor>
    <Company>MADERERIA MARBA S.A.</Company>
    <Phone>01 878 782 0499</Phone>
    <Address>AV. LOPEZ MATEOS 415</Address>
  </Distributor>
  <Distributor>
    <Company>ALMAGUER GUANAJUATO MANUEL</Company>
    <Phone>01 81 8341 3483</Phone>
    <Address>GUINEA 204 LOMAS DEL PEDREGAL</Address>
  </Distributor>
  <Distributor>
    <Company>MADERERIA VILLARREAL</Company>
    <Phone>01 827 285 0053</Phone>
    <Address>CARRETERA NACIONAL KM. 226EL CERCADO</Address>
  </Distributor>
  <Distributor>
    <Company>DISTRIBUIDORA FERRETERA GARZA SANTOS S.A</Company>
    <Phone>824 2421316</Phone>
    <Address>S.A. DE C.V.CUAUHTEMOC PTE. 390CENTRO</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES INTERNACIONALES, S.A DE C.V.</Company>
    <Phone>01 867 714 4225</Phone>
    <Address>JESUS CARRANZA 2717 GUERRERO</Address>
  </Distributor>
  <Distributor>
    <Company>MAT. Y ACEROS DE CADEREYTA (MATRIZ)</Company>
    <Phone>01 81 8284 0407</Phone>
    <Address>CUAUHTEMOC NTE. 311</Address>
  </Distributor>
  <Distributor>
    <Company>PRODUCTOS IND. DE CONC. DE P.N. S.A.C.V.</Company>
    <Phone>01 878 782 2854</Phone>
    <Address>PROLONGACION VICTORIA NTE. 2710 MUNDO NUEVO</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES CONSTRUCTORES DE MONCLOVA, S.A.</Company>
    <Phone>01 866 636 6261/ 636 5339</Phone>
    <Address>PINO SUAREZ # 313 FRACCIONAMIENTO MONCLOVA</Address>
  </Distributor>
  <Distributor>
    <Company>MATERIALES BRISAN DE MONCLOVA</Company>
    <Phone>866 6335242</Phone>
    <Address>.V.JESUS SILVA 468 CENTRO</Address>
  </Distributor>
  <Distributor>
```

```
        <Company>MADERAS Y MATERIALES SALDAÑA</Company>
        <Phone>894 8422496 Y 894 8421360</Phone>
        <Address>KM .82 BRECHA 117.500 RUMBO AL REALITO</Address>
    </Distributor>
    <Distributor>
        <Company>MATERIALES Y BLOCKS SN MIGUEL (MATRIZ)</Company>
        <Phone>01 81 8364 5210</Phone>
        <Address>AVENIDA PILASTRA 610 COL. VILLA DE SAN MIGUEL</Address>
    </Distributor>
    <Distributor>
        <Company>FERRETERA GIRASOL/DAVILA SANCHEZ JOSE GUADALUPE</Company>
        <Phone>844 4170302</Phone>
        <Address>W. GONZALEZ # 2268 AMPLIACIÓN GIRASOL</Address>
    </Distributor>
    <Distributor>
        <Company>CONPAES</Company>
        <Phone>868 8161113</Phone>
        <Address>DIVISION DEL NORTE 4 DELICIAS</Address>
    </Distributor>......................................................................................................................... ETC...............................
ETC.............
```

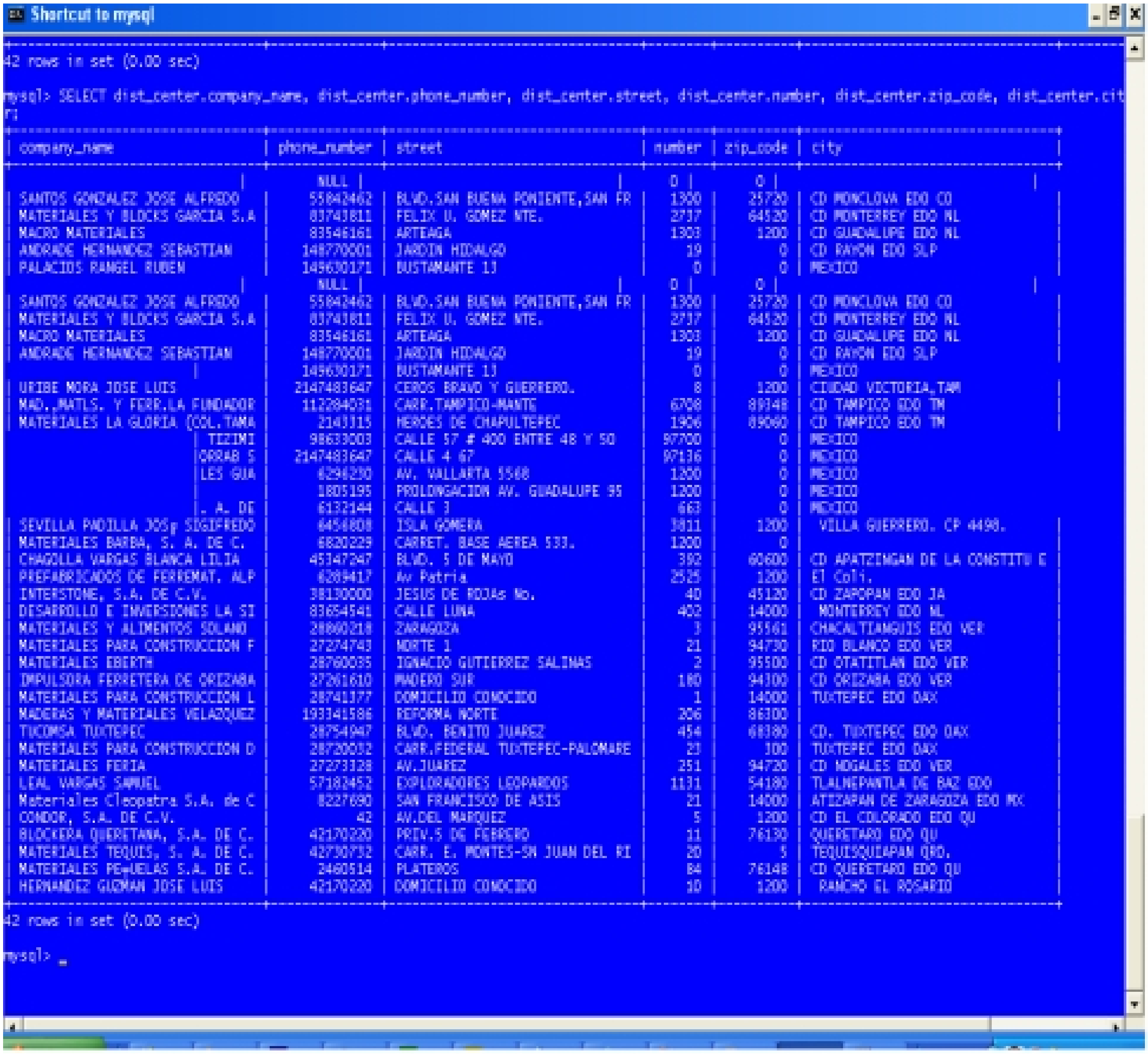
Appendix B



```
mysql> SELECT dist_center.company_name, dist_center.phone_number, dist_center.street, dist_center.number, dist_center.zip_code, dist_center.cit
untry FROM dist center;
+--------------------------------+-----------+------------------------------+---------------+----------------------------------
| company_name                   | phone_number | street                    | number  zip_code | city                       country
+--------------------------------+-----------+------------------------------+---------------+----------------------------------
                             |      NULL |                              |    0 |      0 |                                |      |
|SANTOS GONZALEZ JOSE ALFREDO    |  55842462 | BLVD.SAN BUENA PONIENTE,SAN FR |  1300    25720 | CD MONCLOVA EDO CO          MEXICO
|MATERIALES Y BLOCKS GARCIA S.A  |  83743311 | FELIX U. GOMEZ NTE.          |  2737    64520 | CD MONTERREY EDO NL         MEXICO
|MACRO MATERIALES                |  83546161 | ARTEAGA                      |  1303     1200 | CD GUADALUPE EDO NL         MEXICO
|ANDRADE HERNANDEZ SEBASTIAN     | 148770001 | JARDIN HIDALGC               |    19        0 | CD RAYON EDO SLP            MEXICO
| PALACIOS RANGEL RUBEN          | 149630171 | BUSTAMANTE 13                |     0        0 | MEXICO
                             |      NULL |                              |    0 |      0 |                                |      |
|SANTOS GONZALEZ JOSE ALFREDO    |  55842462 | BLVD.SAN BUENA PONIENTE,SAN FR |  1300    25720 | CD MONCLOVA EDO CO          MEXICO
|MATERIALES Y BLOCKS GARCIA S.A  |  83743311 | FELIX U. GOMEZ NTE.          |  2737    64520 | CD MONTERREY EDO NL         MEXICO
|MACRO MATERIALES                |  83546161 | ARTEAGA                      |  1303     1200 | CD GUADALUPE EDO NL         MEXICO
|ANDRADE HERNANDEZ SEBASTIAN     | 148770001 | JARDIN HIDALGC               |    19        0 | CD RAYON EDO SLP            MEXICO
                     |           149630171 | BUSTAMANTE 13                |     0        0 | MEXICO
|URIBE MORA JOSE LUIS            | 2147483647 | CEROS BRAVO Y GUERRERO.     |     8     1200 | CIUDAD VICTORIA,TAM         MEXICO
|MAD.,MATLS. Y FERR.LA FUNDADOR  | 112284031 | CARR.TAMPICO-MANTE           |  6708    89348 | CD TAMPICO EDO TM           MEXICO
|MATERIALES LA GLORIA (COL.TAMA  |   2143315 | HEROES DE CHAPULTEPEC        |  1906    89060 | CD TAMPICO EDO TM           MEXICO
                     |            98633007 | CALLE 57 # 400 ENTRE 48 Y 50 | 97700        0 | MEXICO
                     |          2147483647 | CALLE 4 67                   | 97136        0 | MEXICO
                     |             6296230 | AV. VALLARTA 5563            |  1200        0 | MEXICO
                     |             1805195 | PROLONGACION AV. GUADALUPE 95 |  1200        0 | MEXICO
                     |             6132144 | CALLE 3                      |   663        0 | MEXICO
|SEVILLA PADILLA JOS╓ SIGIFREDO  |   6456308 | ISLA GOMERA                  |  3811     1200 |  VILLA GUERRERO. CP 4498.   MEXICO
| MATERIALES BARBA, S. A. DE C.  |   6620229 | CARRET. BASE AEREA 533.      |  1200        0 |
|CHAGOLLA VARGAS BLANCA LILIA    |  45347247 | BLVD. 5 DE MAYO              |   392    60600 | CD APATZINGAN DE LA CONSTITU E   MEXICO
|PREFABRICADOS DE FERREMAT. ALP  |   6289417 | Av Patria                    |  2525     1200 | El Culi.                    MEXICO
|INTERSTONE, S.A. DE C.V.        |  38130000 | JESUS DE ROJAS No.           |    40    45120 | CD ZAPOPAN EDO JA           MEXICO
|DESARROLLO E INVERSIONES LA SI  |  83654541 | CALLE LUNA                   |   402    14000 |  MONTERREY EDO NL           MEXICO
|MATERIALES Y ALIMENTOS SOLANO   |  28860218 | ZARAGOZA                     |     3    95561 | CHACALTIANGUIS EDO VER      MEXICO
|MATERIALES PARA CONSTRUCCION F  |  27274743 | NORTE 1                      |    21    94730 | RIO BLANCO EDO VER          MEXICO
|MATERIALES EDERTH               |  28760005 | IGNACIO GUTIERREZ SALINAS    |     2    95500 | CD OTATITLAN EDO VER        MEXICO
|IMPULSORA FERRETERA DE ORIZABA  |  27261610 | MADERO SUR                   |   180    94300 | CD ORIZABA EDO VER          MEXICO
|MATERIALES PARA CONSTRUCCION L  |  28741377 | DOMICILIO CONOCIDO           |     1    14000 | TUXTEPEC EDO OAX            MEXICO
|MADERAS Y MATERIALES VELAZQUEZ  | 193341586 | REFORMA NORTE                |   206    86300 |                             MEXICO
|TUCOMSA TUXTEPEC                |  28754947 | BLVD. BENITO JUAREZ          |   454    68380 | CD. TUXTEPEC EDO OAX        MEXICO
|MATERIALES PARA CONSTRUCCION D  |  28720032 | CARR.FEDERAL TUXTEPEC-PALOMARE |    20      300 | TUXTEPEC EDO OAX          MEXICO
|MATERIALES FERIA                |  27273328 | AV.JUAREZ                    |   251    94720 | CD NOGALES EDO VER          MEXICO
|LEAL VARGAS SAMUEL              |  57182452 | EXPLORADORES LEOPARDOS       |  1131    54180 | TLALNEPANTLA DE BAZ EDO     MEXICO
|Materiales Cleopatra S.A. de C  |   8227390 | SAN FRANCISCO DE ASIS        |    21    14000 | ATIZAPAN DE ZARAGOZA EDO MX MEXICO
|CONDOR, S.A. DE C.V.            |        42 | AV.DEL MARQUEZ               |     5     1200 | CD EL COLORADO EDO QU       MEXICO
|BLOCKERA QUERETANA, S.A. DE C.  |  42170220 | PRIV.5 DE FEBRERO            |    11    76130 | QUERETARO EDO QU            MEXICO
|MATERIALES TEQUIS, S. A. DE C.  |  42730732 | CARR. F. MONTES-SN JUAN DEL RT |    20        5 | TEQUISQUITAPAN QRO.       MEXICO
|MATERIALES PE╤UELAS S.A. DE C.  |   2460514 | PLATEROS                     |    84    76148 | CD QUERETARO EDO QU         MEXICO
|FERNANDEZ GUZMAN JOSE LUIS      |  42170220 | DOMICILIO CONOCIDO           |    10     1200 |  RANCHO EL ROSARIO          MEXICO
+--------------------------------+-----------+------------------------------+---------------+----------------------------------
42 rows in set (0.00 sec)

mysql>
```

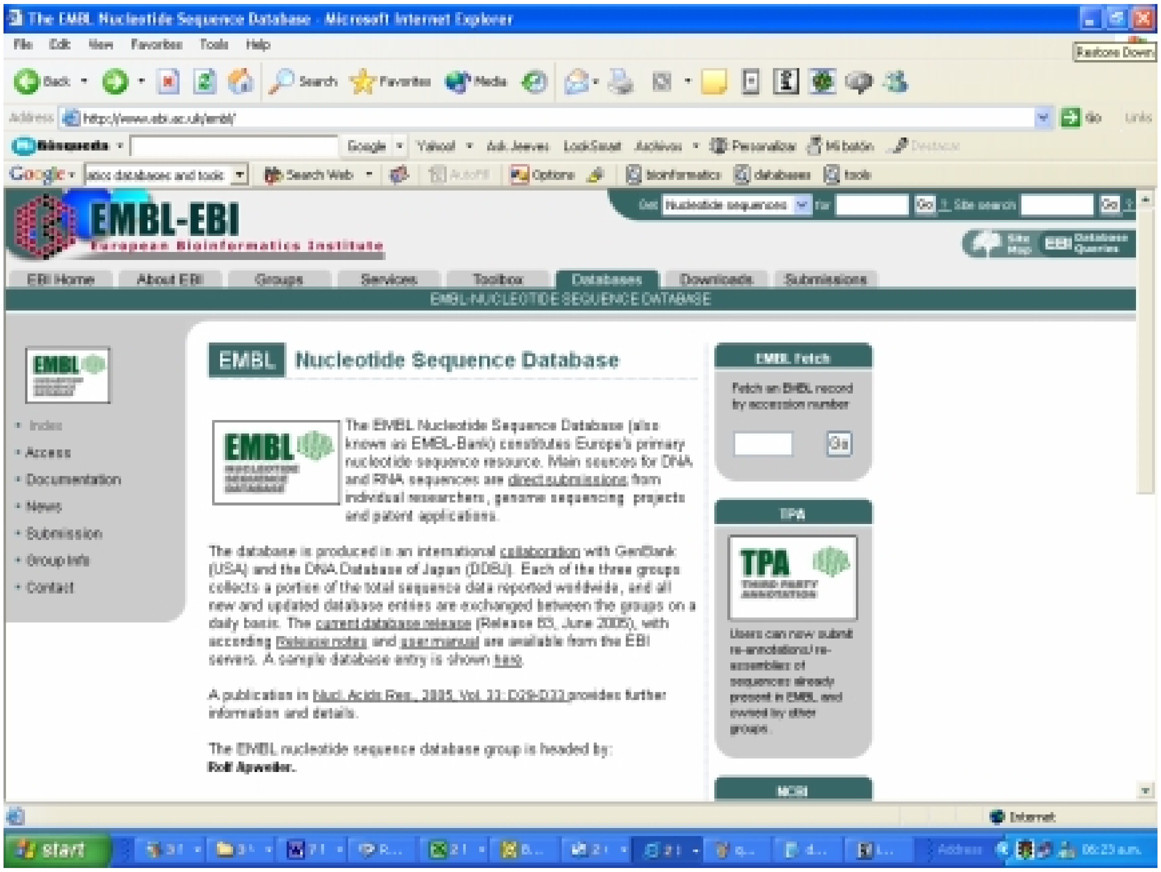


```
mysql> CREATE TABLE dist_center
    -> (
    ->     company_name    CHAR(30) NOT NULL,
    ->     phone_number    INT NULL,
    ->     street          CHAR(30) NOT NULL,
    ->     number          INT NOT NULL,
    ->     zip_code        INT NOT NULL,
    ->     city            CHAR(30) NOT NULL,
    ->     country         CHAR(30) NOT NULL
    -> );
Query OK, 0 rows affected (0.24 sec)

mysql> DESCRIBE dist_center;
+--------------+----------+------+-----+---------+-------+
| Field        | Type     | Null | Key | Default | Extra |
+--------------+----------+------+-----+---------+-------+
| company_name | char(30) |      |     |         |       |
| phone_number | int(11)  | YES  |     | NULL    |       |
| street       | char(30) |      |     |         |       |
| number       | int(11)  |      |     | 0       |       |
| zip_code     | int(11)  |      |     | 0       |       |
| city         | char(30) |      |     |         |       |
| country      | char(30) |      |     |         |       |
+--------------+----------+------+-----+---------+-------+
7 rows in set (0.11 sec)

mysql>
mysql> _
```



```
+--------------------------------+--------------+--------------------------------+--------+----------+----------------------------------+
42 rows in set (0.00 sec)

mysql> SELECT dist_center.company_name, dist_center.phone_number, dist_center.street, dist_center.number, dist_center.zip_code, dist_center.cit
r;
+--------------------------------+--------------+--------------------------------+--------+----------+----------------------------------+
| company_name                   | phone_number | street                         | number | zip_code | city                             |
+--------------------------------+--------------+--------------------------------+--------+----------+----------------------------------+
                                |         NULL |                                |      0 |        0 |                                |
| SANTOS GONZALEZ JOSE ALFREDO   |     55842462 | BLVD.SAN BUENA PONIENTE,SAN FR |   1300 |    25720 | CD MONCLOVA EDO CO               |
| MATERIALES Y BLOCKS GARCIA S.A |     83743811 | FELIX U. GOMEZ NTE.            |   2737 |    64520 | CD MONTERREY EDO NL              |
| MACRO MATERIALES               |     83546161 | ARTEAGA                        |   1303 |     1200 | CD GUADALUPE EDO NL              |
| ANDRADE HERNANDEZ SEBASTIAN    |    148770001 | JARDIN HIDALGO                 |     19 |        0 | CD RAYON EDO SLP                 |
| PALACIOS RANGEL RUBEN          |    149630171 | BUSTAMANTE 1J                  |      0 |        0 | MEXICO                           |
                                |         NULL |                                |      0 |        0 |                                |
| SANTOS GONZALEZ JOSE ALFREDO   |     55842462 | BLVD.SAN BUENA PONIENTE,SAN FR |   1300 |    25720 | CD MONCLOVA EDO CO               |
| MATERIALES Y BLOCKS GARCIA S.A |     83743811 | FELIX U. GOMEZ NTE.            |   2737 |    64520 | CD MONTERREY EDO NL              |
| MACRO MATERIALES               |     83546161 | ARTEAGA                        |   1303 |     1200 | CD GUADALUPE EDO NL              |
| ANDRADE HERNANDEZ SEBASTIAN    |    148770001 | JARDIN HIDALGO                 |     19 |        0 | CD RAYON EDO SLP                 |
                         |    149630171 | BUSTAMANTE 1J                  |      0 |        0 | MEXICO
| URIBE MORA JOSE LUIS           |   2147483647 | CEROS BRAVO Y GUERRERO.        |      8 |     1200 | CIUDAD VICTORIA,TAM              |
| MAD.,MATLS. Y FERR.LA FUNDADOR |    112284031 | CARR.TAMPICO-MANTE             |   6708 |    89348 | CD TAMPICO EDO TM                |
| MATERIALES LA GLORIA (COL.TAMA |      2143315 | HEROES DE CHAPULTEPEC          |   1906 |    89060 | CD TAMPICO EDO TM                |
                         | TIZIMI |     98633003 | CALLE 57 # 400 ENTRE 48 Y 50   |  97700 |        0 | MEXICO
                         |ORRAB S |   2147483647 | CALLE 4 67                     |  97136 |        0 | MEXICO
                         |LES GUA |      6296230 | AV. VALLARTA 5568              |   1200 |        0 | MEXICO
                         |        |      1805195 | PROLONGACION AV. GUADALUPE 95  |   1200 |        0 | MEXICO
                         |. A. DE |      6132144 | CALLE 3                        |    663 |        0 | MEXICO
| SEVILLA PADILLA JOS¿ SIGIFREDO |      6456808 | ISLA GOMERA                    |   3811 |     1200 |  VILLA GUERRERO. CP 4498.        |
| MATERIALES BARBA, S. A. DE C.  |      6820229 | CARRET. BASE AEREA 533.        |   1200 |        0 |                                  |
| CHAGOLLA VARGAS BLANCA LILIA   |     45347247 | BLVD. 5 DE MAYO                |    392 |    60600 | CD APATZINGAN DE LA CONSTITU E   |
| PREFABRICADOS DE FERREMAT. ALP |      6289417 | Av Patria                      |   2525 |     1200 | El Coli.                         |
| INTERSTONE, S.A. DE C.V.       |     38130000 | JESUS DE ROJAs No.             |     40 |    45120 | CD ZAPOPAN EDO JA                |
| DESARROLLO E INVERSIONES LA SI |     83654541 | CALLE LUNA                     |    402 |    14000 |  MONTERREY EDO NL                |
| MATERIALES Y ALIMENTOS SOLANO  |     28860218 | ZARAGOZA                       |      3 |    95561 | CHACALTIANGUIS EDO VER           |
| MATERIALES PARA CONSTRUCCION F |     27274743 | NORTE 1                        |     21 |    94730 | RIO BLANCO EDO VER               |
| MATERIALES EBERTH              |     28760035 | IGNACIO GUTIERREZ SALINAS      |      2 |    95500 | CD OTATITLAN EDO VER             |
| IMPULSORA FERRETERA DE ORIZABA |     27261610 | MADERO SUR                     |    180 |    94300 | CD ORIZABA EDO VER               |
| MATERIALES PARA CONSTRUCCION L |     28741377 | DOMICILIO CONOCIDO             |      1 |    14000 | TUXTEPEC EDO OAX                 |
| MADERAS Y MATERIALES VELAZQUEZ |    193341586 | REFORMA NORTE                  |    206 |    86300 |                                  |
| TUCOMSA TUXTEPEC               |     28754947 | BLVD. BENITO JUAREZ            |    454 |    68380 | CD. TUXTEPEC EDO OAX             |
| MATERIALES PARA CONSTRUCCION D |     28720032 | CARR.FEDERAL TUXTEPEC-PALOMARE |     23 |      300 | TUXTEPEC EDO OAX                 |
| MATERIALES FERIA               |     27273328 | AV.JUAREZ                      |    251 |    94720 | CD NOGALES EDO VER               |
| LEAL VARGAS SAMUEL             |     57182452 | EXPLORADORES LEOPARDOS         |   1131 |    54180 | TLALNEPANTLA DE BAZ EDO          |
| Materiales Cleopatra S.A. de C |      8227690 | SAN FRANCISCO DE ASIS          |     21 |    14000 | ATIZAPAN DE ZARAGOZA EDO MX      |
| CONDOR, S.A. DE C.V.           |           42 | AV.DEL MARQUEZ                 |      5 |     1200 | CD EL COLORADO EDO QU            |
| BLOCKERA QUERETANA, S.A. DE C. |     42170220 | PRIV.5 DE FEBRERO              |     11 |    76130 | QUERETARO EDO QU                 |
| MATERIALES TEQUIS, S. A. DE C. |     42730732 | CARR. E. MONTES-SN JUAN DEL RI |     20 |        5 | TEQUISQUIAPAN QRO.               |
| MATERIALES PE¤UELAS S.A. DE C. |      2460514 | PLATEROS                       |     84 |    76148 | CD QUERETARO EDO QU              |
| HERNANDEZ GUZMAN JOSE LUIS     |     42170220 | DOMICILIO CONOCIDO             |     10 |     1200 |  RANCHO EL ROSARIO               |
+--------------------------------+--------------+--------------------------------+--------+----------+----------------------------------+
42 rows in set (0.00 sec)

mysql> _
```

Appendix C


The EMBL Nucleotide Sequence Database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ebi.ac.uk/embl/
Go Links
EMBL-EBI
European Bioinformatics Institute
EBI Home
About EBI
Groups
Services
Toolbox
Databases
Downloads
Submissions
EMBL-NUCLEOTIDE SEQUENCE DATABASE
Index
Access
Documentation
News
Submission
Group Info
Contact
EMBL Fetch
TPA
NCBI
start
Internet

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 83, June 2005), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in Nucl. Acids Res., 2005, Vol. 33: D29-D33 provides further information and details.

The EMBL nucleotide sequence database group is headed by:
Rolf Apweiler.

Fetch an EMBL record by accession number

Users can now submit re-annotations/ re-assemblies of sequences already present in EMBL and owned by other groups.

Appendix D

EBI Dbfetch

```
ID   AF033812   standard; genomic RNA; VRL; 5894 BP.
XX
AC   AF033812;
XX
SV   AF033812.1
XX
DT   23-JAN-1998 (Rel. 54, Created)
DT   15-APR-2005 (Rel. 83, Last updated, Version 4)
XX
DE   Abelson murine leukemia virus, complete genome.
XX
KW   .
XX
OS   Abelson murine leukemia virus
OC   Viruses; Retro-transcribing viruses; Retroviridae; Orthoretrovirinae;
OC   Gammaretrovirus; 1-Mammalian type C virus group.
XX
RN   [1]
RP   1-5894
RA   Petropoulos C.J.;
RT   "Appendix 2: Retroviral taxonomy, protein structure, sequences, and genetic
RT   maps";
RL   (in) Coffin J.M. (eds.);
RL   RETROVIRUSES:757-0;
RL   Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, NY, USA
RL   (1997)
XX
RN   [2]
RP   1-5894
RA   Chappey C.;
RT   ;
RL   Submitted (12-NOV-1997) to the EMBL/GenBank/DDBJ databases.
RL   NIH, NLM, Rockville Pike, Bethesda, MD 20894, USA
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..5894
FT                   /db_xref="taxon:11788"
FT                   /mol_type="genomic RNA"
FT                   /organism="Abelson murine leukemia virus"
FT   mRNA            69..5070
FT                   /gene="gag-abl"
FT                   /product="p120 polyprotein"
FT   5'UTR           69..145
FT   CDS             621..3566
FT                   /codon_start=1
FT                   /db_xref="GOA:O92809"
FT                   /db_xref="HSSP:1OPJ"
FT                   /db_xref="InterPro:IPR000719"
FT                   /db_xref="InterPro:IPR000840"
FT                   /db_xref="InterPro:IPR000980"
FT                   /db_xref="InterPro:IPR001245"
FT                   /db_xref="InterPro:IPR002079"
FT                   /db_xref="InterPro:IPR003036"
FT                   /db_xref="InterPro:IPR008266"
FT                   /db_xref="UniProt/TrEMBL:O92809"
FT                   /gene="gag-abl"
FT                   /product="p120 polyprotein"
FT                   /protein_id="AAC82569.1"
FT                   /translation="MGQTVTTPLSLTLGHWKDVERIAHNQSVDVKKRRWVTFCSAEWPT
FT                   FNVGWPRDGTFNRDLITQVKIKVFSPGPHGHPDQVPYIVTWEALAFDPPPWVKPFVHPK
FT                   PPPPLPPSAPSLPLEPPLSTPPRSSLYPALTPSLGAKPKPQVLSDSGGPLIDLLTEDPP
FT                   PYRDPRPPPSDRDGNGGEATPAGEAPDPSPMASRLRGRREPPVADSTTSQAFPLRTGGN
FT                   GQLQYWPFSSSDLYITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQ
FT                   RSISLRYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPK
FT                   RNKPTIYGVSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTM
FT                   EVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVSA
FT                   VVLLYMATQISSAMEYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAH
FT                   AGAKFPIKWTAPESLAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKD
FT                   YRMERPEGCPEKVYELMRACWQWNPSDRPSFAEIHQAFETMFQESSISDEVEKELGKRG
FT                   TRGGAGSMLQAPELPTKTRTCRRAAEQKASPPSLTPKLLRRQVTASPSSGLSHKKEATK
FT                   GSASGMGTPATAEPAPPSNKVGLSKASSEEMRVRRHKHSSESPGRDKGRLAKLKPAPPP
FT                   PPACTGKAGKPAQSPSQEAGEAGGPTKTKCTSLAMDAVNTDPTKAGPPGEGLRKPVPPS
FT                   VPKPQSTAKPPGTPTSPVSTPSTAPAPSPLAGDQQPSSAAFIPLISTRVSLRKTRQPPE
FT                   RIASGTITKGVVLDSTEALCLAISRNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRN
FT                   KFAFREAINKLESNLRELQICPATASSGPAATQDFSKLLSSVKEISDIVRR"
FT   mat_peptide     624..1016
FT                   /gene="gag-abl"
FT                   /product="MA"
FT   mat_peptide     1014..1268
FT                   /gene="gag-abl"
FT                   /product="CA"
FT   mat_peptide     1326..3563
FT                   /gene="gag-abl"
FT                   /product="ABL"
FT   3'UTR           4621..5070
XX
SQ   Sequence 5894 BP; 1361 A; 1704 C; 1550 G; 1279 T; 0 other;
     gcgccagtcc tccgagtgac tgagtcgccc gggtacccgt gtatccaata aaccctcttg        60
     cagttgcatc cgacttgtgg tctcgctgtt ccttgggagg gtctcctctg agtgattgac       120
     tacccgtcag cgggggtctt tcatttgggg gctcgtccgg gatcgggaga cccctgccca       180
     gggaccaccg acccaccacc gggaggtaag ctggccagca acttatctgt gtctgtccga       240
     ttgtctagtg tctatgactg attttatgcg cctgcgtcgg tactagttag ctaactagct       300
     ctgtatctgg cggacccgtg gtggaactga cgagttcgga acacccggcc gcaaccctgg       360
     gagacgtcca agggacttcg ggggccgttt ttgtggcccg acctgagtcc aaaaatcccg       420
     atcgttttgg actctttggt gcacccccct tagaggaggg atatgtggtt ctggtaggag       480
     acgagaacct aaaacagttc ccgcctccgt ctgaattttt gctttcggtt tggaaccgaa       540
     gccgcgccgc gcgtcttgtc tgctgcagca tcgttctgtg ttgtctctgt ctgactgtgt       600
     ttctgtagtt gtctgaaaat atgggccaga ctgttaccac tcccttaagt ttgactttag       660
     gtcactggaa agatgtcgag cggatcgctc acaaccagtc ggttgatgtc aagaagagac       720
     gttgggttac cttctgctct gcagaatggc caacctttaa cgtcggatgg ccgcgagacg       780
     gcacctttaa ccgagacctc atcacccagg ttaagatcaa ggtcttttca cctggcccgc       840
     atggacaccc agaccaggtc ccctacatcg tgacctggga agccttggct tttgaccccc       900
     ctccctgggt caagcccttt gtacacccta agcctccgcc tcctcttcct ccatccgccc       960
     cgtctctccc ccttgaacct cctctttcga ccccgcctcg atcctccctt tatccagccc      1020
     tcactccttc tctaggcgcc aaacctaaac ctcaagttct ttctgacagt gggggccgc      1080
     tcatcgacct acttacagaa gaccccccgc cttataggga cccaagacca ccccccttccg      1140
     acagggacgg aaatggtgga gaagcgaccc ctgcgggaga ggcaccggac cctccccaa      1200
     tggcatctcg cctgcgtggg agacgggagc cccagtggc cgactccact acctcgcagg      1260
     cattccccct ccgcacagga ggaaacggac agcttcaata ctggccgttc tcctcttctg      1320
     acctttacat cacccccgtc aacagcctgg agaaacattc ctggtatcat ggccctgtat      1380
     ctcggaatgc tgctgagtat ctgctgagca gcggaatcaa cggcagcttc ttagtgcggg      1440
     agagtgagag tagccctggc cagagatcca tctcgctgcg gtatgaaggg agggtgtacc      1500
     actacaggat caacactgcc tctgatggca agctgtacgt gtcctccgag agccgcttca      1560
     acactctggc tgagttagtt caccatcact ccacggtggc tgatggcctc atcaccacac      1620
     tccactaccc agctcccaag cgcaacaagc ccactatcta cggtgtgtcc cccaactacg      1680
     acaagtggga aatggagcgc accgacatca ccatgaagca caagttgggt ggaggccagt      1740
     acggggaggt gtacgagggc gtttggaaga agtacagcct cactgtggcc gtgaagacct      1800
     tgaaggagga caccatggag gtggaggagt tcctgaagga agcggcggtg atgaaggaga      1860
     tcaaacaccc taacctggtg cagctgctag ggtgtgtac ccgggaacca ccattctaca      1920
     taatcactga gttcatgacc tatgggaacc tgctggacta cctgagggag tgtaaccggc      1980
     aggaggtgag cgccgtggta ctgctctaca tggccacaca gatctcatca gccatggagt      2040
     acttggagaa gaagaacttc atccacagag accttgctgc ccggaactgc ctggtagggg      2100
     aaaaccactt ggtgaaggtg gctgattttg gcctgagcag gttgatgaca ggggacacct      2160
     acacggccca tgctggagcc aaattcccca tcaaatggac cgcacctgag agcctggcct      2220
     acaacaagtt ctccatcaag tcggacgtgt gggcatttgg agtattgctc tgggagattg      2280
     ctacctatgg catgtcacct tacccgggaa ttgacctgtc tcaggtttat gagctgctgg      2340
     aaaaagacta ccgcatggag cgccctgaag gctgcccgga gaaggtctac gagctcatgc      2400
     gagcatgttg gcagtggaac ccctctgacc ggccctcctt tgctgaaatc caccaagcct      2460
     ttgaaaccat gttccaggaa tccagtatct cagatgaggt ggagaaggag ctggggaaac      2520
     gaggcacgag aggaggtgct gggagtatgc tgcaggcccc agagctgccc accaagacca      2580
     gaacctgcag gagagcagct gagcagaaag ccagccctcc cagcttgact cccaaactcc      2640
     tccgcaggca ggtcactgcc tctccttcct ctggcctctc tcacaagaaa gaggccacca      2700
     agggcagtgc ctcaggcatg gggactccgg ccactgcaga gccagcaccc cccagcaaca      2760
     aagtgggcct cagcaaggcc tcctctgagg agatgcgcgt aaggaggcac aagcacagct      2820
     cggagtcccc agggagagac aaggggcgac tggctaagct caagcctgcc ccgccgcctc      2880
     ctcctgcctg cacaggaaaa gcaggcaagc ccgcacagag ccccagccaa gaggccgggg      2940
     aggcaggggg gcccacaaag acaaaatgca cgagtctggc tatggatgct gtgaacactg      3000
     accccaccaa ggccggccca cctggagaag gactgagaaa gcctgtgccc ccatctgtgc      3060
     caaagcccca gtcgacggct aagcctccag ggactcccac cagcccggtc tccacccccct      3120
     ccacagcacc agctccttca cccctggctg gggaccagca gccatcttct gccgccttca      3180
     tccccctcat atcaacccgt gtgtctctta ggaagacccg ccagccgcca gagcgcattg      3240
     ccagtggcac catcaccaag ggtgtggttc tggacagtac tgaggccctg tgccttgcca      3300
     tctcccggaa ctcagagcag atggccagcc acagtgctgt actggaggct ggcaagaacc      3360
     tgtacacttt ctgtgtgagc tatgtggact ctatccagca gatgaggaac aagtttgcct      3420
     tccgtgaggc tatcaacaag ctggagagca acctccgaga gctgcagatc tgccctgcca      3480
     cagcctccag tgggccagct gccacccaag acttcagcaa gctgctcagc tctgtgaagg      3540
     agatcagcga cattgtccgg aggtagcagc aaccagtgta tgtcagcaag agatgttgca      3600
     gttcacaggg ctcttgtgcc tataaagatg ggacagggg actggggagc tggcgtcttt      3660
     ccccaggagc tttaaagaga gacaagcaga gcctgaggga gacctggatg gagcctggtg      3720
     gagttggctc ttcctcctgt gttgtgcacc agctgccctg cacctttcct gcccagccca      3780
     ggcgtcagcc acctctcctc actgcctgtg gatgggtctc ctgctctgaa gactacatct      3840
     ggcctgcctg gccaccaggc ttctcactcc ccggtgcctc agacccagct cccaggtcag      3900
     cctggagtgc tcttccctgt ccttgcagaa cgacctcctc tgatggacct tcttgtcacc      3960
     aaggcatggg agcccctgtg cttactgtac ctgcaccttt gatgcttaca aactgtcccc      4020
     gagagcctgt gctcactgtg ttttcattgg aaggaagctg tcgctttaag ggtcatgagg      4080
     tgctaaagcc aggggcccag atgggtgggc actggaaaca ggagctgggc agtgtggtct      4140
     gtcacctgct ctcagtatct tcagcagtgt gcccggcaga tcttggacag caagcttgag      4200
     ttttatgggt ggcagtcact ggctggctag gcacatagcc aggccaaacc taggcctcca      4260
     agggctcccc aaaatctgaa tttctgagta gtcttcatcc cctctcctgc tctaaggtca      4320
     ggtccatcct ctctggtcct taccttgatg acaaggatcc agccttctgg tgtttttgag      4380
     catttcaaag gtctgcatag aaaggaacag ccactaggtc cattatggga cccctcattg      4440
     tactcctaat gattttgctc ttcggaccct gcattcttaa tcgattagtc caatttgtta      4500
     aagacaggat atcagtggtc caggctctag ttttgactca acaatatcac cagctgaagc      4560
     ctatagagta cgagccatag ataaaataaa agattttatt tagtctccag aaaaaggggg      4620
     aatgaaagac cccacctgta ggtttggcaa gctagcttaa gtaacgccat tttgcaaggc      4680
     atggaaaaat acataactga gaatagagaa gttcagatca aggtcaggaa cagagaaaca      4740
     gctgaatatg ggccaaacag gatatgctgt ggtaagcagt tcctgccccg gctcagggcc      4800
     aagaacagtt ggaacagctg aatatgggcc aaacaggata tctgtggtaa gcagttcctg      4860
     ccccggctca gggccaagaa cagatggtcc ccagatgcgg tccagccctc agcagtttct      4920
     agagaaccat cagatgtttc cagggtgccc caaggacctg aaatgaccct gtgccttatt      4980
     tgaactaacc aatcagttcg cttctcgctt ctgttcgcgc gcttctgctc cccgagctca      5040
     ataaaagagc ccacaacccc tcactcggcg cgccagtcct ccgagtgact gagtcgcccg      5100
     ggtacccgtg tatccaataa accctcttgc agttgcatcc gacttgtggt ctcgctgttc      5160
     cttgggaggg tctcctctga gtgattgact acccgtcagc gggggtcttt catgggtaac      5220
     agtttcttga agttggagaa caacattctg agggtaggag tcgaatatta agtaatcctg      5280
     actcaattag ccactgtttt gaatccacat actccaatac tcctgaaata gttcattatg      5340
     gacagcgcag aagagctggg gagaattccc caagcaatta atttcaatgg ggtcagtaag      5400
     gagcaccctg cagtcttgaa aactgtatat ctttgcacga ttctgggtga aagaccccac      5460
     ctgtaggttt ggcaagctag cttaagtaac gccattttgc aaggcatgga aaaatacata      5520
     actgagaata gagaagttca gatcaaggtc aggaacagag aaacagctga atatgggcca      5580
     aacaggatat gctgtggtaa gcagttcctg ccccggctca gggccaagaa cagttggaac      5640
     agctgaatat gggccaaaca ggatatctgt ggtaagcagt tcctgccccg gctcagggcc      5700
     aagaacagat ggtccccaga tgcggtccag ccctcagcag tttctagaga accatcagat      5760
     gtttccaggg tgccccaagg acctgaaatg accctgtgcc ttatttgaac taaccaatca      5820
     gttcgcttct cgcttctgtt cgcgcgcttc tgctccccga gctcaataaa agagcccaca      5880
     acccctcact cggc                                                        5894
//
```

Bibliography

[1] Aaker,D.A. Strategic Market Management. New York: John Wiley and Sons. 1998.

[2] Abitebouls, S. "Querying semi-structured data." Proceedings of the Sixth International Conference in Data Theory, January 8-10, 1997, Delphi,Greece. Ed. F.N. Afrati and P. Kolaitis. Lecture Notes in Computer Science. Vol. 1186, pp. 1-18.

[3] Adelberg, Brad and Matt Denny. "Building Robust Wrappers for Text Sources." Chicago: Department of Computer Science, Northwestern University, 1999. Technical Report. http://www.cs.northwestern.edu/~adelber/papers/www.pdf (Aug. 2002).

[4] Adelberg, Brad. "NoDoSE- A tool for semi-automatically extracting structured and semistructured data from text documents" Proceedings of the ACM SIGMOD International Conference on Management of Data. Seatle, WA, 1998. 283-294.

[5] Alta Vista. http://altavista.com

[6] Arocena, G.O., and A. O. Meldelzon, A.O. "Web OQL: restructuring documents, databases, and webs." Procedding of the Fourteenth International Conference on Data Engineering. Orlando, FL, 1998 24-33.

[7] Attansio, D.B. "The multiple benefits of competitor intelligence." The Journal Business Strategy 9.3 (1998): 16-19.

[8] Atzeni P., G. Mecca, and P. Merialdo. "Semistructured and Structured Data in the Web: Going Back and Forth." Proc. Workshop on Management of Semi-structured Data. Tucson, Arizona, 1997 1-9.

[9] Baumgartner, R., S. Flesca, G. Gottlob. Declarative Information Extraction, Web Crawling, and recursive Wrapping with Lixto.

[10] Baumgartner, R., Flesca, S., Gottlob, G.Visual Web Information Extraction with Lixto. Proceeding of the twenty-seventh International Conference on Very Large Data Bases. Rome, Italy, September 2001.

[11] Baumgartner, R., S. Flesca, and G. Gottlob. Visual Web Information Extraction with Lixto. In Proceedings of the 26th International Conference on Very Large Data Bases. Rome, Italy, 2001 119-128.

[12] Baumgartner,R., O. Frölich, G. Gottlob, P. Harz, M. Herzog, P. Lehmann. Web Data extraction for Business Intelligence: The Lixto Approach. BTW2005, Karlsruhe, Germany, 2005 30-45.

[13] Berners-Lee, R. Cailliav, A.Luotonen, H.F Nielsen, and A Secret. The World Wide Web. Communications of the ACM, 37.8 (1994): 76-82

[14] Bertaccini, E., J. Shapiro, D. Brutlag, J. Trudell. “Homology Modeling or a Human Glycine Alpha 1 receptor Reveals a Pausible Anesthetic Blinding Site.” American Chemical Society 2005, published on web 12/02/2004.

[15] Braganholo, V., S. Davidson, and C. Heuser. “On the updatability of XML views over relational databases” Proceedings of WEBDB June 2003: San Diego, California. 2003

[16] Bray, J., P. Paoh, and C.M Sperberg – McQueen. “Extensible markup language (XML)1.0” February 1998 http://www.w3.org/TR/REC-xml.

[17] Brin, S., R. Motwani, R., L. Page, and T. Winograd. “What can you do with a Web in your pocket?” Data Engineering Bulletin 21, 2 (1998): 37-47.

[18] Buneman, P., S. Davidson, M. Fernandez, and D. Siciu. “Adding Structure to Unstructured Data.” Proceedings of the International Conference on Data Theory (ICDT). Delphi, Greece, 1997 336-350.

[19] Califf. M.E., and R.J. Mooney. “Relational Learning of Pattern-Match Rules for Information Extraction.” Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence. Orlando, FLorida, 1999.

328-334.

[20] Chaudhuri,S., R. Kaushik,and J. Naughton, J. “On relational support for XML Beyond sorting and tagging.” Proceedings of SIGMOD 2003, San Diego, California. Jun 2003.

[21] Cheng, J., and J. Xu, “XML and DB2.” Proceedings of ICDE’00. San Diego, California, 2000.

[22] Atkinson, Colin and Thomas Kühne. “Rearchitecting the UML infrastructure.” ACM Transactions on Modeling and Computer Simulation, October 2003.

[23] Atkinson, Colin and Thomas Kühne. “The essence of multilevel metamodeling” Proceedings of the Fourth International Conferences on the UML 2000, Toronto, Canada, Ed. Martin Gogolla and Cris Kobryn. LNCS 2185 Springer Verlag, October 2001. 19-33

[24] Conferences Board, Competitive Intelligence, 1998, Conferences Board Report No.913. New York: The Conference Board.

[25] Crecenzi, V., and G. Mecca. “Grammar have exceptions.”Information Systems 23.8 (1998): 539-565.

[26] Crescenzi,V., G. Mecca, and P. Merialdo. “Road Runner: Towards automatic data extraction for large Web sites.” Proceedings of the Twenty-sixth International Conference on Very Large Data Bases. Rome, Italy, 2001. 109-118.

[27] Cvitkovic, E. (1989). “Profiling your competitors.” Planning Review 17.3 (1989): 28-30.

[28] D. Florescu, A.Y. Levy, and A.O. Mendelzon. “Database techniques for the world-wide web: A survey.” SIGMOD Record, 27.3 (1998): 59-74

[29] D. Konopnicki and O. Shemuelli. “W3QS: A query system for the World Wide Web” Proceesings of the Twenty-first International Conference on Very Large Databases. Zurich, Switzerland, 1995.

[30] E.P. Lim, W.K. Ng, S.s. Bhowmick, F.Q. Qin, and X.Ye. “A data warehousing system for web information.” Proceedings of the First Asia Digital Library Workshop. Hong Kong, August 1998.

[31] Eikvil, L. “Information Extraction from World Wide Web – A Survey.” Report No. 945, ISBN 82-539-0429-0, July 1999.

[32] Embley, D.W., D.M. Campbell, Y.S. Jian, S.W. Liddle, NG Kai, D. Quass, and R.D. Nad Smith. “Conceptual-model based data extraction from multiple-record Web pages.” Data and Knowledge Engineering 31.3 (1999): 227-251.

[33] Excite. http://www.excite.com

[34] Fernandez, M., D. Florescu, A. Levy, and D. Siciu. “ A Query Language and Processor for a Web-Site Management System.” Proc. Workshop on Management of Semi-structured Data. Tucson, Arizona, 1997.

[35] Fernández, M., Y. Kadiyska, D. Siciu, A. Morishima, and W. Tan. “Silkroute: A Framework from publishing relational data in XML.” ACM Transaction on Databases Systems (TODS) 27.4 (2002): 438-493.

[36] Freitag D. “Information Extraction from HTML: Application of a General Machine Learning Approach.” Proceedings on the Fifteenth National Conference on Artificial Intelligence. Madison, Wisconsin, 1998 517-523.

[37] Freitag, D. “Machine Learning for Information Extraction in Informal Domains.” Machine Learning 39. 2-3 (2000): 169-2002

[38] G. Arocena., and A. Mendelzon. “WebOQL: Restructuring Documents, Databases, and Webs.” Proceedings on the International Conference on Data Engineering, 1998.

[39] Wiederhold, G. “Mediators in the architecture of future information systems.” IEEE Computer, March 1992.

[40] Go (Infoseek). http://www.go.com

[41] Golgher, P., A. Laender, A. Silva and B. Ribeiro-Neto, B. “An Example-Based Environment for Wrapper Generation.” Proceedings of the Second International Workshop on The World Wide Web and Conceptual Modeling, Salt Lake City, Utha, USA, 2000. 152-164

[42] Gruser, J., L. Raschid, M. Vidal, L.Bright. “Wrapper Generation for Web Accesible Data Source.” Proceedings on the Third IFCIS Conference on Cooperative Information Systems (CoopIS). New York, NY , 1998. 14-23.

[43] Hammer, J., H. Garcia-Molina, R. Cho, R. Aranha, and A. Crespo. "Extracting Semistructured Information from the Web". Proceding Workshop on Management of Semi-structured Data. Tucson, Arizona, 1997. 18-25.

[44] Hammer,J., H. Garcia-Molina, S. Nestorov, R. Yerneni, M. Breunig, and V. Vassalos. "Template-based wrappers in the TSIMMIS system." Proceedings of the ACM SIGMOD International Conference on Management of Data. Tucson, Arizona, 1997. 532-535.

[45] Herbert Stachowiack. *Allemeine Modelltheorie*. Spring-Verlag, Wien and New York, 1973.

[46] Herring,J. 1996). "Creating the intelligences system that produces analytical intelligence." Ed. B. Gilad & J.Herring. The Art and sciences of business intelligence analysis Greenwich, CT:JAI Press 53-81

[47] Herzog, M., and G. Gottlob. InfoPipes: A Flexible Framework for M-Commerce Applications, Workshop at VLDB, 2001.

[48] HotBot. http://www.hotbot.com

[49] Hsu, C.-N., and M-T Dung. "Generating finite-state transducers for semi-structured data extraction from the web Information Systems." 23.8 (1998): 521-538.

[50] I.Winship. Web services feautures. Online, February 2001.

[51] INEGI. http://www.inegi.gob.mx/inegi

[52] J.Widow. "Research problems in data warehousing." Proceedings of the Fourth International Conference on Information and Knowledge Management. Baltimore, Maryland, 1995. 25-30,

[53] Thierry- Mieg, Jean and Richard Durbin. AceDB-. AC. "Elegance Database: Syntactic definitions for the AceDB data base management" 1992.

[54] Jochen Ludewing. "Models in software enginerring – an introduction." Journal on Software and System Modeling 2.1 (Mar 2003): 5-14.

[55] Kenneth, T.R., and J.P. Wong, J.P. "The competitive marketing profile of banks, saving and loans, and credit union." Journal of Professional Services

Marketing 3.4 (1998): 107-121.

[56] Kushmerick.N. “Wrapper Induction. Efficiency and expressiveness.” Artificial Intelligence Journal 118. 1-2 (2000): 15-68.

[57] Kusmerick, N., and B. Thomas. “Adaptive Information Extraction: Lore Technologies for Information Agents” Intelligent Information Agents R&D in Europe. An Egent Link perspective, 2002.

[58] Haas, L., D. Kossmann, E. Wimmers, and J. Yang. “Don´t scrap it, wrap it! A wrapper architecture of legacy data source.” Proceedings of the Twenty-third International Conference on Very Large Data Bases, 1997.

[59] Laender, A., B. Ribeiro-Neto, A. Silva, and J. Teixeira. “A Brief Survey of Web Data Extraction Tools.” SIGMOD Record. 31. 2 (June 2002)

[60] Li,C., and J. Wang. “New Invariant of DNA sequences.” American Chemical Society 2005. Published on web 18/08/2004.

[61] Lia, L., C. Pu and W. Han. “XWRAP: An XML- enabled wrapper construction system for Web Information Sources.” Proceedings of the Sixteenth International Conference on Data Engineering. San Diego, California, 2000. 611-621.

[62] Lixto Software GmbH, http://www.lixto.com

[63] Lycos. http://www.lycos.com

[64] Herzog M. and G. Gottlob. InfoPipes: A Flexible Framwork for M- commerce Applications.

[65] Bergman, M.K. “The Deep Web: Surfacing Hidden Value” September 2001.

[66] MapInfo. http://www.mapinfo.com

[67] Mc Callum A., and K. Nigam. “A Comparison of Event Models for Naive Bayes Text Classification.” Proceedings of AAAI-98. Worshop on Learning for Text Categorization, 1998.

[68] Harkavy, Michael et al. Eds. Webster´s new encyclopedic dictionary. New York: Black Dog & Leventhal publishers Inc., 1994.

[69] Mintzberg, H. <u>The rise and fall os trategic planning</u>. New York. NY: Free Press, 1994.

[70] Muslea, I., S. Minton, and C. Knoblock. “Hierarchical wrapper induction for semistructured information sources.” Autonomous Agent and Multi-Agents Systems. 4.1-2 (2001): 93-114.

[71] MySQL. http://www.mysql.com

[72] Kushmerick, N., D.S. Weld, and R. Doorenbos. “Wrapper Induction for Information Extraction.” International Joint Conference on Artificial Intelligence (IJCAI), Nagoya, Japan, 1997.

[73] Ohmae. <u>The mind of strategist.- The Art of Japanese Business</u>. New York: Mac Graw Hill, 1982.

[74] Madsen, Ole L., Kristen Nygaard, and Birger Möller-Pedersen. <u>Object-Oriented Programming in the BETA Programming Language</u>. Addison-Wesley and ACM Press, 1993.

[75] OMG. “Unified Modeling Language Specification, Version 1.4” 2000. Version 1.4, OMG document ad00-11-01.

[76] P. Atzeni, and G. Mecca. “Cut and Paste” Procedings on the Sixteenth ACM Symposium On Principles of Database Systems (PODS). Tucson, Arizona, 1997. 144-153.

[77] Bernstein, P., M.Brodie, S.Ceri, D. DeWitt, M. Franklin, H.Gracia-Molina, J.Gray, J.Hellerstein, H.V. Jagadish, M. Lesk, D.Maier, J.Naughton, H. Pirahesh, M. Stonebraker, and J.Ullman. “The Asilomar Report on Database Research.” ACM SIGMOND Record 27.4 (1998): 74-80

[78] Porter, M. <u>Competitive Advantage: Creating and sustaining competitive advantage</u>. New York: Free Press, 1980.

[79] Porter, M. <u>Competitive strategy: Techniques for analyzing industries and competitors</u>. New York: Free Press. 1980.

[80] Porter, M. “What is strategy?” <u>Harvard Business Review</u>, November-December 96. 6 (1996): 61-78.

[81] R. Baumgartner, S. Flesca, G. Gottlob. "The Elog Web Extraction Language." Logic for Programming, Artificial Intelligence, and Reasoning, Proceedings on the Eighth International Conference, LPAR 2001, Havana Cuba, December 3-7, 2001. Lectures Notes in Computer Sciences 2250 Springer 2001 ISBN 3-540-42957-3.

[82] R. Baumgartner, S. Flesca, G. Gottlob. Visual Web Information Extraction with Lixto. Proceedings of the Twenty-seventh VLDB Conference, Rome, Italy, 2001.

[83] Ram,S., and I.T. Samir. "Competitor analysis practices of US. Companies: An empirical investigation.". Management International Review 38. 1 (1998):7-23.

[84] Ribeiro-Neto, B., A.H.F. Laender,and A.S. Da Silva. "Extracting semi-structured data through examples." Proceedings of the 1999 ACM CIKM International Conference and Information and Knowledge Management. Kansas City, MO, 1999. 94-101.

[85] Baumgartner, Robert, Michael Cesna, Georg Gootlob, Marcus Herzog, and Victor Zigo. "Web Information Acquisition with Lixto Suite." ICDE 2003: 747-749.

[86] Brugger, Rodolf, Fréderic Bapst, and Rolf Ingold. "A DTD Extension for Document Structure Recognition." (1998): 343-354 [DBLP: conf/er/Scherf/B01].

[87] Roy J, and A. Ramanujan. "XML: Data´s Universal Language." IT Pro May 1, June (2000) 32-36

[88] Abiteoul, S., D. Quass, J. MacHugh, J. Widow, and J. Weiner. "The Lorel query language for semistructured data." Journal on Digital Libraries.

[89] Chauhurri, S.and V.Dayal. "An overview of data warehousing and OLAP technology." ACM SIGMOD Record 26.1 (1997): 65-74

[90] S. Lawrence and C.L. Giles. "Accessibility of information on the Web." Nature 400 (July 1999): 107-109

[91] S. Lawrence and C.L. Giles. "Searching the World Wide Web." Science 280. 5360 (April 1998): 98-100

[92] S.S. Bhowmick. WHOM: A Data Model and Algebra for a Web Warehouse. Ph.D dissertation, School of Computer Engineering, Nanyang Technological Univeristy, Singapore, 2001.

[93] Sahughet, A., and F. Azauant. "Building Intelligent Web Applications using Lightweight." Data and Knowledge Engineering 36.3 (2001): 283-316.

[94] Sahuguet, A. and F. Azavant. "Web Ecology-Recycling HTML pages as XML documents using W4F." ACM International Worksho on the Web and Databases (WebDB'99), Philadelphia, Pennsylvania, USA, June, 1999.

[95] Sahuguet, A. and F. Azavant. "Building Intelligent Web Applications Using Lightweight Wrapper" July 2000, http://db.cis.upenn.edu/DL/dke.pdf (Aug. 2002).

[96] Shaker,S., & M. Gembicki. The War-Room Guide To Competitive Intelligence. New York: McGraw Hill, 1999.

[97] Shanmugasundaram, J., J. Kiernan, E. Shekita, C. Fan, and J. Funderburk. "Querying XML views of relational data." Proceedings of VLDB 2001 Rome, Italy, Sept. 2001.

[98] Shanmugasundaram, J., E.J. Shekita, R. Barr, M.J. Carey, B.G. Lindsay, H. Pirahesh and B. Reinwald. "Effciciently publishing relational data as XML documents." The VLDB Journal (2000): 65-76.

[99] The World Wide Web Consortium (W3C)´s DOM (document object model) webpage, 1999. http://www.w3.org/DOM/.

[100] Unified Modeling Language, http://www.uml.org/

[101] Christophides V., S. Abiteboul, S. Cluet, and M. Scholl. "From structured documents to novel query facilities." Proceedings of the ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, 1994.

[102] Buganholo, V.P., S.B. Davidson., and C.A. Heuser. "Propaging XML View Updates to Relational Database."

[103] Labio, W.J., Z.Yue, J.L. Wiener, H.G. Molina, and J.Widow. "The WHIPS prototype for data warehouse creation and maintenance." Proceedings ACM SIGMOD International Conference on Management of Data, Tucson,

Arizona: ACM Press (1997): 557-559.

[104] Ng, W.K, E.P. Lim, C.T. Huang, S.S. Bhowmick, and F.Q. Qin. "Web warehousing: An Algebra for Web Information." Advances in Digital Libraries, (1998): 228-237

[105] Wide Web Consortium (W3C): http://www.w3c.org.

[106] Arens, Y., C.A. Knoblock, and W. –M. Shen. "Query Reformulation for Dynamic Information." Journal of Intelligent Information Integration. 6.2-3(1996): 99-130

[107] Yahoo!. http://www.yahoo.com

[108] Yang. Y., and X Liu. "A Re-Examination of Text Categorization Methods." Proceedings ACM SIGIR, 1999.

Acknowledgements

First of all, I would like to express full appreciation to my thesis advisor, Georg Gottlob, without whom this work would not have been possible and who always had a word of encouragement. It has been a pleasure to work with him and with the staff of the Database and Artificial Intelligence Department.

Thanks to the secretaries and members of the DBAI at the TU Wien for providing a competitive working environment.

I am grateful to Robert Baumgartner, Markus Herzog, Adrian Brooks, and Nysret Muslia. Finally I’d like to further acknowledge Dr. Werner Schimanovich for his friendship as well as all others from whom I learned and with whom I exchanged knowledge and ideas.

José Aldo Díaz Prado

Resume

Personal Data:

Full Name: José Aldo Díaz Prado.
Date of Birth: April 06, 1966, in Toluca, México
Parents: Lawyer. Agapito Díaz and Prof. Maria Guadalupe Prado
Sister: Elda Luisa Díaz Prado
Brothers: Fernado Edgar, Samuel Mauricio, Vidal A. Díaz-Prado
Nationality: Mexican
Address: Forsthausgasse 2-8/2136, A-1200 Wien, Austria
Emails**:** jadiaz@tuwien.ac.at ; jadiaz@itesm.mx

Education:

1972-1980	Primary School "Justo Sierra", Toluca, Mexico
1980-1983	Middle School "Miguel Hidalgo", Toluca, Mexico
1983-1086	High School " Universidad del Valle de Toluca", Toluca, Mexico
1986-1991	Industrial Engineering. ITESM; Monterrey, Mexico
1994-1997	Master's in Information Systems, ITESM, Monterrey, Mexico

Professional Experiences:

01/93- 2004 ITESM CAMPUS MONTERREY
Instituto Tecnológico y de Estudios Superiores de Monterrey
Technology Transfer & Consulting Services Coordinator

11/97 – 04/00 VITRO CORPORATION
Assimilation and Technolgy Transfer Manager.
Monterrey, Mexico

01/90 – 12/92 SPECTRUM CONSULTORES.
Associate Consultant.

04/92 – 12/92 GRUPO LUGA S.A de C.V.
Logistic Manager

Publications:

1. Web Knowledge Extraction for Visual Business Intelligence Approach using Lixto, International Conference on Knowledge Management (ICKM2005), Charlotte, North Carolina

2. A knowledge-based entrepreneurial approach for business intelligence in strategic technologies: Bio-MEMS, Proceedings of the Eleventh Ameritas Conference on Information Systems, Omaha, NE, USA August 11-14, 2005.

3. Business Intelligence Process through Information Extraction with Lixto. EUROCAST, Spain 2005,Febrero 2005.

4. La reinvención de servicios gubernamentales a través de e-goverment. Transferencia Enero 2002.

5. Grupos de Estrategia Tecnológica, Transferencia 1999.

6. Centro de Inteligencia Artificial Incursiona en desarrollo de e-books, Transferencia Julio 2001

7. Centro de Inteligencia Artificial desarrolla aplicaciones para la industria televisora, Revista Transferencia1999.

8. Evolucionar hacia los negocios de alta tensión, Revista Transferencia 1999.

Projects and Scientific Activities

- Knowledge Mangement Roll-out (Cemex – Philippines, Thailand, Bangladesh and Indonesia)
- Technology RoadMap: MEMS Technologies(Mexican Secretary of Economy and The Mexico-USA Sciences Foundation).
- MEMS Lab (Micro Electro Mechanical Systems).
- Vehicles Automation for Mining (Minera Peñoles; Torreón)
- Intelligent Tutors for Training (IBM de México, Guadalajara)
- Rocket Simulator (Mexican Army, Mexico City)
- Olympic Games Portal, Sydney 2000 (TV Azteca, Mexico City)
- Binary Databases (TV Azteca, Mexico City)
- Rating Measurement for Radio Stations and Broadcast (TV Azteca, Mexico City)
- Maintenances Expert System (Ford Motor Company, Chihuahua, Mexico).
- Technology RoadMap, Technological trends in the TV Industry. TV Azteca, México.
- Intelligent Distribution Center (Cemex – Monterrey, Mexico)
- Peoplemeters Development (radio frequency sonda and audio sonda).