The Lixto systems applications in Business Intelligence and Semantic Web

Robert Baumgartner [1] , Oliver Frölich [1] , and Georg Gottlob [2]

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

[2] Oxford University Computing
Laboratory
Wolfson Building
Parks Road
Oxford, OX1 3QD
United Kingdom
georg.gottlob@comlab.ox.ac.uk

Abstract: This paper shows how technologies for Web data extraction, syndication and integration allow for new applications and services in the Business Intelligence and the Semantic Web domain. First, we demonstrate how knowledge about market developments and competitor activities on the market can be extracted dynamically and automatically from semi-structured information sources on the Web. Then, we show how the data can be integrated in Business Intelligence Systems and how data can be classified, re-assigned and transformed with the aid of Semantic Web ontological domain knowledge. Existing Semantic Web and Business Intelligence applications and scenarios using our technology illustrate the whole process.

1 Introduction

1.1 Motivation

Data available on the Web is a crucial asset in the enterprise world today, such as for making decisions on product prices, collecting opinions and getting an overview in fast-changing markets. To make use of Web data, methodologies and software components for harvesting structured facts from the Web are needed. Semantic Web Ontologies can be populated with Web data and sophisticated rule systems can help to leverage data analysis in Business Intelligence scenarios to a new level. In this paper we address the advantages of Web data extraction for Business Intelligence scenarios. Additionally, we consider how Semantic Web technologies can provide helpful means in such a setting.

1.2 Competitive Intelligence and Business Intelligence

Today, the time available for making operative decisions in a business environment is decreasing: decisions must be made within days or even hours. Just two decades, similar decisions still took weeks or months [Ti95]. Therefore, business management is interested both in increasing the internal data retrieval speed and in broadening the external data sources considered to improve information quality. Fortunately, new technologies like the internet and Business Intelligence systems are available to supply this data. 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, and to 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” (CI) or “competitive analysis” [SCIP04]. Nowadays, much information about competitors can be retrieved legally from public domain information sources, such as Web sites, press releases or public data bases [Ka98]. 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.

CI can be seen as a part of “Business Intelligence” (BI). The term BI is often used as a method box for collecting, representing and analyzing enterprise data to support the decision-making process within a company’s management. More generally, BI can be understood as a process providing better insight in a company and its chains of actions.

Technically, the Business Intelligence process covers three main process steps: *data integration*, *data storage* and *data usage* (see fig.1). The most important step concerning our research is *Data integration,* which covers methods to extract data from internal or external data sources. Traditionally, the data is derived for example from database systems in a so-called ETL process (extract, transform, load). We propose using Integrated Wrapper Technologies [Fr06] for Web data extraction and integration in a process we call Web-ETL [BFG05]. This step will be more closely described in the further course of this paper and may also contain data transformations like data “cleaning” and data normalization. A load process with a scheduler regularly uploads (e.g. daily, weekly, or monthly) the processed data into the final data base storage of the BI system, the data warehouse. This *data storage* holds the relevant data for decision makers in a dedicated, homogeneous database. An important characteristic of the data warehouse is the physical storage of data in a single, centralized data pool. It also covers the subject-oriented clustering of data organized by business processes, such as sales, production, or finance. With the information being well-organized in the data warehouse, *Data usage* now can support decision making with predefined reporting for occasional users, ad-hoc data analysis for knowledge workers, or data mining for data analysts.

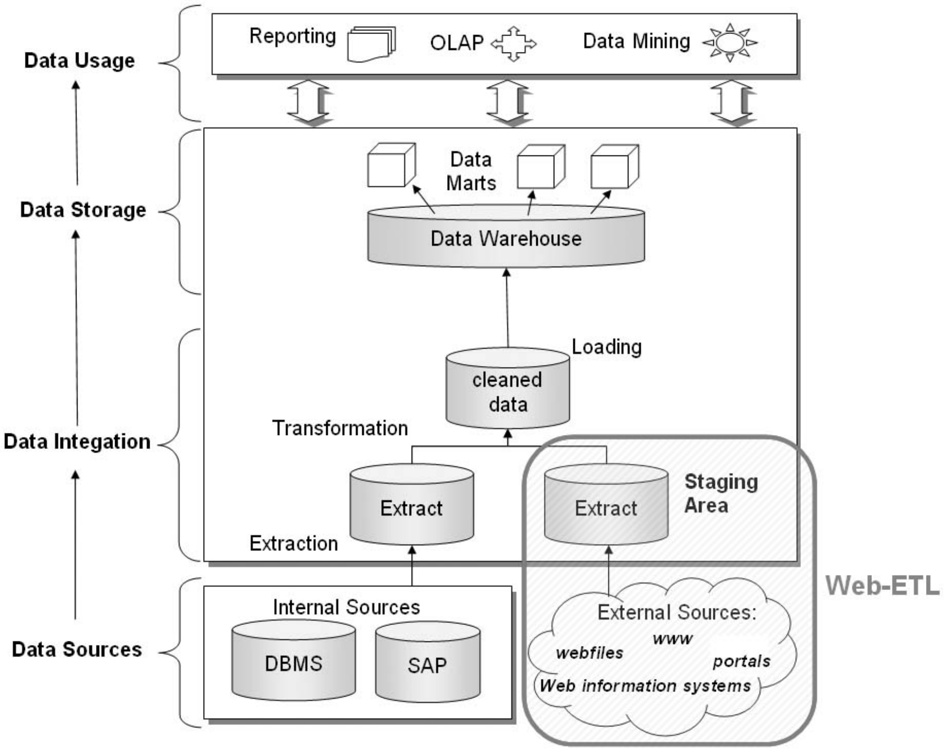


Figure 1: The Business Intelligence reference process [BFG05].

1.3 Semantic Web

Today, the realization of the Semantic Web idea of "an extension of the current web in which information is given in well-defined meaning, better enabling computers and people to work together" [BHL01] is technically still in an early stage: W3C recommendations exist for machine-readable semantics, appropriate markup and description languages, and sharable knowledge representation techniques, but the logical definition and technical implementation of the upper layers of the so-called Semantic Web tower [Be02], e.g. the rule and reasoning layer, or the layers of proof and trust, are still to be explored.

The semi-structured Web of today consists of billions of documents in different formats which are not query-able as a database and heavily mix layout, structure and the intended information to be presented. There is a huge gap between Web information and the well-structured data usually required in corporate IT systems or as envisioned by the Semantic Web.

Until this vision is realized, “translation components” between the Web and corporate IT systems that (semi-)automatically translate Web content (e.g. in HTML) into a structured format (e.g. XML) are necessary. Once transformed, data can be used by applications, stored into databases or populated into ontologies.

1.4 Integrated Wrapper Technologies

A Wrapper is a program that automatically accesses source data (e.g. from the Web in HTML) and then extracts and transforms the data into another format (e.g. XML). A number of classification taxonomies for wrapper development languages and environments have been introduced in various survey papers. A to our knowledge complete overview of the different approaches and systems is given in [Fr06].

Integrated Wrapper Technology (IWT) systems combine the capabilities of wrapping components with Information Integration (II) components [Fr06]. The latter generally transform the extracted data and integrate it in other (business) IT systems. IWT systems are suitable for implementing advanced information systems for the Semantic Web: Partially, they can bridge the gap between the Web existing today and the today not yet existing Semantic Web that might be used as the largest database on earth where data exists in machine-readable formats and can be integrated easily in other IT systems. IWT systems can extract data from semi-structured Web pages, transform it to a semantically useful structure, and integrate it e.g. with a Web ETL-process into a Business Intelligence system. A solution proposition to this problem will be illustrated in the next chapter.

2 The Lixto Solution

The *Lixto Suite* software is an IWT system which provides tools to access, extract, transform, and deliver information from various semi-structured sources like Web pages to many different customer systems. The Lixto software is based on Java technology and standards such as XML schema, XSLT, SOAP and J2EE. Technically, the main distinguishing criteria to many other approaches are that Lixto embeds the Mozilla browser and is based on Eclipse. This allows Lixto to be always at the cutting edge of Web browser technology and access and extract data from all Web pages even using the newest techniques like Web 2.0, e.g. Ajax. Internally, the software uses the logic-based data extraction language *Elog* [GK02].

The Lixto Suite is comprised of three products: The *Visual Developer* for Wrapper generation, the *Metasearch* for real-time Wrapper queries and the *Transformation Server* as runtime environment for scheduled Wrappers queries and as Information Integration component. In this chapter, we successively describe the process steps for creating and delivering structured data from semi-structured sources.

2.1 Wrapper Generation with Visual Developer

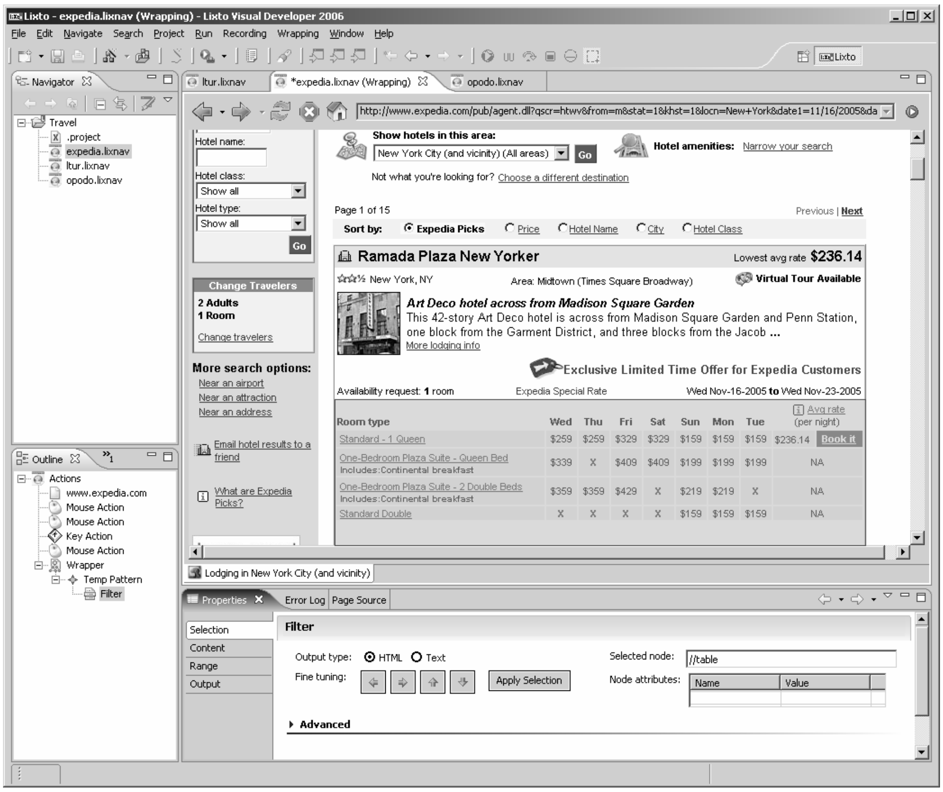
Wrappers generated with Visual Developer extract and translate all relevant information from HTML Web pages to a structured XML format. The underlying extraction language Elog is derived from datalog and is especially designed for wrapper generation. The Elog language operates on Web objects, which are (lists of) HTML elements, and strings. Elog rules can be specified visually in a graphical user interface by a few mouse clicks without knowing the Elog language. Thus, no special programming knowledge is needed, and wrappers can be generated by non-technical personnel, typically by employees with the relevant business expertise for the project, e.g. from a company's marketing department.

Creating a wrapper with the Visual Developer comprises two steps: First, the data model creation. In this phase an XML-schema based model in which extracted data is inserted either is imported (e.g. in case of news extraction typically RSS) or generated from scratch. In the second phase navigation and extraction steps are configured. Such extraction rules are semi-automatically and visually defined by a wrapper designer in an iterative process. Extracted data can subsequently populate an ontology with instance data. The whole wrapper generation process starts by generating a deep Web navigation sequence (such as navigation through forms) and subsequently highlighting the relevant information with two mouse clicks in the integrated standard Mozilla browser window. The software 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 where information is extracted from different trip booking websites like *expedia.com* and *opodo.com*.

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 Dollar-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 over time can be achieved for the wrapper agents.

In addition, navigation to further documents during the wrapping process is possible: For example, starting from a Google result list overview page, it is possible to extract defined information from all result sub-pages and, by clicking on the "next"-button on each result list overview page (this action has to be defined only once) to extract all relevant information from all overview pages and their sub pages. The Visual Developer is capable of action-based recording and replaying, i.e. recording actions of a user based on mouse clicks and key actions. This is highly advantageous compared to a plain request-based macro, as it allows the system to deal with dynamically changing HTML pages.

Visual Developer wrapper agents also support 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), dynamic handling of session IDs, and automatic handling of cookies and SSL. Detailed information on further wrapping capabilities can be found in [BFG01].



Ramada Plaza New Yorker — Lowest avg rate $236.14

☆☆½ New York, NY — Area: Midtown (Times Square Broadway) — Virtual Tour Available

Art Deco hotel across from Madison Square Garden
This 42-story Art Deco hotel is across from Madison Square Garden and Penn Station, one block from the Garment District, and three blocks from the Jacob ...
More lodging info

Exclusive Limited Time Offer for Expedia Customers

Availability request: 1 room — Expedia Special Rate — Wed Nov-16-2005 to Wed Nov-23-2005

Room type	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Avg rate (per night)	
Standard - 1 Queen	$259	$259	$329	$329	$159	$159	$159	$236.14	Book it
One-Bedroom Plaza Suite - Queen Bed Includes:Continental breakfast	$339	X	$409	$409	$199	$199	$199	NA	
One-Bedroom Plaza Suite - 2 Double Beds Includes:Continental breakfast	$359	$359	$429	X	$219	$219	X	NA	
Standard Double	X	X	X	X	$159	$159	$159	NA	

Figure 2: Visual Developer User Interface.

2.2 The Transformation Server

In a second step, XML data generated by wrappers is processed in the *Lixto Transformation Server* [GH01]. A wrapper in the run-time environment of the Transformation Server retrieves the Web data automatically, with no developer interaction, based on events. Events can be for example a defined schedule, such as Web data retrieval *every hour* or *every 5 minutes*. The usual setting for the creation of *services* based on Web wrappers is that information is obtained from multiple wrapped sources and has to be integrated. Often source sites have to be monitored for changes, and changed content has to be automatically extracted and processed.

The Lixto Transformation Server allows for Information Integration by combining, transforming and re-formatting data from different wrappers. Results of this process can be delivered in various formats and channels to other IT systems, e.g. Business Intelligence systems such as SAP Business Information Warehouse or Microsoft Analysis Server. Transformation Server can interactively communicate with these systems using various interfaces, such as special database formats, XML messaging, and Web services.

The whole process of modelling the workflow and dataflow is done in the graphical user interface of 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 (see also fig. 3). A more detailed description of the components is given in [BFG05].

3 Application Business Cases

The following business cases show and illustrate the capabilities of the IWT software *Lixto Suite* in real-world scenarios. The first scenario concentrates more on a BI solution, whereas the second scenario describes an application for the Semantic Web.

3.1 CI Solution for Pirelli

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.[1] 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.

The Lixto Software was integrated in the Pirelli BI infrastructure 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). A simplified screenshot of the Pirelli service shows the data flow as it is defined in the Lixto Transformation Server (see fig. 3).

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

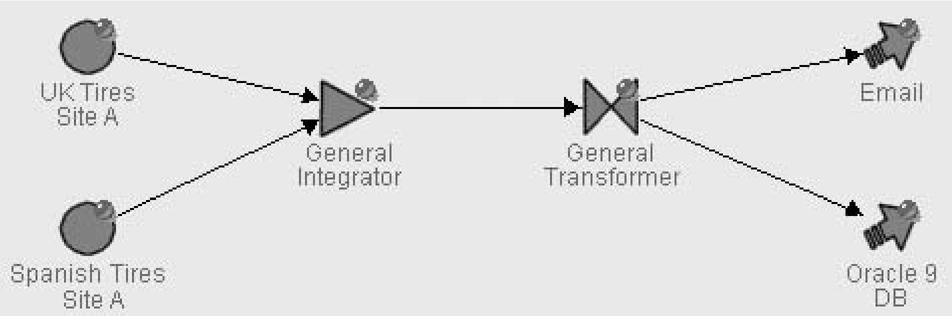
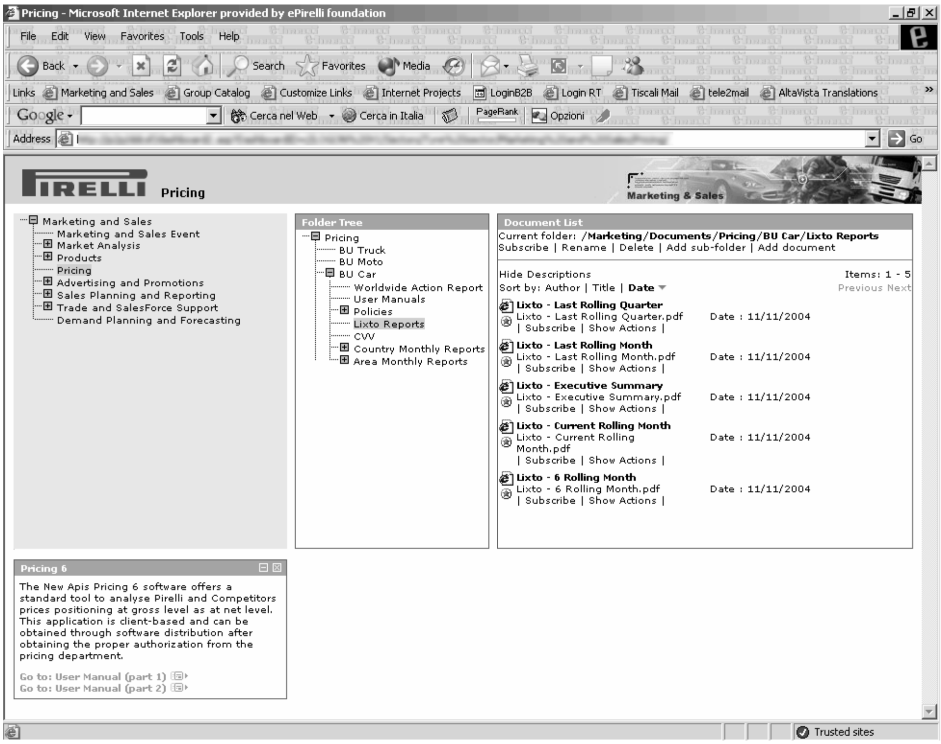


Figure 3: Modelling the data flow in the Transformation Server [BFG05].

The data is normalized in 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 (see fig. 4).



Pricing

- 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 |

Pricing 6

The New Apis Pricing 6 software offers a standard tool to analyse Pirelli and Competitors prices positioning at gross level as at net level. This application is client-based and can be obtained through software distribution after obtaining the proper authorization from the pricing department.

Go to: User Manual (part 1)
Go to: User Manual (part 2)

Figure 4: Lixto Reports within the Pirelli intranet [BFG05].

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. A more detailed description of the Pirelli BI integration project can be found in [BFG05].

3.1 The Personal Publication Reader

The Personal Publication Reader is an advanced personalized information systems using Semantic Web technologies [BHM05] that has been created by the Learning Lab Hannover. It provides a syndicated and personalized view on non-uniform, distributed Web data. The Personal Publication Reader has been designed and implemented in the context of the Network of Excellence "REWERSE – Reasoning on the Web" and syndicates and personalizes information about the REWERSE project structure, people, objectives, and information about research papers in the context of the project.

The Personal Reader Framework was used for the design and implementation of personalized Web Content Readers. Such readers have been realized for e-Learning and for publication browsing. Web services allow for the development of flexible information systems by encapsulating specific functionality and communication capabilities with other Web services. The aim of the Personal Reader Framework was to develop a toolset for designing, implementing and maintaining Personal Web Content Readers. These Personal Readers allow the user to browse information (the "Reader" part), and to access personal recommendations and contextual information on the currently viewed Web resource (the "Personal" part). The framework features a distributed open architecture designed to be easily extensible. It uses standards such as XML and RDF, and technologies like JSP and XML-based-RPC. The communication between all of these components is syntactically based on RDF descriptions, providing a high level of flexibility for the whole framework.

Content syndication and personalization is achieved by reasoning about ontological knowledge and extracted Web data. The Lixto Software is used for the implementation of the information provision part. The process of gathering Web data for the Personal Publication Reader using Lixto is described in the following.

In a first step, wrappers are created for the websites of the REWERSE project participant organizations. Then, these wrappers are loaded to a new service within the Lixto Transformation Server. Here the XML data derived from the various wrappers has to be combined, cleaned, and syndicated into the Framework's ontology. The output format of the Transformation Server is an RDF representation. This process is scheduled regularly and hands over the data to the Personal Publication Reader. Fig. 5 shows the data flow from the wrapper to the RDF output as it is defined in the Transformation Server.

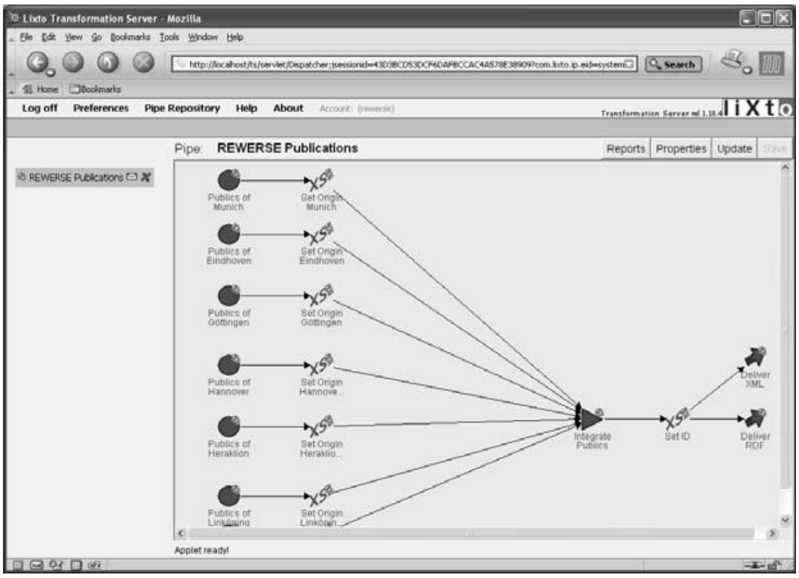
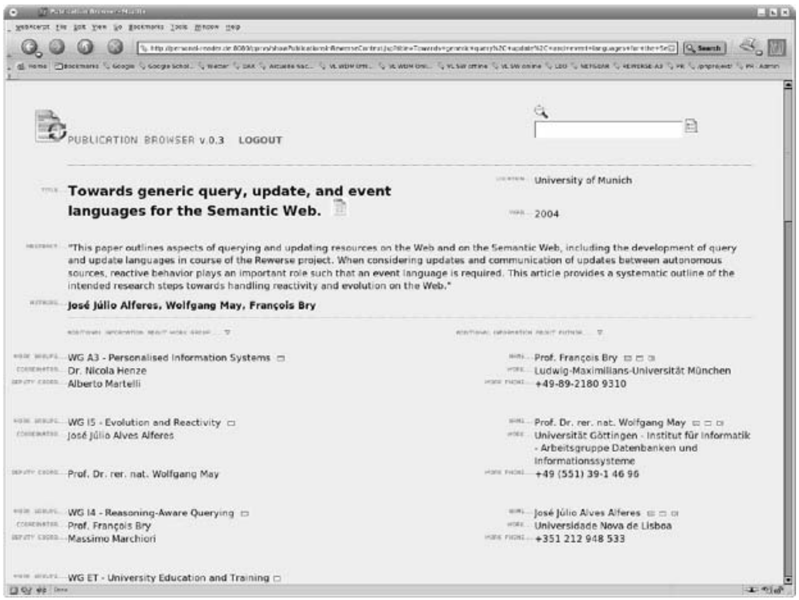


Figure 5: Personal Publication Reader data flow in the Transformation Server [BHM05].

In addition to the extracted information on research papers described above, a “REWERSE-Ontology” has been built in OWL extending the Sementic Web Research Community Ontology (SWRC) [SWRC01]. Here information about research members of the REWERSE project is kept.



Towards generic query, update, and event languages for the Semantic Web.

University of Munich

2004

"This paper outlines aspects of querying and updating resources on the Web and on the Semantic Web, including the development of query and update languages in course of the Rewerse project. When considering updates and communication of updates between autonomous sources, reactive behavior plays an important role such that an event language is required. This article provides a systematic outline of the intended research steps towards handling reactivity and evolution on the Web."

José Júlio Alferes, Wolfgang May, François Bry

WG A3 - Personalised Information Systems
Dr. Nicola Henze
Alberto Martelli

Prof. François Bry
Ludwig-Maximilians-Universität München
+49-89-2180 9310

WG I5 - Evolution and Reactivity
José Júlio Alves Alferes
Prof. Dr. rer. nat. Wolfgang May

Prof. Dr. rer. nat. Wolfgang May
Universität Göttingen - Institut für Informatik - Arbeitsgruppe Datenbanken und Informationssysteme
+49 (551) 39-1 46 96

WG I4 - Reasoning-Aware Querying
Prof. François Bry
Massimo Marchiori

José Júlio Alves Alferes
Universidade Nova de Lisboa
+351 212 948 533

WG ET - University Education and Training

Figure 6: Screenshot of the Personal Publication Reader [BHM05].

Finally, fig. 6 shows the syndicated view on publications in REWERSE together with the corresponding links, and information about the authors of the publications like homepage, phone number, etc. The Personal Publication Reader is available online via the URL *www.personal-reader.de*.

Acknowledgements

This research has been partially supported by REWERSE - Reasoning on the Web (rewerse.net), Network of Excellence, 6th European Framework Program.

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

References

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

[BFG05] Baumgartner, R.; Frölich, O.; Gottlob, G.; Harz, P.; Herzog, M.; Lehmann, P.: Web Data Extraction for Business Intelligence: the Lixto Approach, in: Datenbanksysteme in Business, Technologie und Web (BTW), LNI, Series of the Gesellschaft für Informatik, P-65 (2005), pp. 48–65.

[BHL01] Berners-Lee, T.; Hendler, J.; Lassila, O.: The semantic web. Scientific American, May 2001.

[Be02] Berners-Lee, T.: The semantic web - mit/lcs seminar, 2002.. http://www.w3c.org/2002/Talks/09-lcs-sweb-tbl/.

[BHM05] Baumgartner, R.; Henze, N.; Herzog, M.: The Personal Publication Reader: Illustrating Web Data Extraction, Personalization and Reasoning for the Semantic Web, in: European Semantic Web Conference ESWC 2005, LNCS 3532, Springer, Berlin Heidelberg, 2005, pp. 515–530.

[Fr06] Frölich, O.: Erstellung und Optimierung von Geschäftsprozessen durch Integrierte Wrapper-Technologien mit Anwendungsbeispielen aus den Branchen Mobile Services, Competitive Intelligence und dem Verlagswesen. Dissertation, DBAI, TU Vienna, Vienna, 2006.

[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.

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

[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.

[SWRC01] SWRC - Semantic Web Research Community Ontology, 2001. http://ontobroker.semanticweb.org/ontos/swrc.html.

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