

INTERNET DATA OPPORTUNITY

UTILIZING NEW SOURCES OF REAL-TIME DATA TO STAY
COMPETITIVE AND INFORMED



THE NEW DATA CHALLENGE

Traditional data sources no longer answer all of the questions facing businesses and organizations today. The pace of global competition and rapidly unfolding events dictate that newer, real-time sources of data are needed to keep organizations informed, proactive and competitive.

The Internet is the world's new data source.

Tremendous amounts of information are being published each day on the public Internet – on company websites, blogs, news sites, retail sites, government and municipal sites – just to name a few. The scale of information publishing is accelerating, not just in the United States, but globally.

This data is as diverse as it is valuable, and can benefit any application:

- Business intelligence – Understand key events and capabilities of multiple companies to see trends, opportunities, challenges
- News and publishing – Aggregate news and blog content for redistribution and/or search applications
- People intelligence – Identify key data about potential or existing employees - criminal records, certifications, social media
- Pricing intelligence – Aggregate pricing changes for competitive market analysis
- Risk management and compliance – Collect and aggregate specific information from a diverse set of Internet sources to assess and manage risk

Organizations face a daunting challenge – how do they effectively connect to and utilize this valuable data? Having employees manually aggregate data across thousands of websites would require a veritable army of personnel. And search engines don't provide data in a format that's useful, nor do they index the "deep web" where much of this data resides.

So how can organizations automate this need?

Traditional sources of IT services do not offer much assistance, despite the millions of dollars that organizations often expend to purchase and integrate them. They cannot manage this new "type" of data, instead focusing on internal, traditional sources of data only. For those that can handle Internet data, they only work with very specific types of published data, such as RSS feeds, which represent a mere fraction of the Internet's data and may not be timely or comprehensive enough for an organization's needs.

What is needed in this new information age is a scalable, reliable, automated solution to allow organizations to connect to websites, aggregate accurate and timely information, and integrate this information in a usable format into existing systems or processes.

ESSENTIAL ELEMENTS OF AN EFFECTIVE WEB DATA AGGREGATION SOLUTION

The sheer volume of available data is becoming increasingly incompatible with the traditional options for extracting and sorting data.

Scalable, accurate automation of key data aggregation processes

Given the broad range of data required by organizations, effective organizations need scalable, reliable access to new web data sources. Organizations often initially employ 1 of 3 strategies to acquire this data:

- **Manual** efforts to aggregate this data are fruitless, and require multiple full-time personnel with the sole mission of cutting and pasting from websites.
- **Inexpensive web scraping software** extracts everything on a page, including ads, hidden HTML, and other extraneous content, requiring expensive post processing, and increasing bandwidth and server expenses.
- **Custom, in-house development** efforts focused on the development and maintenance of scripts are expensive and very difficult to maintain. These scripts often result in inaccurate data along with high costs due to script breakage when websites change, requiring expensive, ongoing engineering support to rewrite the scripts.

High performing organizations require reliable, cost-effective access to these new data streams, dictating an automated solution that has the following characteristics:

- **Scalable** – Given an organization’s diverse data needs and the exponential growth of new web data sources, the system must scale to a large number of sites due to the diversity of information and sites.
- **Accurate** – The data returned from this system must be accurate, and not return extra content, such as ads that appear on the page or extraneous HTML that requires post-processing.
- **Robust and reliable** – The Internet is not a traditional database – it is a constantly changing organism. Accessing data from the Internet involves many issues, ranging from network issues to site availability. Any solution must anticipate these changes and be robust enough to work through outages and site changes.
- **Comprehensiveness** – The Internet is made up of websites that utilize a wide range of technologies. A high performing web data solution can handle all of these to maximize the set of accessible data, whether they utilize syndication technologies such as RSS or not. The solution must also work in the deep/queriable web, where a significant amount of valuable information resides.
- **Requires few resources to operate** – Effective organizations need its best people acting on data, not collecting it. An effective data aggregation system empowers organizations by eliminating data gathering work, allowing employees to work on additional value- added activities.

Transformation of data into a usable, normalized format

Websites are inherently unique, which creates a challenge in aggregating data across a large number of sites. Since each website is different, so generally are the formats of the data across sites. An effective data aggregation solution must then include a mechanism of normalizing the collected data into a uniform format, ideally using a nomenclature that data format that can be customized to the specific needs of a customer. This allows the resulting data to be used immediately, integrated into applications and infrastructure that is already in place.

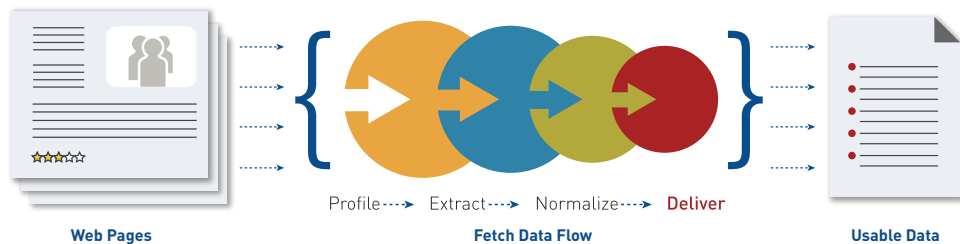
A system that works within the ecosystem of the Internet

The Internet is a complex system of interconnected computer networks that consists of millions of private, public, academic, business, and government networks. Creating reliable connections to data across hundreds or thousands of websites is not an insignificant task, requiring knowledge and capability to deal with the many routine issues that occur in connecting to websites – bandwidth issues, server timeouts, website database errors, etc. Effective systems make the unreliable Internet reliable.

In addition, effective systems work in harmony with the individual websites they come in contact with – respecting site policies, and creating the smallest footprint possible to make the data connection as lightweight to each website as a manual user visit.

FETCH LIVE ACCESS: SEAMLESS ACCESS TO CRITICAL DATA

Fetch Live Access provides a mechanism for organizations to access the critical real-time data essential to today’s competitive environment. Fetch Live Access is a data service that allows organizations to plug-into key data on the web – pricing, people, company, news, product, and more. Its sophisticated technology takes the work out of extracting, aggregating and normalizing data from any website, freeing organizations to use the data, not collect it.



Fetch Live Access provides data access to any website, regardless of the underlying technology that powers it. Whether or not the website has an RSS feed or formal API, Fetch can access the data on each selected sites, precisely extracting only the data that is requested, aggregate it with data extracted from other sites, and transforming it into a customized, uniform format for immediate use.

Based on sophisticated pattern recognition and machine learning technology, combined with over 10 years of operational experience, Fetch makes companies more productive and competitive. With data via Fetch Live Access, companies can augment their traditional data sources and IT infrastructures to understand today's business environment – viewing the latest news and blog posts to understand what's happening in their marketplace, what their competitors are doing, evaluating prospective employees, detecting threats and determining compliance.

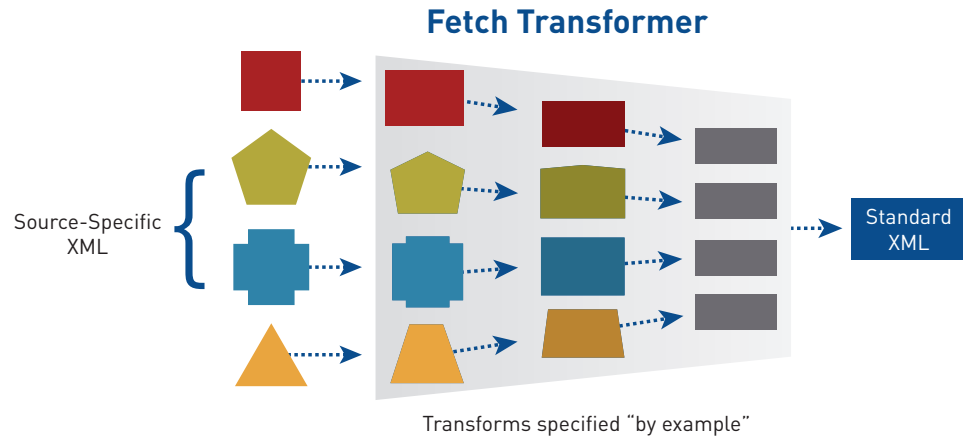
Fetch Live Access is available as a data-as-a-service (hosted web service) model or as licensed software. **The Fetch Live Access service has already processed over 500 million webpages and retrieved over 4 billion data elements.**

Fetch Live Access Technology

Fetch Live Access includes four key technology elements that make it the most scalable, cost-effective mechanism for delivering timely, accurate and critical data to organizations.



- **Profiling the website** – Fetch AgentBuilder builds powerful “web agents” that navigate through websites to extract specific data. These agents locate and extract the targeted data, using a machine learning approach for extracting the right data. Users specify the name of the fields to extract, and provide examples of the fields from sample pages, and the software automatically learns rules for extracting that information. AgentBuilder supports techniques for addressing a wide variety of web site features including forms, cookies, complex navigation, AJAX, frames, malformed HTML and non-HTML data (such as images and structured text).
- **Connect to the data** – Fetch AgentRunner runs and manages agents, using a sophisticated data flow architecture that employs parallel processing to access and extract data from multiple sources at high speed. AgentRunner's load balancing and parallel processing architecture automatically validates agent output, and AgentRunner supports large numbers of agents operating simultaneously in a server side environment. AgentRunner can harvest data on a scheduled basis or in real-time. AgentRunner manages the validation, cleaning, transforming and standardizing of extracted data.
- **Transform the output** – Data from disparate sites isn't truly useful unless it is normalized in a standard format that organizations can use, with their unique data labeling and formats. The Fetch XML Transformer standardizes data collected by Fetch agents, using sophisticated, automated transformation technology to customize the data to each customer's specifications.



- Monitor the sites** – Websites and networks can be unreliable, so ongoing monitoring is critical to ensure accurate and complete data sets. Fetch’s AgentMonitor monitors extracted data for errors, automatically retrying requests if sites are down and sending alerts about potential errors during agent executions, so that agents can be quickly repaired if a web site changes. This enables proactive management of agents and data for ultimate system up-time.

Fetch manages all aspects of data aggregation and normalization. Equipped with Fetch Live Access, organizations can tame the constantly changing anarchy of the Internet and emerge with reliable and predictable sources of accurate, clean, and usable data.

FETCH POWERS ORGANIZATIONS WITH DATA

Fetch solutions are used by a wide variety of organizations to jumpstart their Internet data needs



Dow Jones | Type of data: News

Dow Jones is one of the premier media brands in the world. As Internet-based news grew in importance, Dow Jones knew it needed a scalable, reliable system to gather Internet-based news for its acclaimed Factiva service. The Fetch Live Access Platform has been reliably powering the Factiva service around-the-clock with millions of news stories in a wide variety of languages since 2008.



Shopzilla | Type of data: Prices

Shopzilla is the gold standard for comparative shopping search engines, connecting shoppers with over 80 million products. Given the explosion of Internet-based retailers, Shopzilla came to Fetch to help expand the breadth of their offering. Fetch helps Shopzilla to aggregate millions of pages of product information to offer the most comprehensive online shopping experience on the web.



McGraw Hill | Type of data: Blogs

Organizations and individuals rely on McGraw Hill for independent and unbiased surveys of customer satisfaction, product quality and buyer behavior. McGraw Hill needed a scalable and reliable mechanism to aggregate millions of blogs for insight into a wide variety of business topics. Fetch supplies McGraw Hill with millions of blogs via a hosted service, helping to provide critical insight to its customers.



ADP | Type of data: People

ADP is one of the world's largest providers of business outsourcing solutions, offering a wide range of HR, payroll, tax and benefits administration solutions. The Fetch Live Access service allows ADP to electronically connect with court sites all over the U.S. for real-time criminal history data to power its candidate background screening products.