White Paper

Bringing Order to Data Chaos:
Improving Productivity with Information Extraction Tools

Use It or Lose It

The Web has created an explosion of freely available rich (and not so rich) information. On top of this, even the most basic computer has gigabytes of storage space on which important information is stored and can be hard to find. Yet despite all this available data, there is a problem: How do you get the exact information you want, in the format you want, quickly, easily and without great expense? Solving this problem is critical to staying competitive.

Without Extraction Tools

Tools are needed to manage all available information including the Web, subscription services, and internal data stores. Without an extraction tool (a product specifically designed to find, organize, and output the data you want), you have very poor choices for getting information. Your choices are:

- **Use search engines:** Search engines help find some Web information, but they do not pinpoint information, cannot fill out web forms they encounter to get you the information you need, are perpetually behind in indexing content, and at best, can only go two or three levels deep into a Web site. And they cannot search file directories on your network.
- **Manually surf the Web and file directories:** Aside from the labor-intensive aspect of this option, the work is tedious, costly, error prone, and very time consuming. Humans have to read the content of each page to see if it matches their criteria, whereas a computer is simply matching patterns, which is so much faster.
- **Create custom programming:** Custom programming is costly, can be buggy, requires maintenance, and takes time to develop. Plus the programs must be constantly updated as the location of information frequently changes.

Inefficient methods means the information analyst spends time finding, collecting, and aggregating data instead of analyzing data and gaining the competitive edge. This also affects the application programmer who has to spend time developing extraction tools instead of developing tools for the core business..

New Solutions Improve Productivity

Extraction tools using a concise notation to define precise navigation and extraction rules greatly reduce the time spent on systematic collection efforts. Tools that support a variety of format options provide a single development platform for all collection needs regardless of electronic information source.

Early attempts at software tools for "Web harvesting" and unstructured data mining emerged, and started to get the attention of information professionals. These products did a reasonable job of finding and extracting Web information for intelligence gathering purposes. But this was not enough. Organizations needed to reach the "deep Web" and other electronic information sources, capabilities beyond simplistic Web content clipping.

A new generation of information extraction tools is markedly improving productivity for information analysts and application developers.

Uses for Extraction Tools

The most popular applications for information extraction tools remain competitive intelligence gathering and market research, but there are some new applications emerging as organizations learn how to better use the functionality in the new generation of tools.

- **Deep Web price gathering:** The explosion of e-tailing, e-business, and e-government makes a plethora of competitive pricing information available on Web sites and government information portals. Unfortunately, price lists are difficult to extract without selecting product categories or filling out Web forms. Also, some prices are buried deep in .pdf documents. Automated forms completion and automated downloading are necessary features to retrieve prices from the deep Web.
- **Primary research:** Message boards, e-pinion sites, and other Web forums provide a wealth of public opinion and user experience information on consumer products, air travel, test drives, experimental drugs, etc. While much of this information can be found with a search engine, features like simultaneous board crawling, selective content extraction, task scheduling, and custom output reformatting are only available with extraction tools.
- **Content aggregation for information portals:** Content is exploding and available from Web and non-Web sources. Extraction tools can crawl the Web, internal information sources, and subscription services to automatically populate portals with pertinent content such as competitive information, news, and financial data.
- **Supporting CRM systems:** The Web is a valuable source of external data to selectively populate a data warehouse or a CRM database. To date most organizations focus on aggregating internal data for their data warehouses and CRM systems. Now, however, some organizations are realizing the value of adding external data as well. In the book Web Farming for the Data Warehouse from Morgan Kaufman Publishers, Dr. Richard Hackathorn writes, "It is the synergism of external market information with internal customer data that creates the greatest business benefit."
- **Scientific research:** Scientific information on a given topic (such as a gene sequence) is available on multiple Web sites and subscription services. An effective extraction tool can automate the location and extraction of this information and aggregate it into a single presentation format or portal. This saves scientific researchers countless hours of searching, reading, copying, and pasting.
- **Business activity monitoring:** Extraction tools can continuously monitor dynamically changing information sources to provide real time alerts and to populate information portals and dashboards.

Extraction Tools Versus Search Engines

What is the difference between an information extraction tool and a search engine? The simple answer is that extraction tools pick up where search engines leave off, doing the work the search engine is not capable of. Table 1 displays a side-by-side comparison:

DATA GATHERING MECHANISM	LOCATE	PINPOINT	EXTRACT	INTEGRATE
Extraction Tools	automated	automated	automated	automated
Search Engines	automated	manual	manual	manual

Table 1: Extraction tools versus search engines

Search Engines

Search engines locate information and point to it. They typically go no deeper than two or three levels into a Web site to find information and then return URLs, meta descriptions, and meta keywords. The meta descriptions and keywords can be bogus, because some webmasters load their meta data with popular descriptions and keywords in order to create hits, and search engines cannot distinguish the difference. The

search engine cannot do anything beyond the simple matching of keywords to return URLs. If you are using a search engine for data gathering, tasks such as pinpointing, extraction, and integration of useful information, have to be accomplished by a person or people who complete the following steps:

- Skim the content until the information is found
- Mark the information (usually with a mouse)
- Copy information
- Switch to another application (such as a spreadsheet or database)
- Paste the information into that application

Taking it a step further, many Web sites or subscription services require one or more manual entries prior to retrieving and displaying information, further complicating and elongating the process. Automated forms discovery and completion is something most search engines cannot do, and add another task for the researcher.

Extraction Tools

Extraction tools automate the full process of gathering, pinpointing, and outputting data, thereby freeing resources from these tasks. A robust extraction tool should be able to perform all the tasks you need, quickly and easily.

ATTRIBUTES OF A COMPREHENSIVE EXTRACTION TOOL

Not all information extraction tools are alike. You should consider the following when evaluating an extraction product:

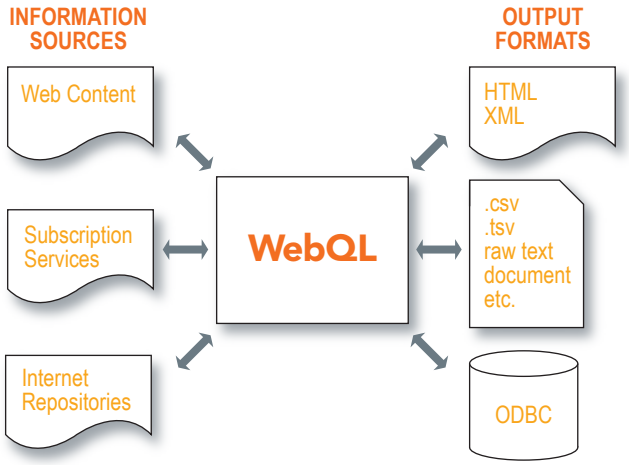
- **Precision and automation:** There are three types of methods you can use to locate and extract data. Some extraction tools use artificial intelligence (AI) techniques. While interesting and highly automated, these tools are often imprecise. Some tools use a drag-and-drop GUIs, which provides precision, but are not automatic and can be impractical if the data you are seeking resides on a variable number of pages with inconsistent formats. The most effective tools use a concise notation capable of calling on a variety of precise extraction techniques depending on the structure of the information solving both the automation and the precision problems.
- **Integration with APIs (Application Programming Interfaces):** Application programmers or Web masters may want to incorporate content into their applications. To enable this, extraction tools should provide robust APIs for the popular computing environments (ActiveX, COM, Java, C++, VBA, SOAP).
- **Integrated navigation and extraction:** An extraction tool should have integrated navigation and extraction. A superior extraction tool uses a seamless method to both navigate and extract data equally well. An inferior extraction tool may be effective once it gets to a page, but cannot navigate well to the page in the first place. This is the issue with tools using both AI and drag-and-drop GUIs. To deal with this deficiency, these tools "bolt on" scripting languages to handle navigation. The result of a mix of technologies for navigation and extraction is an application that is clunky, slow, and difficult to maintain, particularly if the navigation is page-content sensitive.
- **Scalability:** Information extraction applications can be very resource and time intensive, therefore extraction tools must be scalable and distributable. A superior tool can be distributed across multiple processors, does not limit the number of processors per application, and can process multiple applications in parallel.

- **Identity protection:** For a variety of reasons, organizations may not want to be identified when visiting Web sites. Depending on the application and traffic volume, extraction tools need to provide options for identity protection. For high performance and ease of implementation, this capability should be tightly integrated.
- **Multi-format support:** Information comes from a variety of places and in a variety of formats. An effective tool should extract data from all the commonly used file formats (e.g., .pdf, .doc, .xls, .html). It should be able to find them both on the Web as well as from sources outside the Web (internal data stores and subscription services). Lastly it should be able to directly integrate the output into popular presentation or storage formats.

Introducing WebQL

WebQL (Web Query Language) was designed to readily retrieve information from unstructured, semi-structured, and structured information sources, just as Structured Query Language (SQL) does from enterprise databases. Using concise SQL-like constructs, WebQL extracts information from these sources and integrates it into storage formats, presentation formats, and applications.

In summary, WebQL is a software tool for: 1) Querying documents and tabular information, 2) extracting information from these sources into usable presentation or storage formats, 3) for subsequent processing by other applications in real time.



Advanced WebQL Features

- Concise notation based on familiar ANSI standard SQL including full support for joins, grouping, sorting, and set operations.
- Fully featured IDE includes syntax highlighting editor, graphical data flow monitor, and real-time delivery of results
- Standalone or server deployments available for Windows, Linux, and Unix
- Integrated extraction and navigation techniques simplify the development and maintenance of applications
- Supports all modern technologies for parsing and extracting unstructured, semi-structured and structured data, including a number of advanced proprietary algorithms
- Simultaneous and uniform input and output of all common file formats – HTML, XML, PDF, DOC, CSV, TAB, images, databases, proprietary email formats, etc.
- Automatically employs sophisticated algorithms for page request throttling and parallel processing – no special coding required
- A variety of robust integration options – Web Services, Web browser, command line, GUI, ActiveX, and APIs for various languages
- Comprehensive facilities for error trapping and reporting
- "Point-in-time" page archiving via HTML localization or graphical sharpshooting of rendered pages

- Transparently supports all Web functionality – script execution, forms handling, cookies, user agents, frames, tables, authentication, etc.
- Query execution scheduling enabling "lights out" operation
- Integrated identity protection options

Supported File Formats

- HTML
- raw (binary)
- .csv (space delimited text)
- .tsv (tab delimited text)
- .txt
- .doc
- .xls
- .xml (with optional xsl style sheets)
- ODBC compliant databases

WebQL Application Programming Interfaces (APIs)

- Java interface for cross platform environments
- SOAP interface for Web Services
- C++ interface for high performance applications
- ActiveX / COM / .NET interfaces for the Windows environment
- VBA interface for the Microsoft Office environment

WebQL Product Packaging and Services

- **WebQL Enterprise Edition:** This software package enables organizations to create and run their own WebQL applications on their own systems. It is the ideal solution for organizations with varied extraction and integration needs that want complete control and confidentiality of the information.
- **QL2 Solutions:** QL2 Solutions are developed using WebQL. Once developed, the QL2 Solution is installed and runs on a customer's system. QL2 Solutions are suited to organizations wanting control and confidentiality, but do not have the capability or desire to create their own WebQL applications.
- **QL2 Hosted Services:** QL2 Hosted Services are targeted to organizations wanting to completely outsource their extraction and integration requirements. The organization contracts with QL2 to develop and run the application on QL2 systems. QL2 delivers the harvested data in the customer-specified format.

System Requirements and Platforms

- Pentium III or faster processor
- 512MB RAM suggested (256MB minimum)
- 25MB available hard-disk space
- Windows 2000, Windows XP, or Linux
- High speed Internet connectivity

WebQL and Internet Privacy

There have been various attempts to harvest Web information by monitoring the surfing behavior of individuals. Some people consider these attempts invasive snooping and this behavior has been met with strong negative reactions. WebQL functions on very different principles. It only harvests information from publicly accessible Web pages and completely respects the privacy of an individual's Web surfing behavior.