

A decorative graphic on the right side of the page features three overlapping circles of varying sizes, each composed of concentric blue rings. Two thin blue lines intersect at the top left and extend diagonally across the page, framing the circles.

# **Custom-designed Web Crawler for Advanced Web Data Extraction and Data Mining**

A quick guide for Ficstar Web Grabber

Ficstar Web Grabber offers efficient, fully-automated web data extraction capabilities that eliminate lengthy time, errors, and expenses associated with traditional approaches on finding, collecting, and saving web content on today's Internet.

**May 2010**

## *Introduction*

---

Web data extraction and data mining is a mission-critical activity for many businesses today – but one that creates unique and unavoidable challenges. The manual processes most companies use to satisfy their Web information needs are slow and highly error prone, and waste a tremendous amount of both human and financial assets. Additionally, few companies have access to the technical or programming resources needed to design and develop their own automated web data extraction solutions in house, and most of the products on the market today are significantly limited in their ability to deliver the needed results.

In this paper, readers will learn about the drawbacks and disadvantages of using manual processes, in-house programming, or other third-party tools to execute their most important Web data extraction tasks. Additionally, they'll find out about Ficstar Web Grabber, a robust solution designed to help businesses overcome those obstacles, enhancing productivity, increasing accuracy, and minimizing the costs associated with related tasks.

Ficstar Web Grabber offers efficient, fully-automated web data extraction capabilities that eliminate the time, mistakes, and expenses associated with manually finding, collecting, and saving Web content. It dynamically locates and captures information from target sites, and automatically transforms it into text files, Excel spreadsheets, or any database formats. And, Web Grabber can be fully-customized to access any static or dynamic data, including text and images, from virtually any web site.

## *Traditional Web Data Extraction Methods: The Challenges*

---

The Internet revolution has transformed the Web into the largest and most diverse information resource ever created, providing an endless flow of data about virtually any topic. This expansive knowledge base can benefit not only individuals, but corporations, government agencies, educational institutions, and other business entities as well.

Organizations can leverage the content that exists across the Internet for a variety of vital purposes – to obtain competitive intelligence, conduct financial research, gain insight into marketplace trends, enhance e-commerce operations, gather contact information for use in marketing and promotional campaigns, and much, much more. But, getting to that data, and transforming it into something relevant and usable is not as easy as it may sound. Information must be searched for, located, filtered, and extracted into the needed formats.

Search engines provide only a fraction of the features companies need to facilitate effective Web data collection. So, in order to locate, capture, and store the high volumes of information they need from targeted Web sites, businesses typically use one of two cumbersome and error-prone methods:

### *Manual “Cut and Paste”*

In many organizations, employees are hired specifically to spend countless hours copying, pasting, and re-formatting text, images, documents, multi-media files, and other data from thousands of pages across the Web. This approach is both slow and expensive, wasting a significant amount of time and money, and negatively impacting the integrity and accuracy of results due to the risk of man-made mistakes.

A recent study conducted by leading analyst firm IDC shows that the average information worker can spend as much as 15 to 20 hours each week – almost half of his or her time – searching for and copying needed information from the Internet. That same report demonstrated that this labor-intensive browsing and gathering can cost companies a whopping \$26,700 per employee each year.

### *Development of “Homegrown” Tools*

Many companies utilize their valuable IT resources to design and build custom applications to automate the needed manual tasks for Internet content collection and aggregation. But, building these systems from the ground up is quite time-consuming and complex, requiring programmers to abandon other critical corporate technology initiatives to create such functionality and features as Internet navigation, page selection, and content extraction.

Additionally, there are few truly reliable and cost-effective solutions on the market that fully streamline and automate the extraction and storage of Web data. Many of the tools available today use a traditional Socket approach to get to needed content, a method that creates a

connection between the Web server and the client computer using the Server Socket and Client Socket API functions. The client computer sends HTTP requests to the Web server, and receives HTTP responses in return. The HTML source codes are then read from the server by parsing the responses before output data can be generated.

The primary drawback of this approach is that many Web pages today are developed with embedded client-side applications such as Java scripts or AJAX, and as a result, some page data can only be created by those client-side applications in the Web browsers. Those solutions that use the traditional Socket method to support Web data extraction can only capture Internet data that is in HTML, not in scripts. So, collected page content will be incomplete, and the needed results will not be achieved.

## *Critical Success Factors*

---

Truly effective collection, transformation, storage, and mining of Internet data requires a broad range of features and powerful functionality in order to meet the unique, diverse, and sophisticated needs of today's businesses. In order to ensure success and produce the desired results, a Web data extraction software solution must:

- Be fully-automated, eliminating all the manual tasks that can put a strain on administrative staff, waste valuable corporate budget dollars, and produce a high number of errors.
- Work with Java scripts, AJAX, as well as framed Web pages and special HTML 4.0 tags, to capture the most complete results possible from any Web site, regardless of how it is designed.
- Search multiple Web sites, as well as multiple instances within a single Web site, simultaneously, to rapidly deliver the comprehensive content desired.
- Provide maximum flexibility, allowing custom alterations to be made quickly and easily to address changing and emerging business requirements.
- Be built on powerful client/server architecture, so it can run on multiple computers and support multiple users and their separate extraction projects at the same time.
- Deliver results in multiple formats, such as CSV, Microsoft Excel, Access, SQL Server, MySQL and other databases to support the broadest range of Web information needs.

# *Ficstar Software: Providing Today's Most Powerful, Full-Featured Web Data Extraction Solutions*

---

## **Overview**

Ficstar Software Inc. provides robust, cutting-edge software solutions and services for automated web data extraction and data mining. The Toronto-based company was incorporated in 2005 with a mission of leveraging the latest Internet technologies to automate and enhance the retrieval, collection, and storage of web-based content. With a comprehensive portfolio of innovative, yet affordable, products and services, Ficstar helps companies of all types and sizes to increase efficiency and productivity, and reduce overhead expenses, while achieving accurate, high-quality results from their most complex Web extraction efforts. With Ficstar solutions, businesses can rapidly and effectively gather and save any vital information from anywhere on the Web.

## **The Benefits**

Ficstar Web Grabber delivers significant value to organizations of all types and sizes, helping them to satisfy their Web data extraction and mining needs more efficiently and cost-effectively than ever before. Key benefits include:

### *Improved Accuracy*

Man-made mistakes are virtually eliminated through complete, end-to-end automation of the entire extraction and storage process. So, results are always accurate, each and every time.

### *Accelerated Results and Increased Staff Efficiency*

Ficstar Web Grabber has been proven in real-world scenarios to produce results more than 100 times faster than manual Web data extraction. For example, NASA is currently using Ficstar's customized software to monitor all NASA-obtained patents from the Internet. A manual "search and save" process for all existing patent data used to take the agency's employees months of time to complete. But, with Ficstar solutions, NASA can obtain the needed information in merely days and sometimes even hours.

### *Reduced Costs*

With Ficstar Web Grabber, companies can minimize the expenses associated with the collection, aggregation, and storage of Internet-based data. In fact, one Ficstar client, a leading UK-based e-commerce vendor, used to hire a group of full-time marketing analysts to validate product prices on their competitors' Web sites by hand – taking several weeks per site, and creating a huge spike in human resource and payroll costs. But, once Ficstar's customized software was deployed, all manual labor activities were eliminated and replaced by fully-automated computer executable processes. The entire product catalog of more than 50 competitor Web sites can

now be scanned repeatedly on a daily basis, without the need to spend money on additional staff resources.

## How It Works

Ficstar Web Grabber is an independent, Windows-executable application that can run on any Windows-based system. The program is seamlessly and completely integrated with the default Web browser on the user's computer, and can work with Microsoft Internet Explorer, Netscape, Firefox, Mozilla, Opera, Safari, AOL, and other popular browser products.

The Web browsing process is fully automated using revised Windows Socket HTTP components. Pre-set data extraction configuration settings allow for effective navigation of virtually any target Web sites, as well as dynamic parsing of both dynamic and static page content – faster and more effectively, without the need for manual “cut and paste” activities.

Additionally, Ficstar Web Grabber uses pre-defined criteria and a collection of hyperlinks to fully streamline and automate the searching and location of desired data residing across various URLs. This increases the accuracy and integrity of recorded data by minimizing the risk of errors that can result from human data entry.

Ficstar Web Grabber connects only to those servers upon which target Web sites operate. It will, however, also connect to a secure server provided by Ficstar at timely pre-set intervals, to validate and verify the IP addresses of the computers running the Ficstar software. This information is used to prevent target Web sites from “blocking” the end user's true IP address, preventing them from accessing and gathering the needed data. The program can also switch to a “secured” model to hide the real IP address of the executing computer if necessary.

Furthermore, the Ficstar application will also continuously verify the data extraction process, to ensure the smoothest, problem-free operations. Invalid or unsuccessful searches can be automatically reactivated, for example, by using an alternative routing proxy server to re-access the target Web site. When this function is executed, no Web sites other than the one targeted for data retrieval will be accessed. This function can be manually disabled by users, for various computer and network security purposes.

Finally, collected data is then instantly and automatically transformed into the format of the user's choice – CSV, Plain Text, Excel, SQL, local or remote databases etc.

## Built On a Robust Foundation

At the heart of the Ficstar Web Grabber solution is a core class library built over the course of many years. This library serves as a key development kernel of the application, allowing the

program to easily handle internal HTML controls and to dynamically parse HTML text files to retrieve the needed results.

The processing and collection of most requested user data can also be fully automated by Ficstar Web Grabber. Since there are many ways to effectively collect links, they are identified and gathered based on pre-set configurations in which certain rules are set to locate data on given Web sites.

For example, many companies pull data from Internet forums, to assess marketplace perceptions, understand customer needs and opinions, etc. Each forum has an ID parameter contained in its URL that uniquely identifies it. Ficstar Web Grabber will use that specific link – containing the identifying parameter – to determine which data to collect. This ensures maximum accuracy in the output results.

## **Unique Design in Software Architecture**

Ficstar Web Grabber provides industry-leading web data extraction capabilities that have exceeded expectations of its clients, providing the speed, ease, and features needed to satisfy their enterprise-level data mining needs. The system is embedded with full of cutting-edge Internet technologies such as Grid Computing and military mission-critical Web anonymous surfing protection.

Additionally, Ficstar Web Grabber is developed on a flexible software structure that can be easily custom-designed to integrate with virtually any existing data management system. Ficstar's unique model of customized, affordable solutions that are tailored to each company's specific needs has resonated well with its clients, who have successfully and seamlessly integrated Ficstar solutions into their existing enterprise infrastructures.

Lastly, Ficstar Web Grabber is based on an inventive algorithm, combined with superior innovation in artificial intelligence. It offers businesses a faster and more effective way to perform web data extraction and data mining. Ficstar's robust, innovative solutions eliminate repetitive, error-prone manual processes that waste time and money. Required data is quickly and cost-effectively retrieved, whenever the users need it.

## **Optimum Flexibility**

With Ficstar Web Grabber, companies can address virtually any Web data extraction need. The software can be fully customized to work with virtually any Web site, including e-commerce sites such as Amazon.com or Shopping.com, member list directories, search engines such as Google and Yahoo, financial sites such as securities firms and stock exchanges, job posting sites like Monster.com, and many others. Since different types of sites are programmed in different

languages and built on different technologies, each solution is modified specifically to work seamlessly with the unique characteristics of each target Web site. Additionally, future alterations to the software program can be made with ease, to handle changes to target Web sites over time.

Additionally, Ficstar Web Grabber offers unlimited flexibility when it comes to data output. Extracted information can be transformed into a variety of formats to address the broadest range of needs. Users can save their results as CSV files, Excel spreadsheets, or common database formats such as Microsoft Access, SQL Server, MySQL, Oracle and many others.

## *The Ficstar Portfolio of Products and Services*

---

The Ficstar team has developed a comprehensive suite of offerings to enable companies to quickly and effectively address all their Web data extraction and data mining needs. Available products include:

- **Web Grabber** – powerful, fast, fully-automated Web data extraction. Available for e-commerce, contacts, Web extraction, and Web crawling. Web Grabber can also be easily custom-tailored to meet unique and specific requirements.
- **Website Keyword Monitor** – to instantly track and capture changes to target Web sites, such as modifications to text, Meta tags, keywords, page layouts, or dates and times for posted content.
- **Search Engine Ranking Tracker** – for continuous monitoring of rankings in search engines and directories, as well as the results of pay-per-click (PPC) campaign efforts.
- **Web Browsing Automator** – streamlines and accelerates the end-to-end Web browsing process by automating data entry, data synchronization, form completion and submission, order entry, and other important tasks.
- **Web Grabber .Net** – an enhanced Web data extraction solution designed and built entirely on a .NET 2.0/3.0 architecture, providing the ability to scrape data from script-intensive or frame-based pages and other advanced features.

Multi-edition solutions are also provided, to satisfy the specific Web mining needs of various industry sectors, including small businesses, enterprises, government agencies, and non-for-profit organizations, as well as online shopping engine submitting and order entry for online retailers and Internet merchants, and automatic claims form submission for insurance companies. Functional solutions for competitive analysis, sales lead generation, and business intelligence are also available.

Additionally, Ficstar offers a wide range of comprehensive services designed to compliment and enhance its products. These are delivered by some of today's most skilled and knowledgeable professionals, and are designed to ensure that clients achieve optimum return on investment from their Ficstar products. Available services include data extraction of e-commerce data, financial data, contact information from business directories, and bibliographies and resumes from job posting sites, as well as consulting, training, and post-implementation support.

## *Ficstar Web Grabber in Action: Customer Success Case Studies*

---

### **Extracting E-Commerce Data from Competitive Web Sites**

#### *Overview*

The company is a well-known e-commerce firm specializing in a wide array of technology products such as computers, large screen televisions and other consumer electronics. With tens of thousands of products available through its online store, this leading Internet merchant serves customers across the world, processing and fulfilling millions of Web orders every year.

#### *The Challenge*

The Internet retailer was conducting extensive marketplace research by continuously monitoring and collecting data from more than 50 competitive Web sites, each containing as many as 20,000 products. Staff members spend countless hours performing manual Web page scraping and Web data extraction. The entire process was wasting a tremendous amount of employee time and effort, and distracting workers from other important company projects.

#### *The Solution*

With Ficstar Web Grabber, this e-commerce organization was empowered to:

- Leverage a single, fully integrated Web data extraction solution for dynamically searching and locating results on all targeted Web sites.
- Perform searches and Web page scraping from multiple sites simultaneously.
- Scheduled searches to increase speed and avoid high Internet traffic.
- Automatically detect variations between competitors' product catalogs.
- Adjust its own prices, to match or beat those of its competitors, in real-time.

#### *The Results*

- **Increased data accuracy.** Man-made mistakes, and the manual, error-prone tasks that were causing them, were completely eliminated and replaced with rapid, fully automated activities that ensured maximum output integrity.
- **Improved employee efficiency.** Web data extraction activities that used to take days, or even weeks, can now be completed in just a fraction of the time.
- **Reduced costs.** Annual marketing expenditures were dramatically reduced due to increased staff efficiency and effectiveness, and minimize payroll costs by hiring fewer price analysts.

## Collection of Web-based Data about Real Estate Properties

### Overview

The company is a leading contributor to the U.S. housing industry, and one of the country's largest mortgage lenders. It provides a variety of home loans and related funding services, making housing more available and more affordable to prospective buyers in communities across the country.

### The Challenge

The firm maintains a large database of available residential real estate properties throughout the 50 states. In order to compile that information, they must conduct extensive Web data mining, browsing many real estate info Web sites, and collect and save information about current properties, features, sold prices, and new ownership data. One of the challenges to collect the required data is that each Web site is structured in a different layout format, requiring extra work to complete even the simplest search and retrieval activities. Another issue is the complexity of the output results, which includes tax and owner contact data, as well as information about selling transactions – all located in separate HTML tables scattered across multiple Web pages for each property listed.

### The Solution

Ficstar designed a completely customized Web Grabber software solution that

- Provides a single, fully integrated system for searching and retrieving data from all targeted Web sites
- Easily locates needed results using all criteria and criteria combinations.
- Conducts simultaneous searches of many Web sites at the same time.
- Allows for automatic scheduling of searches, so different Web sites are browsed at different times to avoid conflicts or periods of high Internet traffic.
- Leverages a unique and highly sophisticated computer algorithm that compares search results to existing data, and makes additions, updates, and changes to the database where needed.

### The Results

- **Accelerated output.** The time needed to find, collect, and save property data has been dramatically reduced.
- **Increased information integrity.** By automating the entire process from start to finish, and eliminating the risk of human error. The accuracy of the information contained in the company's database is increased significantly.
- **Improved productivity.** Staff members were freed from the time-consuming and labor-intensive "cut and paste"-type processes associated with manual Web data collection.

## Extraction of Data from Online Classified Auto Ads

### *Overview*

The company is the leading provider of vehicle data and related statistics in the United States. It serves millions of used car consumers every year, making valuable information such as current titles and past owners, damage histories, service records, odometer readings, and inspection results readily accessible via the Internet.

### *The Challenge*

The organization maintains a unique database that contains more than five billion records, including online classified ads for available used cars across the country. In order to collect and compile all needed data about those “for sale” autos, staff members must search for and retrieve vehicle details, pricing, and seller or dealer information from numerous partner Web sites. This process proved to be both time-consuming and costly for the company, because each target Web site contains a large number of vehicles, and related information is changed, updated, or added often, forcing employees to re-conduct their searches on a very frequent basis.

### *The Solution*

Ficstar empowered this auto data provider to:

- Utilize pre-defined search criteria to rapidly generate the required results.
- Simultaneously search multiple Web sites at the same time.
- Collect thorough and detailed vehicle information from across the Internet, including make, model, year, VIN, price, and location.
- Compile complete and accurate seller or dealer data, such as name, address, phone number, fax, and email addresses from a wide range of Web sites.

### *The Results*

- **Rapid output and updates.** Target Web sites are automatically and continuously searched and scraped.
- **Reduced errors.** Mistakes from manual processes were fully eliminated.
- **Increased efficiency.** The traditional “copy and paste” activities that previously drained employee time no longer need to be performed, dramatically improving the productivity of staff members.
- **Minimized costs.** Significant increases in worker efficiency, combined with the elimination of data errors and associated re-work, have reduced related expenses and increased cost-efficiency.

## *Ficstar Web Grabber Work for...*

---

<b>Category</b>	<b>Website</b>
Auction	eBay
Business List	CISCO Partners GolfCourse Store_Locator_Walmart
Car Sellers	Autotrader Craigslist
Contact	Blogger Lawyers LinkedIn
E-commerce	Amazon Home-Depot Sears Samsclub JCWhitney
Financial Data	NYMEX
Health Care Providers	AACC Webmd
Insurance	Allstate
Job	Monster Careerbuilder HotJobs
News	CNN MSNBC Yahoo
Patents	USPTO
Real Estate Properties	Realtor
Restaurants	AAA OpenTable Zagat
Science	Pubmed
Travel	Hotels HRS Expedia Travelocity

## Comparison of Ficstar Web Grabber Editions

---

Category	Function	Express	Standard	Enterprise
Web Crawling	Search all pages on a website	●	●	●
	Search pages by category	●	●	●
	Search multiple websites		●	●
	Secured web pages (https)		●	●
	Website with user logins		●	●
	Synchronizing different websites		●	●
	Multiple-instance crawling			●
Data Extraction	Self-verification on output results	●	●	●
	Auto resume from last search	●	●	●
	Data extraction search scheduler		●	●
	Add/remove output data fields		●	●
	Change output database		●	●
Output Database	Excel/CSV	●	●	●
	Access/MySQL		●	●
	Oracle/DB2/SQL Server/Sybase			●
Performance	Daily result update		●	●
	Turbo Website Scrape			●
	Grid Computing			●
Security	Integrated proxy technologies	●	●	●
	Embedded Ficstar IP Hider			●
Data Analysis	Page difference monitoring		●	●
	Homepage tracking		●	●
	Changes on historical data			●
	Page update email alerts			●
Reports	Status log	●	●	●
	Exception report	●	●	●
Languages	Latin/WGL	●	●	●
	Chinese/Japanese/Korean			●
	Arabian/Hebrew			●
Maintenance	User-requested website update	●	●	●
	Add/remove target websites		●	●
	Add/remove monitoring criteria			●
	Auto Scrape Update			●

## *Feature Summary for Ficstar Web Grabber*

---

<b>Value</b>	<b>Web Grabber Feature</b>	<b>Benefits</b>
Fast search results	Fully-automated search on websites, hundreds of times faster than manual Web data extraction	Accelerated web data collection process
No man-made mistakes	Man-made mistakes are eliminated through complete, end-to-end automation of the entire extraction and storage process. Auto-correction function to protect inaccurate results to be saved	Improved accuracy on results
Automate manual data collection process	<ul style="list-style-type: none"> <li>• Schedule searches on target websites at various time range</li> <li>• Resume search from previous stopping point</li> <li>• Search website by category, name, or any possible criteria</li> </ul>	Reduced costs and increased staff efficiency
Various output formats	Supports nearly all possible output database or text format, results can be delivered in multiple formats such as CSV, Excel, MySQL, MS Access, SQL Server, Oracle, DB2, and others	Compatible to nearly any system
Using most advance Internet technologies	Fully compatible in parsing web pages with: <ul style="list-style-type: none"> <li>• Java scripts</li> <li>• AJAX</li> <li>• Frames</li> <li>• Dynamic pages in Java, PHP, ASP.NET, and others</li> </ul>	Getting results from virtually any website
Designed for enterprise-level data extraction capabilities	Searching multiple Web sites, running multiple instances within a single Web site, executing searches on a grid of web data extraction servers, all can be done simultaneously	Powerful multi-layer architecture on centralized relational database system

Customizable	Can be customized for a list of target websites, or for a specific task to generate special results. Easy, fast and accurate outputs can be created in no time	Flexible solution to suit for any custom design need
Security	<ul style="list-style-type: none"> <li>• Fully mimicking human web browsing process</li> <li>• Non-intrusive crawling on websites</li> <li>• Automatic IP address detection and alternation to prevent web crawler being noticed and blacklisted</li> <li>• Data encryption and decryption</li> </ul>	Safe and correct results at all time

## Contact Info

---

### **Ficstar Software Inc.**

245 Fairview Mall Dr., Suite 410

Toronto, ON Canada M2J 4T1

[www.ficstar.com](http://www.ficstar.com)

1.888.666.8865 or 1.416.595.9222

[info@ficstar.com](mailto:info@ficstar.com)

### **Disclaimer**

*Ficstar Software Inc. makes no representations about the suitability of the information contained in this document for any purpose. This document and all of the information it contains is provided "as is" without warranty of any kind whether express or implied. All implied warranties, including, without limitation, implied warranties of merchantability, fitness for a particular purpose, and non-infringement, are hereby expressly disclaimed. In no event shall Ficstar be liable to any person or business entity for any special, direct, indirect, punitive, incidental or consequential damages arising out of or in connection with the use of this document, including, without limitation, any lost profits, business interruption, or loss of programs or information even if Ficstar has been specifically advised of the possibility of such damages whether based on contract, tort, strict liability or otherwise.*