

DATABASE

TRENDS AND APPLICATIONS

Solutions for the Information Project Team • www.dbta.com

Volume 20, Number 3 • March 2006

Data Extraction Automation: The Next Generation

The data you need is out there. WebQL 3.0 lets you gather it efficiently in a useful format.

Here is a typical problem. Each year thousands of bills are introduced to the Georgia General Assembly. As they make their way through the process, they change form and format.

Frequently multiple pieces of legislation would modify the same section of law in different, inconsistent ways.

Each day, the full text and status of the bill is posted to the General Assembly's Web site. Analysts from the Office of Planning and Budget must submit relevant portions to the Governor's Legislative Information System to analyze the impact on budget and existing state programs. But extracting the relevant data from the General Assembly Web site was tedious, time-consuming, and prone to error. The solution was WebQL from QL2 Software.

Automating Data Extraction

WebQL is an award winning Web mining and unstructured data management tool for developing and deploying software agents to extract information from the World Wide Web and unstructured data sources, and for presenting that data in a format appropriate for analysis. There is a wealth of information on the Web, and within internal corporate repositories that cannot be efficiently retrieved in a useable form.

To get the information, companies often turn to labor-intensive methods, either creating custom programs on their own, or manually searching and reformatting the information. WebQL automates the full process of gathering, pinpointing and outputting data. Using SQL-like constructs, it extracts information from structured, semi-structured and unstructured information sources and integrates that information into a useable format.

Typical Applications

WebQL can be effectively deployed in a wide number of competitive intelligence, business intelligence and knowledge management applications including:

• Deep Web price gathering

The explosion of e-business, and e-government has made competitive pricing information widely available on Web sites and government information portals. But, price lists are difficult to extract without selecting product categories or filling out Web forms and prices are often buried deep in documents. Automated form completion and downloading are needed to retrieve prices from the deep Web.

• Primary research

Message boards, blogs, and other Web forums provide a wealth of public opinion and user experience information on consumer products, air travel, test-drives, experimental drugs, and more. WebQL offers simultaneous board crawling, selective content extraction, task scheduling, and custom output reformatting.

• Content aggregation

Content is exploding and available from Web and non-Web sources. WebQL can crawl the Web, internal information sources, and subscription services to automatically populate portals.

• Supporting CRM systems

The Web is a valuable source of external data to selectively populate a data warehouse or a CRM database. Leading organizations are realizing the value of adding external data, too.

• Scientific research

Scientific information, such as a gene sequence data, is available on multiple Web sites and subscription services. WebQL can automate the location and extraction of this information and aggregate it into a single presentation format or portal.

• Business activity monitoring

WebQL can continuously monitor dynamically changing information sources to provide real-time alerts and to populate information portals and dashboards.

WebQL 3.0: The Next Generation

WebQL 3.0 represents the next generation of automated data extraction. All three components of WebQL--the WebQL Server, WebQL Studio and

WebQL Runtime--have been significantly enhanced. Using Optical Character Recognition, WebQL can now extract data embedded in images in .pdf files and Web pages. It can also reformat and resize through images.

Moreover, WebQL 3.0 extracts data from more file types than ever before including PowerPoint slides, RTF, DBF and self-extracting Zip files. Many file formats can be converted into text files that reflect the layout of the original document.

Finally, WebQL 3.0 has a number of new enterprise-grade features including a scheduling tool, real-time event monitoring, and new interfaces for retrieving WebQL output. Strong encryption adds security and grid technology enables parallelization for very large data integration deployments.

Real ROI

Unlike many investments, WebQL provides an immediate, measurable real return on investment. Researchers estimate that on average, analysts spend more than 17 hours a week gathering the information they need at an annual cost of more than \$26,000 per analyst. WebQL automates that activity.

As importantly, many companies simply do not seek out information available to them because the process is too cumbersome and costly. Automating complex data extraction provides many companies with the information and competitive edge they need to win.

CONTACT INFORMATION

QL2 Software, Inc.

316 Occidental Ave. S., Suite 410
Seattle, WA 98104

sales@QL2.com • www.QL2.com

www.webql.com

Phone: +1.206.443.6836

Toll Free: 800.750.8830

Fax: +1.206.269.0694