# Web Data Extraction for Business Intelligence: the Lixto Approach

Robert Baumgartner [1] , Oliver Frölich [1] , Georg Gottlob [1] , Patrick Harz [2] ,

Marcus Herzog [1] and Peter Lehmann [2]

[1] DBAI
TU Wien
Favoritenstr. 9
1040 Vienna, Austria
{gottlob, froelich, baumgart, herzog}
@dbai.tuwien.ac.at

[2] Department of Information and
Communication
Hochschule der Medien, FH Stuttgart
Wolframstr. 32
70191 Stuttgart, Germany
{lehmann, harz}@hdm-stuttgart.de

**Abstract:** Knowledge about market developments and competitor activities on the market becomes more and more a critical success factor for enterprises. The World Wide Web provides public domain information which can be retrieved for example from Web sites or online shops. The extraction from semi-structured information sources is mostly done manually and is therefore very time consuming. This paper describes how public information can be extracted automatically from Web sites, transformed into structured data formats, and used for data analysis in Business Intelligence systems.

## 1 Introduction

### 1.1 Motivation

Companies from all branches and sizes are forced nowadays to make operative decisions within days or even hours – just 25 years ago, similar decisions took weeks or months [Ti95]. Thus, business management is interested in increasing the internal data retrieval speed. At the same time, the external data sources considered should be broadened to improve information quality. This fast, high-quality data is also needed to satisfy increasing investor demands for transparency, and to satisfy today’s better informed customers. Fortunately, new technologies like Business Intelligence systems and the internet are available to supply this data. Furthermore, the growing relevance of the internet in developed and developing countries creates new and dynamic sales channels and business opportunities.

Based on the described competitive pressure, a systematic observation of competitor activities becomes a critical success factor for business to:

- early identify chances in the market,
- anticipate competitor activities,
- recognize new and potential competitors,
- learn from errors and success stories of competitors, and
- validate and enhance own strategic goals, processes and products.

This process of collecting and analyzing information about competitors on the market is called "competitive intelligence" or "competitive analysis" [Ka98; SCIP04]. Nowadays, a lot of basic information about competitors can be retrieved legally from public information sources (public domain information [Ka98, p. 59]), such as Web sites, annual reports, press releases or public data bases.

In this paper we describe a solution how data from public information sources, in particular from the World Wide Web, can be retrieved and normalized automatically. We also illustrate how this data can be automatically integrated afterwards in an (often complex) Business Intelligence environment.

This paper is structured as follows: chapter 1 describes the problem situation and explains the Business Intelligence process. The architecture of the Lixto Software, a toolset for semi-structured data extraction and transformation, is introduced in chapter 2. In chapter 3, we take a closer look at two business scenarios applying the technologies introduced in the preceding chapters. Related work is discussed in chapter 4. Finally, we conclude with a short summary in chapter 5.

### 1.2 About Business Intelligence

Over the last 10 years, the term "Business Intelligence" (BI) has developed from an ambiguously used buzzword to a well-defined, real market. Also, the term BI is often used as a method box for collecting, representing and analyzing enterprise data to support decision makers. Taking a closer look at the word "intelligence", synonyms such as "knowledge, message, reconnaissance, clarification" can be found in a dictionary.[1] Thus, in the further course of this paper, "Business Intelligence" will be understood as a process providing better insight in a company and its chains of actions.

The Business Intelligence process covers three main process steps: *data integration*, *data storage* and *data usage* (see fig.1).

[1] See for example http://dict.leo.org.

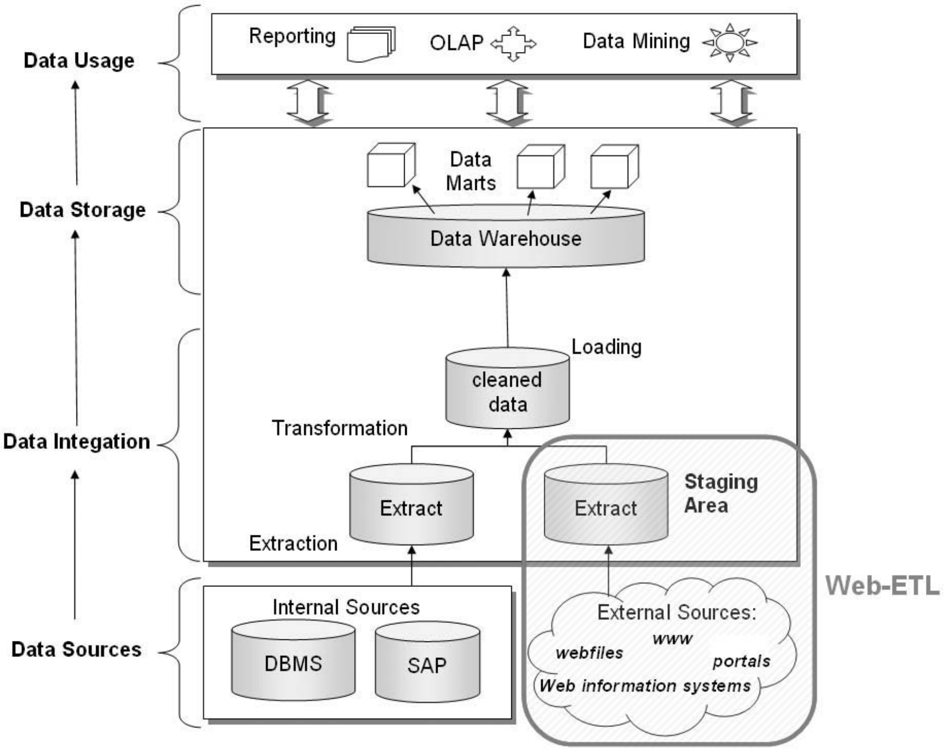


Figure 1: The Business Intelligence reference process.

*Data integration* covers methods to extract data from internal or external data sources such as ERP systems[2] or database systems. The data is transferred into a processing area allowing further data transformations like data "cleaning" and data normalization. A load process contains a scheduler which regularly (e.g. daily, weekly, or monthly) uploads the processed data into the final data base storage, the data warehouse.

*Data storage* in a data warehouse: the basic idea of a data warehouse is to store the relevant data for decision makers in a dedicated, homogeneous database. An important characteristic of the data warehouse is the integration of heterogeneous, distributed, internal and external data. This covers the physical storage of data in a single, centralized data pool, and it also covers the subject-oriented clustering of data organized by business processes, such as sales, production, or finance. The subject oriented organisation of data is called a data mart.

*Data usage*: to support decision making, data in a data warehouse has to be well-organized to fulfil different end-user requirements: predefined reporting for occasional users, ad-hoc data analysis for knowledge workers, or data mining for data analysts.

[2] ERP stands for Enterprise Resource Planning. An example of an ERP system is SAP R/3.

### 1.3 Problem Definition

On the one hand, powerful tools for Extracting, Transforming and Loading (ETL-tools) data from source systems into a data warehouse are available today. They support the data extraction from *internal* applications in an efficient way. On the other hand, there is a growing need to integrate also *external* data, such as market information, into these systems. The World Wide Web, the largest database on earth, holds a huge amount of relevant information. Unfortunately, this data exists in formats primarily intended for human users and cannot be easily processed by computer programs. Advanced data extraction and information integration techniques are required to process Web data automatically. Increasing demand for such data leads to the question of how this information can be extracted, transformed to a semantically useful structure, and integrated with a "Web-ETL" process into a Business Intelligence system. A solution proposition to this problem will be illustrated in chapter 2.

## 2 The Lixto Solution

The *Lixto Suite* software provides tools to access, extract, transform, and deliver information from various semi-structured sources like Web pages to various customer systems. The Lixto software is 100% based on Java technology and standards like XML schema, XSLT, SOAP and J2EE. Internally, the software uses the logic-based data extraction language ELOG [GK02]. In this chapter, we successively describe the process steps for creating and delivering structured data from semi-structured sources.

At first, so-called *wrappers* are generated. Dynamically and independently, these "intelligent"[3] software agents extract and translate all relevant information from HTML Web pages to a structured XML format that can be queried and processed by other programs. With Lixto, wrappers are generated in a graphical user interface with a few mouse clicks. Thus, no special programming knowledge is needed, and *wrappers* can be generated by non-technical personnel. Wrapper agents are typically generated by employees with the relevant business expertise for the project, e.g. from a company's marketing department.

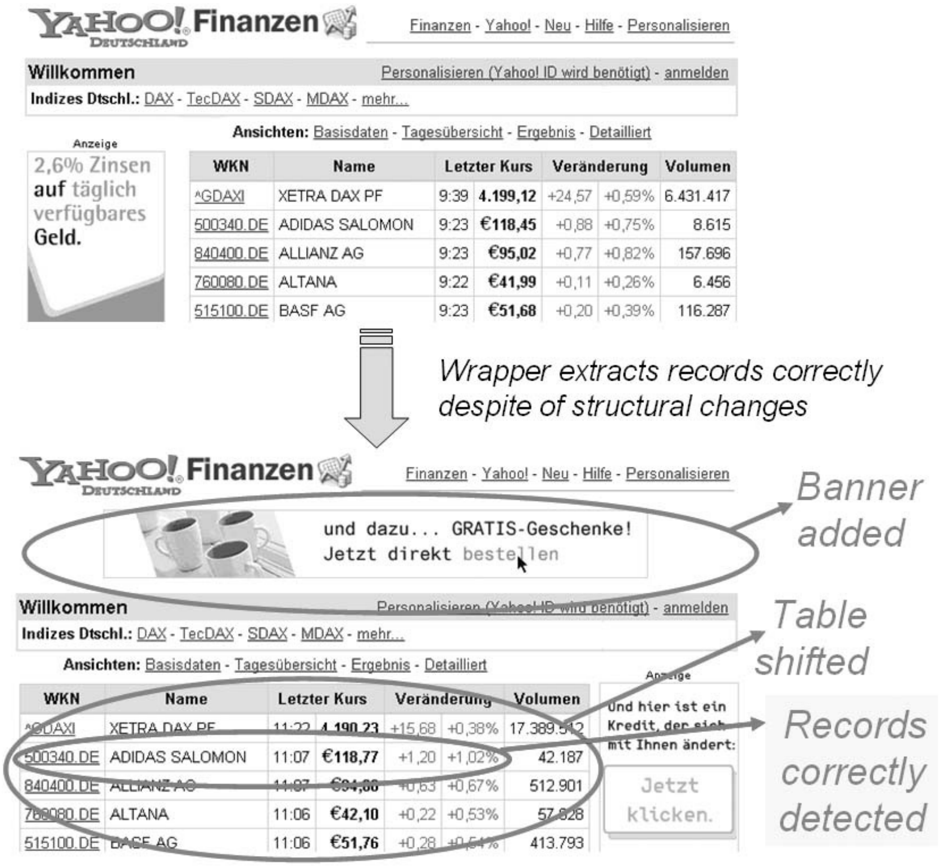
In a second step, XML data generated by wrappers is processed in the *Lixto Transformation Server* [GH01], the run-time environment for Lixto wrapper agents. A wrapper in the Transformation Server retrieves the Web data automatically, with no developer interaction, based on events. Events are for example a Web page content change, or a defined schedule, such as Web data retrieval *every hour* or *every 5 minutes*. Additionally, the Lixto Transformation Server can combine, transform and re-format data from *different* wrappers. The Transformation Server also supports the run-time administration and supervision of the whole process. For example, if a wrapper cannot extract data from a specific Web site because the Web server is down, the wrapper generates an error message for the administrator.

[3] See also chapter 2.1.

Finally, the Transformation Server delivers the extracted, aggregated information into the desired formats to other Business Intelligence systems such as SAP Business Information Warehouse or Microsoft Analysis Server. Also, the Transformation Server interactively communicates with these systems using various interfaces, such as special database formats, XML messaging, and Web services.

### 2.1 Extracting Data from External Sources

Creating a wrapper with Lixto starts by highlighting the relevant information with two mouse clicks in a standard internet browser window. Lixto then marks the data in a different colour. Conditions can be defined, allowing the program to identify the desired data even if the structure of the Web page slightly changes. Fig. 2 shows an example: the share prices from the companies listed in the German share index DAX are to be extracted from the Web site *finance.yahoo.de*.


YAHOO! Finanzen
DEUTSCHLAND
Finanzen - Yahoo! - Neu - Hilfe - Personalisieren
Willkommen
Personalisieren (Yahoo! ID wird benötigt) - anmelden
Indizes Dtschl.: DAX - TecDAX - SDAX - MDAX - mehr...
Ansichten: Basisdaten - Tagesübersicht - Ergebnis - Detailliert
Anzeige
2,6% Zinsen auf täglich verfügbares Geld.

| WKN | Name | Letzter Kurs | | Veränderung | | Volumen |
|---|---|---|---|---|---|---|
| ^GDAXI | XETRA DAX PF | 9:39 | 4.199,12 | +24,57 | +0,59% | 6.431.417 |
| 500340.DE | ADIDAS SALOMON | 9:23 | €118,45 | +0,88 | +0,75% | 8.615 |
| 840400.DE | ALLIANZ AG | 9:23 | €95,02 | +0,77 | +0,82% | 157.696 |
| 760080.DE | ALTANA | 9:22 | €41,99 | +0,11 | +0,26% | 6.456 |
| 515100.DE | BASF AG | 9:23 | €51,68 | +0,20 | +0,39% | 116.287 |

*Wrapper extracts records correctly despite of structural changes*

YAHOO! Finanzen
DEUTSCHLAND
Finanzen - Yahoo! - Neu - Hilfe - Personalisieren
und dazu... GRATIS-Geschenke!
Jetzt direkt bestellen
Willkommen
Personalisieren (Yahoo! ID wird benötigt) - anmelden
Indizes Dtschl.: DAX - TecDAX - SDAX - MDAX - mehr...
Ansichten: Basisdaten - Tagesübersicht - Ergebnis - Detailliert
Anzeige
Und hier ist ein Kredit, der sich mit Ihnen ändert:
Jetzt klicken.

| WKN | Name | Letzter Kurs | | Veränderung | | Volumen |
|---|---|---|---|---|---|---|
| ^GDAXI | XETRA DAX PF | 11:22 | 4.190,23 | +15,68 | +0,38% | 17.389.512 |
| 500340.DE | ADIDAS SALOMON | 11:07 | €118,77 | +1,20 | +1,02% | 42.187 |
| 840400.DE | ALLIANZ AG | 11:07 | [illegible] | +0,63 | +0,67% | 512.901 |
| 760080.DE | ALTANA | 11:06 | €42,10 | +0,22 | +0,53% | 57.828 |
| 515100.DE | BASF AG | 11:06 | €51,76 | +0,28 | +0,54% | 413.793 |

Banner added

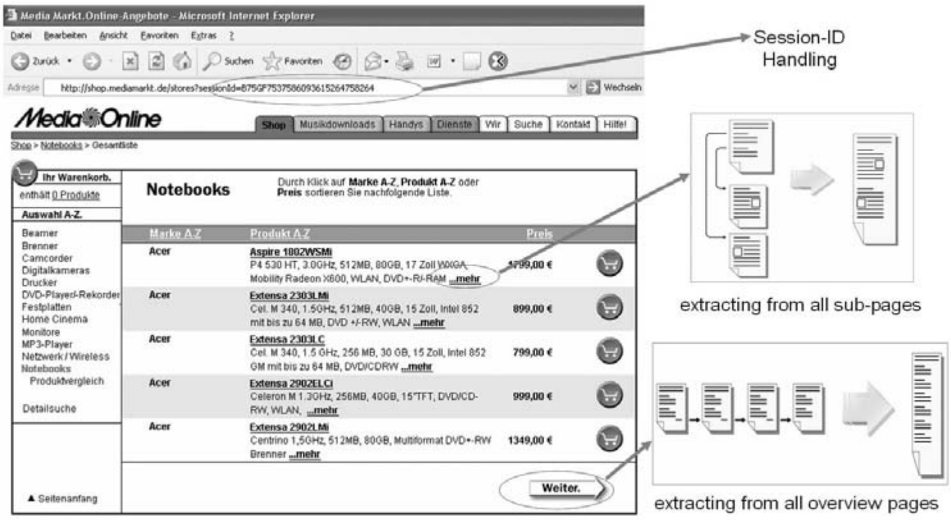
Table shifted

Records correctly detected

Figure 2: Wrapper robustness.

After a wrapper agent was successfully generated, the layout of the Web site changed after some weeks: the table with the quoted stocks moved from left to right, and additional banners were added. The existing wrapper still extracts all relevant information from the Web site. For a wrapper, an internet page is an HTML tree structure. A wrapper does *not* extract just the text from a specified HTML tree node, but uses "intelligent" conditions, so-called *logical patterns*. For the wrapper of fig. 2, such conditions could be *„the relevant area should contain the €symbol in each line"* or *"some specified company's names should occur"* (these names are stored in a system database). For the *logical pattern* comprised of the conditions, the software searches for the best match within the HTML tree using heuristic methods. So a very high robustness to changes within Web pages can be achieved for the wrapper agents.

Other capabilities of the Lixto software during the wrapper generation process are shown in fig. 3. Here, information about notebooks is to be extracted from the online shop *shop.mediamarkt.de*. Of special interest shall be the information about *manufacturer*, *model name*, *model price* and *model description*. On the overview page shown in fig. 3, only the first two lines of the *model description* are displayed. For each model name a linked sub-page with the whole description text exists. Furthermore there is a "next"-link ("weiter") leading to the next article overview page. The Lixto Software allows to record navigation sequences in a kind of macro recorder. During wrapper generation, only one sub-page needs to be accessed as an example and the "next"-link needs to be followed only once. The system then recognizes the similarly structured Web pages and extracts all complete model descriptions from all overview pages. The results are transformed to structured XML.



Notebooks

Durch Klick auf **Marke A-Z**, **Produkt A-Z** oder **Preis** sortieren Sie nachfolgende Liste.

| Marke A-Z | Produkt A-Z | Preis |
|---|---|---|
| Acer | **Aspire 1802WSMi**<br>P4 530 HT, 3.0GHz, 512MB, 80GB, 17 Zoll WXGA, Mobility Radeon X600, WLAN, DVD+-R/-RAM ...mehr | 1799,00 € |
| Acer | **Extensa 2303LMi**<br>Cel. M 340, 1.5GHz, 512MB, 40GB, 15 Zoll, Intel 852 mit bis zu 64 MB, DVD +/-RW, WLAN ...mehr | 899,00 € |
| Acer | **Extensa 2303LC**<br>Cel. M 340, 1.5 GHz, 256 MB, 30 GB, 15 Zoll, Intel 852 GM mit bis zu 64 MB, DVD/CDRW ...mehr | 799,00 € |
| Acer | **Extensa 2902ELCi**<br>Celeron M 1.30Hz, 256MB, 40GB, 15"TFT, DVD/CD-RW, WLAN, ...mehr | 999,00 € |
| Acer | **Extensa 2902LMi**<br>Centrino 1,5GHz, 512MB, 80GB, Multiformat DVD+-RW Brenner ...mehr | 1349,00 € |

Figure 3: Extraction of all article data.

In addition, the wrapper agents support dynamic handling of session IDs, automatic logon to password-protected pages, filling in form pages and processing the extraction from corresponding result pages (i.e. for Web interfaces of data bases) as well as automatic handling of cookies and SSL. Detailed information on further wrapping capabilities can be found in [BFG01].

### 2.2 The Transformation Server

Lixto wrapper agents are embedded in the runtime-environment of the Lixto Transformation Server. This server allows post processing the XML data generated by wrapper agents. Here data from different wrappers can be aggregated, re-formatted, transformed and delivered.

The whole process of modelling the workflow and dataflow is done in a graphical user interface in the Lixto Transformation Server. Graphical objects symbolize components, such as an *integrator* for the aggregation of data or a *deliverer* for the transmission of information to other software applications. By drawing connecting arrows between these objects, the flow of data and the workflow are graphically defined. A more detailed description of the components will be given in chapter 3.2.4 within the context of a business case example.

## 3 Application Business Cases

### 3.1 BI Scenario for a Consumer Electronics Online Shop

This example demonstrates the Web data integration process using Lixto for data extraction, transformation, and delivery to the data warehouse of the SAP Business Information Warehouse (SAP BW), a Business Intelligence system.

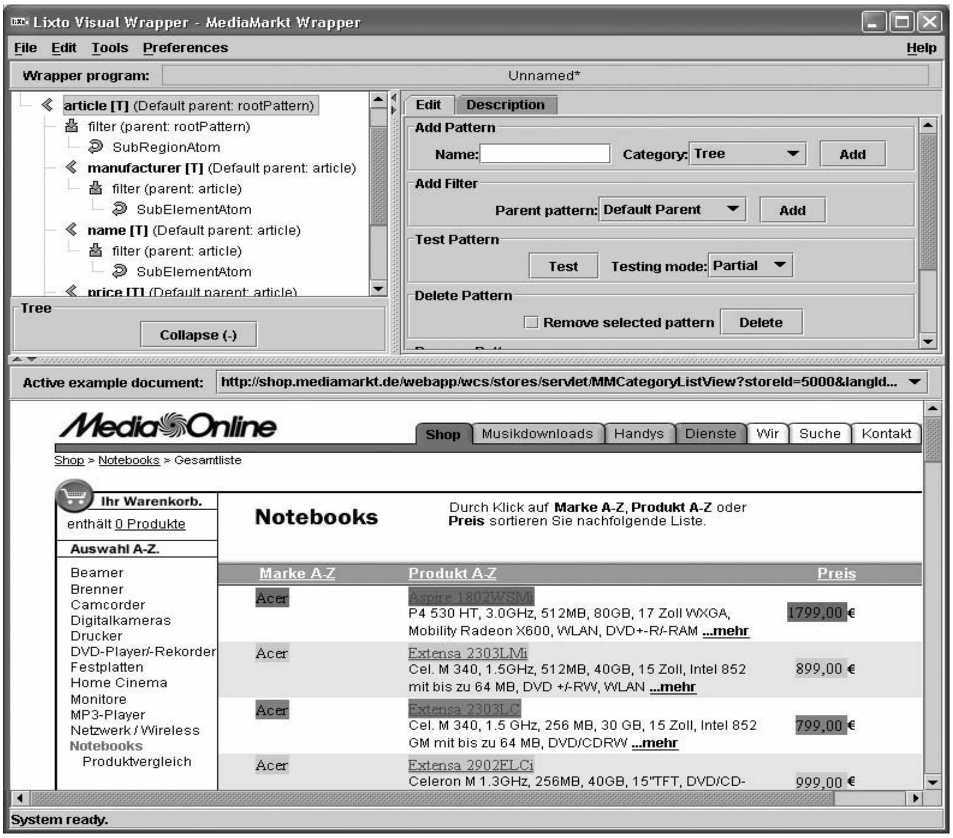
#### *3.1.1 Case Description*

A company sells consumer electronics, such as computers, digital cameras, TV sets and cellular phones. In an online shop, customers can order these goods. Many employees of the company spend many hours a day searching the Web to collect price information about their competitors from the vendor's Web sites. The price information retrieved is used for monthly price definitions. Product availability and regional price differences are also included in the data analysis.

In our scenario, price and product information from online shops is extracted from Web sites with Lixto. This includes data from competitor's Web sites to be able to compare their pricing with the pricing of our online shop. Data is then automatically transferred into SAP BW.

### 3.1.2 Data Extraction and Transformation with Lixto

As an example, fig. 4 shows how Web page data is selected for extraction using the Lixto software. In the lower window of fig. 4, a Web page from *shop.mediamarkt.de* has already been loaded in a Web browser, and the relevant information has been marked with two mouse clicks. The upper windows shows already defined *logical patterns,* such as *article* and *price*, arranged in a hierarchical structure. This structure corresponds to the XML output that will later be generated by the wrapper. After loading a Web page with relevant data into the Lixto software, at first a pattern named *article* is defined. This pattern later recognizes lines with article information. Within this line, other patterns are created, identifying information such as article *manufacturer* and article *price*. For this, structures of the HTML document are used, or regular expressions representing logical structures. For example, *price* can be defined as a number followed by a currency symbol (e.g. the euro symbol "€"). No programming is necessary because all selections are made visually in the browser window. With a "test"-button, all steps during wrapper generation can be evaluated immediately.



Wrapper program: Unnamed*

```
article [T] (Default parent: rootPattern)
  filter (parent: rootPattern)
    SubRegionAtom
  manufacturer [T] (Default parent: article)
    filter (parent: article)
      SubElementAtom
  name [T] (Default parent: article)
    filter (parent: article)
      SubElementAtom
  price [T] (Default parent: article)
```

Name:

Category: Tree

Parent pattern: Default Parent

Testing mode: Partial

□ Remove selected pattern

Active example document: http://shop.mediamarkt.de/webapp/wcs/stores/servlet/MMCategoryListView?storeId=5000&langId...

Shop > Notebooks > Gesamtliste

enthält 0 Produkte

Beamer
Brenner
Camcorder
Digitalkameras
Drucker
DVD-Player/-Rekorder
Festplatten
Home Cinema
Monitore
MP3-Player
Netzwerk / Wireless
Notebooks
Produktvergleich

Notebooks

Durch Klick auf **Marke A-Z**, **Produkt A-Z** oder **Preis** sortieren Sie nachfolgende Liste.

| Marke A-Z | Produkt A-Z | Preis |
|---|---|---|
| Acer | Aspire 1802WSMi<br>P4 530 HT, 3.0GHz, 512MB, 80GB, 17 Zoll WXGA, Mobility Radeon X600, WLAN, DVD+-R/-RAM **...mehr** | 1799,00 € |
| Acer | Extensa 2303LMi<br>Cel. M 340, 1.5GHz, 512MB, 40GB, 15 Zoll, Intel 852 mit bis zu 64 MB, DVD +/-RW, WLAN **...mehr** | 899,00 € |
| Acer | Extensa 2303LC<br>Cel. M 340, 1.5 GHz, 256 MB, 30 GB, 15 Zoll, Intel 852 GM mit bis zu 64 MB, DVD/CDRW **...mehr** | 799,00 € |
| Acer | Extensa 2902ELCi<br>Celeron M 1.3GHz, 256MB, 40GB, 15"TFT, DVD/CD- | 999,00 € |

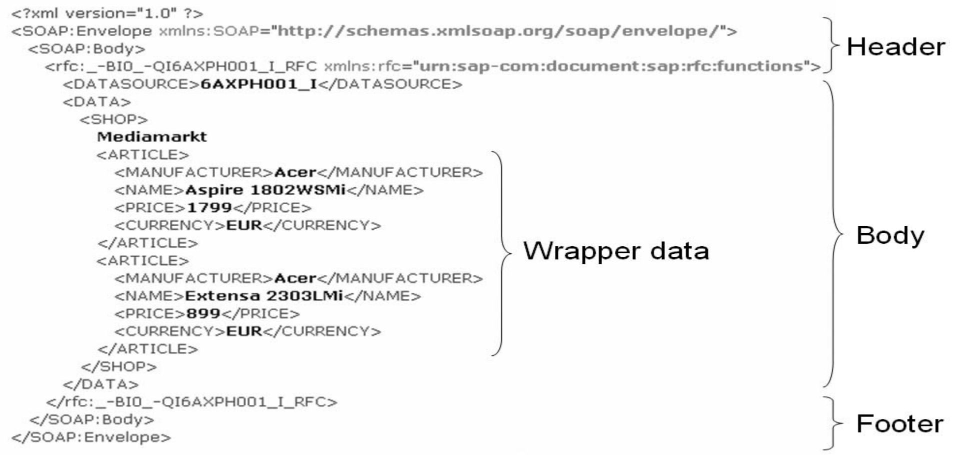
System ready.

Figure 4: Visual wrapper generation with Lixto.

SAP Business Information Warehouse supplies many interfaces for uploading data from different source systems into the data warehouse. For example, interfaces exist for:

- flat files with delimiter separated data or a fixed file format
- many relational database systems, such as Oracle, IBM DB2 or Microsoft SQL Server
- XML files processed by a SOAP interface (Simple Object Access Protocol).

The Lixto Transformation Server can use all these interfaces for data transmission. Due to the strict separation of content, logic and presentation, XML is becoming more and more important as a data source for Business Intelligence solutions [SB04]. In our business case, the resulting data from the wrappers is transformed and re-formatted in the Lixto Transformation Server in order to create SOAP messages based on XML.[4] These messages are used for loading the extracted data into the SAP BW. A SOAP message contains a header with meta information such as routing information or security parameters. The message body contains the XML data created by the wrappers. The following fig. 5 shows a typical SOAP message created by the Lixto Transformation Server.



```
<?xml version="1.0" ?>
<SOAP:Envelope xmlns:SOAP="http://schemas.xmlsoap.org/soap/envelope/">
  <SOAP:Body>
    <rfc:_-BI0_-QI6AXPH001_I_RFC xmlns:rfc="urn:sap-com:document:sap:rfc:functions">
      <DATASOURCE>6AXPH001_I</DATASOURCE>
      <DATA>
        <SHOP>
          Mediamarkt
          <ARTICLE>
            <MANUFACTURER>Acer</MANUFACTURER>
            <NAME>Aspire 1802WSMi</NAME>
            <PRICE>1799</PRICE>
            <CURRENCY>EUR</CURRENCY>
          </ARTICLE>
          <ARTICLE>
            <MANUFACTURER>Acer</MANUFACTURER>
            <NAME>Extensa 2303LMi</NAME>
            <PRICE>899</PRICE>
            <CURRENCY>EUR</CURRENCY>
          </ARTICLE>
        </SHOP>
      </DATA>
    </rfc:_-BI0_-QI6AXPH001_I_RFC>
  </SOAP:Body>
</SOAP:Envelope>
```

Figure 5: SOAP message with wrapper data.

### *3.1.3 Uploading XML in SAP BW*

XML data is uploaded in SAP BW using a SOAP interface [SAP01; HBE04]. The SOAP-/XML-interface is a Web service based on standard services like http for system communication over TCP/IP. Using SOAP, a source system can create a direct communication link and send data to a target system. The SOAP Web service acts as a receiver, which permanently waits for data messages. As soon as a data message arrives, the data structure is validated and the data package is stored in a temporary data queue.

[4] See also chapter 3.1.3.

The queue and the validation of the XML file structure is managed on a Web application server and processed by the SAP BW server engine driven by the data staging scheduler. The staging process extracts, transforms and loads the data from the data queue into data marts of the SAP BW server, triggered by user defined time slots.
An overview of the process from data extraction to BI integration is shown in fig. 6.

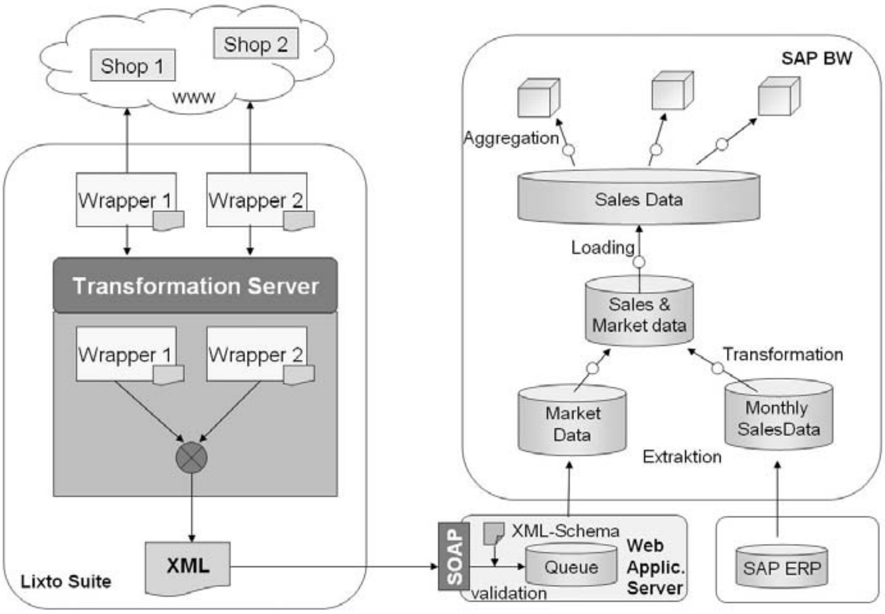
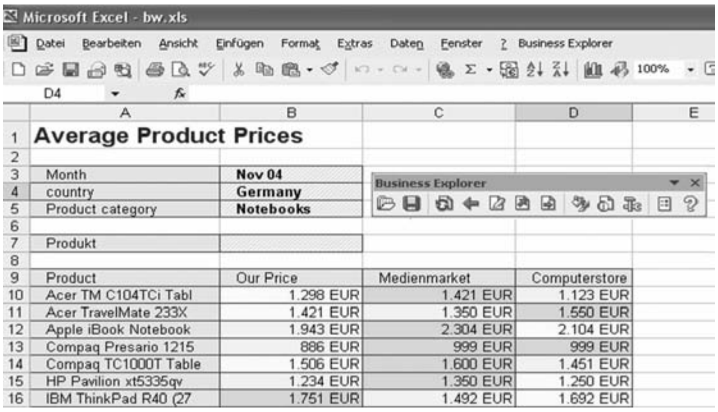


Figure 6: The data extraction and the BI integration process.

Once the data extracted by Lixto is sent and uploaded into the data warehouse by the described "Web-ETL" process, it can be accessed with an end-user tool which in our case is based on Microsoft Excel. Fig. 7 shows the result of a data query to SAP BW.



|  | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Average Product Prices |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 | Month | Nov 04 |  |  |  |
| 4 | country | Germany |  |  |  |
| 5 | Product category | Notebooks |  |  |  |
| 6 |  |  |  |  |  |
| 7 | Produkt |  |  |  |  |
| 8 |  |  |  |  |  |
| 9 | Product | Our Price | Medienmarket | Computerstore |  |
| 10 | Acer TM C104TCi Tabl | 1.298 EUR | 1.421 EUR | 1.123 EUR |  |
| 11 | Acer TravelMate 233X | 1.421 EUR | 1.350 EUR | 1.550 EUR |  |
| 12 | Apple iBook Notebook | 1.943 EUR | 2.304 EUR | 2.104 EUR |  |
| 13 | Compaq Presario 1215 | 886 EUR | 999 EUR | 999 EUR |  |
| 14 | Compaq TC1000T Table | 1.506 EUR | 1.600 EUR | 1.451 EUR |  |
| 15 | HP Pavilion xt5335qv | 1.234 EUR | 1.350 EUR | 1.250 EUR |  |
| 16 | IBM ThinkPad R40 (27 | 1.751 EUR | 1.492 EUR | 1.692 EUR |  |

Figure 7: End-user reporting within SAP BW.

## 3.2 Business Case: Pirelli

### *3.2.1 General Description of the Project*

Pirelli is one of the world market leaders in tire production, but also active in other sectors such as cables (energy and telecommunication cables). With headquarters in Milan/Italy, the company runs 21 factories all over the world and has more than thirty-five thousand employees.[5] On account of the growing amount and relevance of Web sites selling tires on the internet (both B2B and B2C), Pirelli analyzed the possibilities of monitoring retail and wholesale tire prices from competitors for their major markets. This external data should be automatically uploaded to their existing BI solution. After an extensive market research concerning available tools for Web data extraction and transformation, Pirelli selected the Lixto software because of its high scalability for back office use, its high robustness concerning data extraction quality and its straightforward administration.

In this business case, the Lixto Software was integrated in the Pirelli BI infrastructure in 2003 within a timeframe of two months. Tire pricing information of more than 50 brands and many dozens of tires selling Web sites are now constantly monitored with Lixto (Pirelli prices and competitor prices). The data is normalized in the Lixto Transformation Server and then delivered to an Oracle 9 database. From here, the Pirelli BI solution fetches the data and generates i.e. reports in PDF format and HTML format. These reports are automatically distributed to the Pirelli intranet for marketing and sales departments. An overview of the whole system structure is shown in fig. 8.

The success of the project can be measured by the more than 1.000 self-registered Pirelli users receiving the Lixto PDF reports regularly by email. Since its introduction, the Lixto reports are in the top 5 list of all most accessed files from the Pirelli intranet.

[5] See http://uk.biz.yahoo.com/p/p/peci.mi.html and [Pi03].

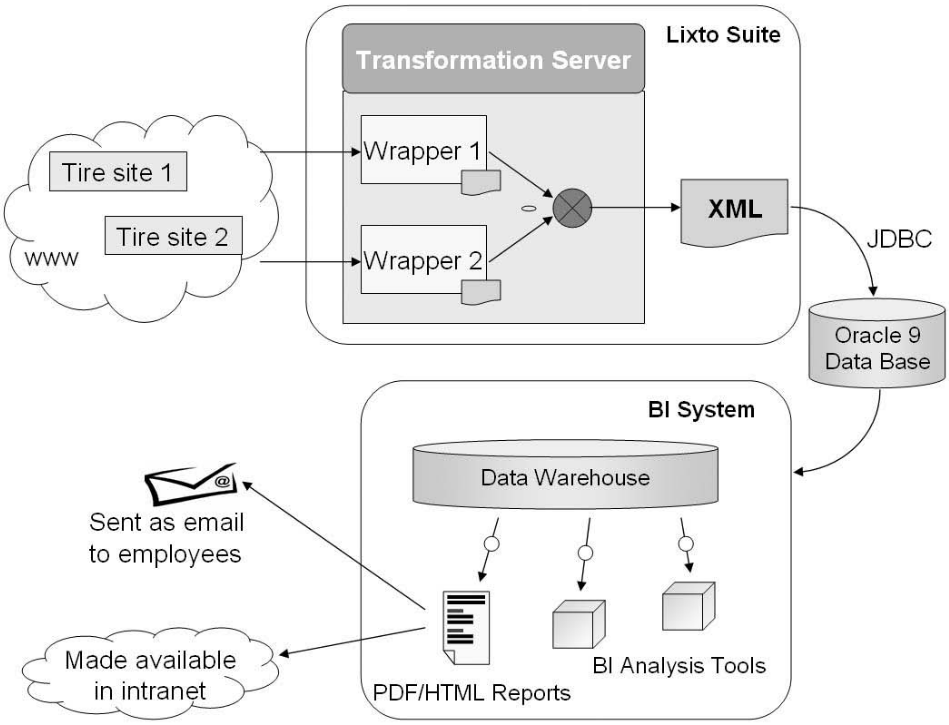
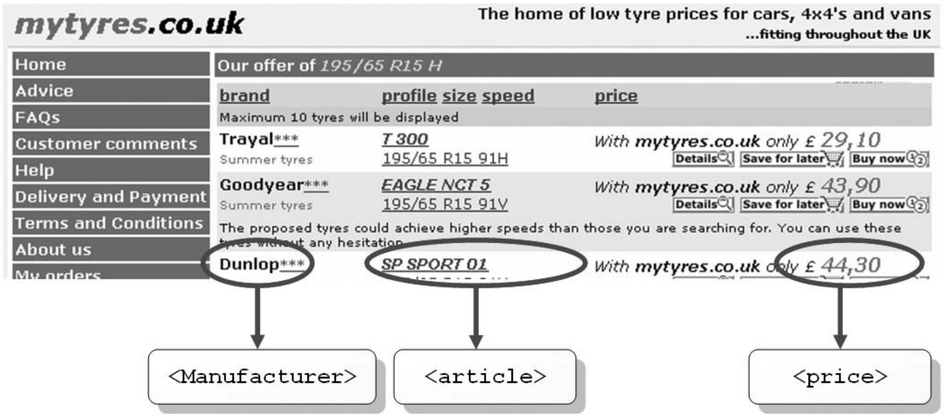


Figure 8: System structure overview.

### *3.2.3 Tire Data Extraction with Lixto*

The generation of the wrapper agents to extract data in online pricelists from tire selling Web sites is conducted analogous to the procedure described in chapter 3.1.2. In addition, many of these Web sites require logging in to the site (authentication), and then filling out request forms (what tires are of interested) before the result page with the information needed is displayed and can be extracted. The extracted data can even be inserted *iteratively* in other forms to extract more detailed data from the corresponding result pages. As depicted in chapter 2.1 and [GH01] all described processes are completely supported by Lixto. A typical tire Web page containing relevant data is shown in fig. 9.



mytyres.co.uk

The home of low tyre prices for cars, 4x4's and vans
...fitting throughout the UK

Home
Advice
FAQs
Customer comments
Help
Delivery and Payment
Terms and Conditions
About us
My orders

Our offer of 195/65 R15 H

| brand | profile size speed | price |
|---|---|---|
| Maximum 10 tyres will be displayed | | |
| Trayal***<br>Summer tyres | T 300<br>195/65 R15 91H | With mytyres.co.uk only £ 29,10<br>Details Save for later Buy now |
| Goodyear***<br>Summer tyres | EAGLE NCT 5<br>195/65 R15 91V | With mytyres.co.uk only £ 43,90<br>Details Save for later Buy now |
| The proposed tyres could achieve higher speeds than those you are searching for. You can use these tyres without any hesitation. | | |
| Dunlop*** | SP SPORT 01 | With mytyres.co.uk only £ 44,30 |

Figure 9: Tire data extraction.

### *3.2.4 Service Generation with the Lixto Transformation Server*

In the Lixto Transformation Server, a new *service* named *PirelliTireMonitor* is created. For this service, components are defined. Every component has a defined input and output behaviour. Configuration data for the components is generated by the system and automatically saved in XSLT stylesheets inside the *PirelliTireMonitor* service.

Components are graphically connected by arrows. Every arrow represents a flow of XML data. A simplified data flow as it is created in the Transformation Server is shown in fig. 10.

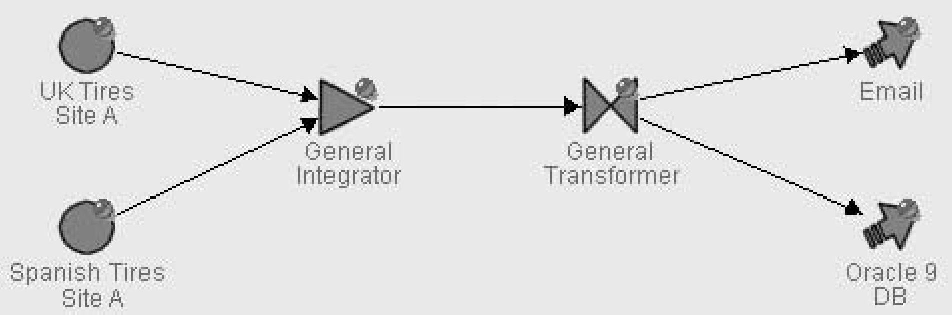


Figure 10: Modelling the data flow in the Transformation Server.

In the following, the further steps during the process of service configuration are described.

**a. Embedding wrappers in the service:** At first, the generated wrappers are loaded into the *PirelliTireMonitor* service. They are represented as circular objects (see fig. 10). In the Transformation Server additional wrapper attributes can be set, such as how often a Web site will be queried by a wrapper, for example every 5 minutes or every day.

**b. Integrator:** In the next step, an *Integrator* component is defined ("General Integrator" in fig. 10). An integrator allows XML data with different structures from different wrappers to be brought together in a uniform XML format. This is done by drawing arrows from the wrappers to the integrator, and by connecting XML elements by using graphical dialogs. Additionally, content adoptions can be made here, such as converting all *prices* from all currencies to euros, or subtracting local VAT and other taxes to get generally comparable prices.

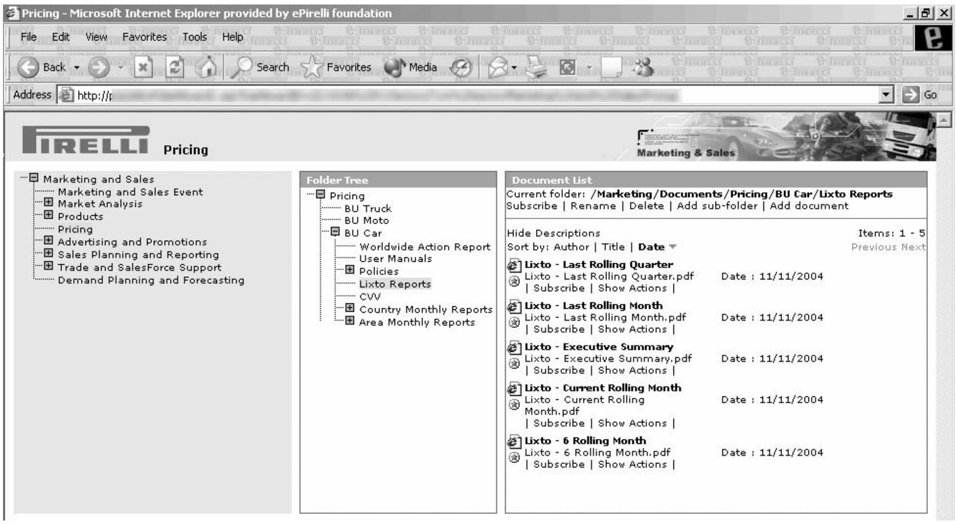
**c. Transformer:** The *Integrator* transfers the normalized data to a *Transformer* component ("General Transformer" in fig. 10). It can further restructure the XML data and combine it with other data, e.g. from internal databases. The most important job of a transformer component is to filter incoming data by defining queries, e.g. selecting only the tires from France and Spain, or removing double data entries.

**d. Deliverer:** After filtering the relevant data, it is passed on to a *Deliverer* component ("Email" and "Oracle 9 DB" in fig. 10). This component reformats the information for delivery in the desired output format. Here, data is converted to a valid data stream for transmission via JDBC and SQL store procedures to an Oracle 9 database. If an error occurs during extraction, e.g. if a Web Site is inaccessible, an email notification is sent to the administrator to allow quick response. Furthermore, the error is logged in the internal logs and reports of the Lixto Transformation Server.

After activating the *PirelliTireMonitor* service in the Transformation Server, the Oracle 9 database is incessantly supplied with new data.

### *3.2.5 Loading and Processing the Data in Pirelli's BI System*

Data in the Oracle 9 data base extracted by the Lixto software is loaded into the BI data warehouse following a predefined schedule. Once integrated in the BI software, the data can be analyzed with integrated analysis tools. For example, Pirelli prices are automatically compared with competitor prices, and regional sales and marketing employees can define conditions triggering an alert – if all winter tires from competitors are sold out in Southern Austria due to unexpected heavy snowfall, Pirelli can slightly increase their own prices in this region. Furthermore, the BI system creates different kinds of PDF reports and HTML reports based on the Lixto data, e.g. quarterly reports, monthly reports and executive summaries. These reports are available on the Pirelli intranet. In addition, Pirelli employees can self-subscribe on the intranet to different kinds of Lixto newsletters, and then the corresponding reports will be emailed to them. A screenshot from the Pirelli intranet showing the Lixto reports is illustrated in fig. 11.



⊟ Marketing and Sales
Marketing and Sales Event
⊞ Market Analysis
⊞ Products
Pricing
⊞ Advertising and Promotions
⊞ Sales Planning and Reporting
⊞ Trade and SalesForce Support
Demand Planning and Forecasting

Folder Tree
⊟ Pricing
BU Truck
BU Moto
⊟ BU Car
Worldwide Action Report
User Manuals
⊞ Policies
Lixto Reports
CVV
⊞ Country Monthly Reports
⊞ Area Monthly Reports

Document List
Current folder: **/Marketing/Documents/Pricing/BU Car/Lixto Reports**
Subscribe | Rename | Delete | Add sub-folder | Add document

Hide Descriptions Items: 1 - 5
Sort by: Author | Title | **Date** ▾ Previous Next

**Lixto - Last Rolling Quarter**
Lixto - Last Rolling Quarter.pdf Date : 11/11/2004
| Subscribe | Show Actions |

**Lixto - Last Rolling Month**
Lixto - Last Rolling Month.pdf Date : 11/11/2004
| Subscribe | Show Actions |

**Lixto - Executive Summary**
Lixto - Executive Summary.pdf Date : 11/11/2004
| Subscribe | Show Actions |

**Lixto - Current Rolling Month**
Lixto - Current Rolling Month.pdf Date : 11/11/2004
| Subscribe | Show Actions |

**Lixto - 6 Rolling Month**
Lixto - 6 Rolling Month.pdf Date : 11/11/2004
| Subscribe | Show Actions |

Figure 11: The Lixto PDF reports on the Pirelli intranet.

## 4 Related Work

Many non-commercial and some commercial solutions exist in the area of *wrapper generation frameworks.* An extensive overview is given in [KT02] and [LRS02]. Some of these are *stand-alone wrapper programming languages,* such as Jedi [AFH98] or Florid [HLL99], and do not provide any visual support, hence they are difficult to use for non-technical personnel. Tools for interactive wrapper generation such as W4F [AS99] or Wiccap [LLN02] require in general either manual post processing or do not offer the browser-displayed document for labelling.

In the area of *XML-based data integration frameworks*, various solutions are available, and several papers describe theoretical research on schema matching (e.g., [BMR01]). There are, however, very few frameworks that are designed for Web data extraction *and* integration, transformation and syndication of Web data. Especially in the commercial domain, they do not offer such an integrated solution as the Lixto framework.

The integration of external data into a data warehouse or BI solution has received little coverage in literature. A scientific, semi-automatic approach based on the object model MIX is described in [ZB02]. Another semi-automated methodology for designing Web warehouses based on XML sources is discussed in [VBR03].

## 5 Summary and Outlook

In this paper we showed how data can be automatically extracted from semi-structured web sites to obtain competitor information for decision support. We have introduced the architecture of the Lixto software for processing web data in an efficient manner. The result of the process is a structured XML file that can be used by a Business Intelligence system. We described the data upload into a SAP Business Information Warehouse using the SOAP/XML interface, and the upload into another BI system via an Oracle 9 database.

We took a closer look at two business scenarios for a Consumer Electronics online shop and for the company Pirelli. General advantages of the integration of external data to BI systems can be summarized as follows:

- fast data integration to support immediate reaction to market needs and changes
- active alerts by BI reporting agents in order to detect major changes
- real-time picture of the market
- low training costs through intuitive graphical user interface
- tailor services to end-user needs through end-user personalization
- reduce time, cost and personnel efforts for manual information retrieval
- reduce data collection errors caused by manual data input
- more data sources can be considered in high granularity
- better data transparency und data quality

Data analysts are able to obtain knowledge about the market situation in nearly real-time. This leads to better pricing decisions, a better positioning of the company and its products on the market, and a faster reaction to competitive activities, such as product innovations, price dumping, or promotions.

In future we will concentrate on automating wrapper repairing technologies and also focus on more unstructured formats such as PDF and plain text, using domain ontologies to support rational data validation.

## Acknowledgements

The authors would like to thank Giacomo del Felice from Pirelli Pneumatici S.p.A. for his continuous and reliable project support.

## References

[AFH98] Aberer, K.; Fankhauser, P.; Huck, G.; Neuhold, E.: JEDI: Extracting and Synthesizing Information from the Web, in: Proc. of COOPIS, 1998, pp. 32–43.

[AS99] Azavant, F.; Sahuguet, A.: Building light-weight wrappers for legacy Web datasources using W4F, in: Proc. of VLDB, 1999, pp. 738–741.

[BFG01] Baumgartner, R.; Flesca, S.; Gottlob, G.: Visual web information extraction with Lixto. In: Proc. of VLDB, 2001, pp. 119–128.

[BMR01] Bernstein, P.A.; Madhavan, J.; Rahm, E.: Generic Schema Matching with Cupid, in: The VLDB Journal, 2001, pp. 49–58.

[GH01] Gottlob, G.; Herzog, M.: Infopipes: A Flexible Framework for M-Commerce Applications, in: Proc. of TES workshop at VLDB, 2001, pp. 175–186.

[GK02] Gottlob, G.; Koch, C.: Monadic datalog and the expressive power of languages for Web Information Extraction, in: Proc. of PODS, 2002, pp. 17–28. Full version: Journal of the ACM 51(1), 2004, pp. 74 – 113.

[HBE04] Hahne, M.; Burow, L.; Elvers, T.: XML-Datenimport in das Information Warehouse bei Bayer MaterialScience, in: Schelp, Winter, Robert (Hrsg.): Auf dem Weg zur Integration Factory, Heidelberg, 2004, pp. 231-251.

[HLL99] Himmeröder, R.; Lausen, G.; Ludäscher, B.; May, W.: A Unified Framework for Wrapping, Mediating and Restructuring Information from the Web, in: WWWCM. Sprg. LNCS 1727, 1999, pp. 307–320.

[Ka98] Kahaner, L.: Competitive Intelligence: How to Gather, Analyse Information to Move your Business to the Top. Touchstone, New York, 1998.

[KT02] Kuhlins, S.; Tredwell, R.: Toolkits for Generating Wrappers, in: Net.ObjectDays, 2002, pp. 184–198.

[LLN02] Li, F.; Liu, Z.; Ng, W. K.: Wiccap Data Model: Mapping Physical Websites to Logical Views, in: Proc. of the 21st International Conference on Conceptual Modelling, 2002, pp. 120–134.

[LRS02] Laender, A. H.; Ribeiro-Neto, B. A.; da Silva, A. S.; Teixeira, J. S: A brief survey of web data extraction tools, in: Sigmod Record 31/2, 2002, pp. 84–93.

[Pi03] Pirelli & C. SpA: Annual Report 2003. http://www.pirelli.com//investor_relation/bilanciocompl2003.pdf, accessed on 2004-09-28, p. 7.

[SCIP04] Society of Competitive Intelligence Professionals (SCIP): What is CI? http://www.scip.org/ci/index.asp, accessed on 2004-09-28.

[SAP01] SAP AG: How to send XML Data to BW, ASAP for BW Acceleration, 2001, http://www.sdn.sap.com/documents/a1-8-4/HowtoSendXMLDatatoBW.pdf, accessed on 2004-09-28.

[SB04] Schwalm, S.; Bange, C.: Einsatzpotentiale von XML in Business-Systemen; Wirtschaftsinformatik 46, 2004, pp. 5-14.

[Ti95] Tiemeyer, E.; Zsifkovitis, H.E.: Information als Führungsmittel: Executive Information Systems. Konzeption, Technologie, Produkte, Einführung; 1st edition; Munich, 1995, p. 95.

[VBR03] Vrdoljak, B., Banek, M., Rizzi, S.: Designing Web Warehouses from XML Schemas, in: Proc. of the 5th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2003), Springer-Verlag, Lecture Notes in Computer Science, 2003.

[VdV03] Van der Vlist, E.: XML Schema, O'Reilly, 2003.

[ZB02] Zhu,Y., Buchmann, A.: Evaluating and Selecting Web Sources as External Information Resources of a Data Warehouse, in: Proc. of the 3rd International Conference on Web Information Systems Engineering (WISE02), 2002..