CITO Research
Advancing the craft of technology leadership

From Data to Dollars

How Website Content Can Power Your Data Business

SPONSORED BY

Connotate

CONTENTS

Introduction

Website content is arguably the most underused source of publicly available data—even though it can lead companies to better decisions, deeper insights about the competitive landscape, and new revenue streams. Website content—both textual and visual—can fuel a wide array of valuable applications and data monetization opportunities, including information products such as company and financial databases and directories, competitive intelligence, background checks, and compliance, to name just a few.

This valuable resource is often not leveraged fully because it is difficult to harvest it reliably and at scale. Website content is visual and unstructured. As such, it is designed for individual human consumption, not for systematic, large-scale automated harvesting, or subsequent transformation into machine-readable data.

However, new technology and machine learning science now make it possible to take websites' visual, unstructured content and turn it into a high-scale flow of usable data. This technology automates many processes, enabling companies to rapidly and cost-effectively build extraction Agents, automatically harvest and QA the data on a set schedule, pre-process and normalize content from hundreds or thousands of websites, and automatically deliver a clean flow of data into its existing production environment.

This CITO Research paper describes the latent opportunities that lie waiting to be unleashed in the world's 1 billion+ websites—and highlights the most effective ways to apply technology and automation to transform this huge and underleveraged resource into information products and services, revenue streams, and fresh insights.

What Is Website Content?

Website content refers to the vast trove of information that appears on web pages, such as pricing, contact details, hours of operation, product information, reviews, and more. Website content also includes data and documents that can be acquired through the website, such as databases, pdfs, and other downloadable content. It is worth noting that the same automated harvesting technology can also be used to obtain content from any source accessed via browser, such as web portals.

Website content is not to be confused with web analytics (page statistics, site analytics, clickthroughs or any other such metrics) which are used to optimize websites.

Website Content Is Already Making An Impact

Website content can be hugely valuable, and for leading-edge companies, it is already a vital part of many of their data strategies and operations. These companies have created the capabilities to automatically and reliably harvest, integrate, and draw insights from the dynamic content of many different websites—driving new revenues and competitive differentiation.

Here are just a few examples:

- **Looking for brand violations**. A premium brand monitors product and pricing details on 90 retailer/distribution websites to find out whether its products are being overly discounted by these retailers, in violation of their pricing agreements.

- **Aggregating data**. A multimillion-dollar jobs aggregator monitors job postings on 1,000 sites and updates its content in near real-time, ensuring that job seekers who click on a posting are not disappointed by stale links.

- **Gaining competitive intelligence**. Industry leaders monitor competitors' websites for changes such as new offerings and pricing, changes in key personnel, press releases, new collateral, marketing activities, and more. In addition, they monitor review sites to keep tabs on the health of their competitors' brands (as well as their own).

- **Monetizing data**. The information products industry now tops three-quarters of a trillion dollars. Demand for data products such as company information and news, directories, financial data, regulatory information, and much more continues to grow. In addition, website content about products and pricing also powers shopping services, travel sites, recommendation engines, and much more.

These use cases barely scratch the surface of the ways companies profitably leverage website content. It can also be used to enhance existing information products, perform verifications to ensure compliance, serve as a key input into pricing and distribution optimization applications, and much more.

If Website Content Is So Valuable, Why Isn't Everyone Using It?

80%
of companies say that they're underutilizing website content

According to a survey performed by Connotate, a provider of technology and managed data services for high-scale harvesting of website content, most companies already recognize the potential value in leveraging website content. Still, 80% say that they're underutilizing it, in part because of the difficulty of extracting and making sense of such data.

Why is extracting this content so hard? There are three key reasons.

Reason 1: The Web Was Designed for People—One Page at a Time

Creating a robust data stream from the world's websites requires that machines perform extractions and move through websites at high scale and with high precision. Unfortunately, it's very difficult to train machines to perform this task. Most websites are visual, designed to be looked at—by humans. We humans intuitively know the difference between a headline and a price, between one entry and another, and between specifications and descriptive text. We know how to click the "next" button to see additional entries. We know if a site is bringing back the results we want, and we can change the search criteria if it is not. But these are all daunting tasks for machines.

Reason 2: Search Engines Only Scratch the Surface

Another reason why companies underutilize website content in their operations is that so much of that content is hidden from their view. Today, more and more websites are dynamic; website content is generated from a database that isn't visible to the site visitor. These sites require that a user enter search terms or other parameters; the sites then serve up pages in response to that particular set of inputs. Take court records, for example. They aren't indexed, but if you supply a parameter such as last name, the site will dynamically serve up that record.

Such sites are part of what is called "the Deep Web."[1] This is a huge, hidden part of the Web—7,500 terabytes, compared with just 19 terabytes in the Surface Web—that search engines cannot adequately crawl and index. To get to the Deep Web, users need specialized tools and approaches that go beyond search engines. The good news: with the right approaches, the Deep Web can be mined: the vast majority of this "hidden" content, about 95%, is, in fact, publicly available.[2]

[1] http://money.cnn.com/2014/03/10/technology/deep-web/index.html

[2] http://hewilson.wordpress.com/what-is-the-deep-web/statistics/

Reason 3: Most Extraction Approaches Cannot Provide the Data Flow Companies Require

Website content is typically extracted in one of five ways, but for companies seeking an efficient way to create a high-scale, clean, and easily ingestible data flow, most of these approaches have fatal flaws.

- **People: A Page at a Time**. This approach, where researchers extract website content and assimilate it, generally involves humans cutting and pasting data from sites into spreadsheets or other databases. Often individuals monitoring several companies may do it themselves; as demands increase, the company may hire offshore resources. Because it's easy to get started at relatively low cost, this is often the first approach that companies adopt.

  However, as demands grow, this approach becomes unwieldy—and costs grow rapidly. Moreover, it takes humans a long time to get the data under this highly manual approach. Quality control suffers, the company is spending more and more money on these operations, data is not flowing fast enough—and there is still the problem of normalizing and transforming all this data so that it can be easily ingested and integrated. In short, the approach does not scale, is prone to error, and is extremely time-consuming.

- **Open Source or Non-Commercial-Grade Software Solutions: Small Scale for Periodic Extractions**. These solutions help automate the extraction of information from one site at a time—but extracting information from many sites and aggregating it into a single database becomes problematic because there are few built-in pre-processing routines. Data must be integrated and normalized via post-processing. In addition, there are virtually no automated work processes built into these platforms. For example, users can't set up extraction schedules and automatically harvest the data they want at specified times.

- **Web Spiders or Crawlers: Taking Everything**. These tools typically harvest all the data on a site, resulting in a huge flow of undifferentiated information that requires significant post-processing and data transformation to be usable. In addition, they often leave very heavy footprints on a site, disrupting website operations and prompting sites to try to block them.

- **Programmers: Multiple Pages, Using Scripts**. Programmers write scripts that extract data from particular websites. This approach is expensive because programmers who can write such scripts are hard to find and relatively expensive. In addition, scripts are very brittle, and tend to stop working even with minor website changes. Some customer using scripts report that they spend so much time fixing existing scripts that they lack the resources to approach the new sites they're being asked to harvest. And to get data from highly dynamic sites (using Javascript or Ajax), scripts must be very complex. In short, this approach is high cost, often unreliable, and limits the number of new sources that can be accessed.

- **Platforms: Extraction at Scale**. Businesses that rely on high volumes of website content as input for their data products often turn to powerful platforms that combine scale, automation, efficiency, and high data precision and quality. The platforms have been designed to create a highly robust and flexible means of acquiring website content, as well as to simplify and improve companies' downstream data product manufacturing processes by reducing personnel and processing costs, maintenance requirements, complexity, and data volumes. Such platforms may require upfront investment, but ultimately prove their worth by providing easier access to more sites and more content, by reducing operating costs, and by allowing companies to bring to market a broader array of data products at a faster pace.

According to a survey by Connotate, **75%** of businesses seeking to use website content are **dissatisfied** with the tools they're using to aggregate it.

Extracting And Using Website Content At Scale: Best Practices

Once companies conclude that their current methods are insufficient for their needs, they should examine approaches and technologies used by leading information product companies to determine best practices for creating a robust means of acquiring website content and feeding it into their complex operations. By and large, these businesses have adopted automated approaches with the following characteristics:

- **Scalable**. Their platforms have the ability to extract high volumes of data, from many different sites, swiftly and without glitches, no matter how frequently content or the site itself changes.

- **Cost-effective**. Their platforms reduce costs and improve efficiencies from acquisition all the way through the downstream data products manufacturing process. Industry-leading platforms deliver cost savings in labor (e.g., by automating extractions, by reducing the time it takes to build extraction routines and allowing non-programmers to build Agents); reduction in complexity (robust pre-processing of data and superior site change detection that reduces overall incoming data volumes and post-processing requirements); lower maintenance and less Agent breakage; and more.

- **Automated**. The most advanced solutions automate many of the operations that must be manually performed in less sophisticated approaches. This includes automated scheduling of harvests at the specific site level; monitoring of harvesting operations; alerts with reason codes if there are any issues; automated QA, ongoing change detection and change alerts, and seamless delivery into a wide range of different databases.

- **Integrated**. The best solutions easily transform unstructured website content into a smooth flow of structured, normalized data. This data flow is aggregated and turned into a form that can be easily ingested by existing systems and workflows to maximize its value and timeliness (see "Incorporating Website Content into Data Flows" for an example).

- **Resilient**. The extraction routines continue to work even with minor website changes.

- **Flexible deployment**. Solutions can be offered on-premise or as a managed service, with professional services available to accelerate time to value.

Incorporating Website Content into Data Flows

Extracting data is really only half the battle. The other half is getting that content into a normalized, clean format that can be incorporated seamlessly into analytical processes and data flows.

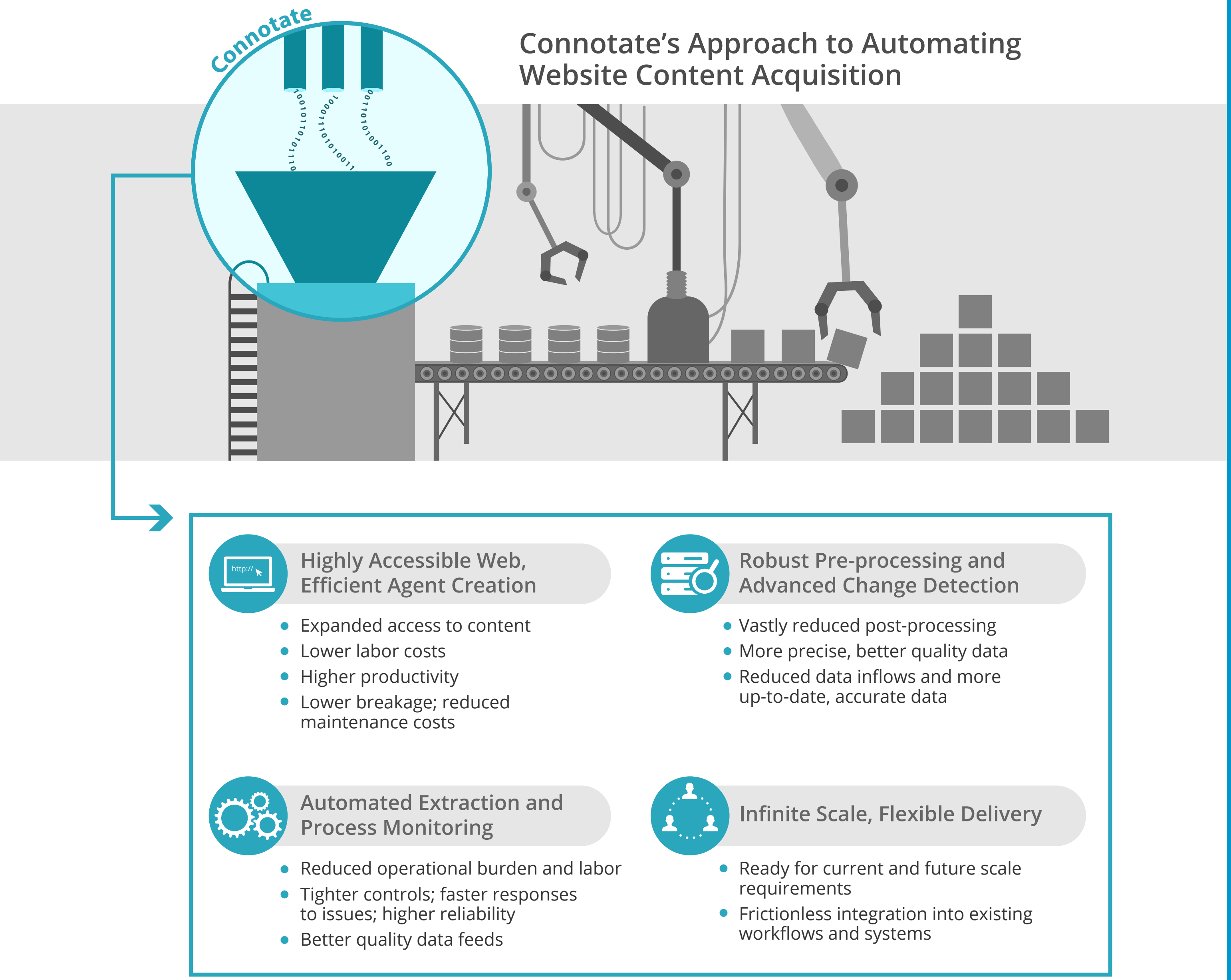
For example, one financial services firm integrates website content with Salesforce for sales intelligence. The firm finds and vets leads by reviewing changes in an organization's legal status that are posted on its website. Leads from Salesforce are exported to the platform in a CSV file at night. The platform verifies the legal status of each lead and submits an updated CSV file to Salesforce, which is imported before sales reps arrive at work. Sales reps obtain updated leads daily, and those updates include information that enables better-informed conversations with prospects at the optimal time. By integrating website content with other systems, the firm was able to double the productivity of its sales team.

The takeaway: platforms for website content extraction should be vetted not only on their ability to acquire data, but also on their ability to normalize that data and enable its use with existing systems, to make it part of an organization's entire production flow.

Connotate: Crafted For High-Scale Website Content Harvesting

While there are a variety of options on the market for companies looking to automate website content acquisition, few offer a platform with all of the aforementioned best-practice characteristics. Connotate is an exception. It offers as an industry-leading website content harvesting platform as well as managed data services such as custom datapipes that they build and manage on behalf of their customers.

Connotate's Approach to Automating Website Content Acquisition



Highly Accessible Web, Efficient Agent Creation

- Expanded access to content
- Lower labor costs
- Higher productivity
- Lower breakage; reduced maintenance costs

Robust Pre-processing and Advanced Change Detection

- Vastly reduced post-processing
- More precise, better quality data
- Reduced data inflows and more up-to-date, accurate data

Automated Extraction and Process Monitoring

- Reduced operational burden and labor
- Tighter controls; faster responses to issues; higher reliability
- Better quality data feeds

Infinite Scale, Flexible Delivery

- Ready for current and future scale requirements
- Frictionless integration into existing workflows and systems

Connotate relies on advanced, patented machine learning technology to allow users to get the most out of the Internet's abundance. Connotate users create Agents that extract content from Web pages. Users train the Connotate Agents, which create a website content extraction model. An example data set of one or two pages is usually enough to train an Agent to create an accurate model. No programming expertise is required.

To get an idea of the cost and operational benefits that Connotate delivers, consider the resources required to create and maintain a data flow of robust content from 10,000 websites. Using manual methods, this would take 60 to 80 full-time researchers. Given the high breakage rate of script-based approaches, it would take 6 to 8 programmers to continually fix the non-performing scripts. With Connotate, it typically takes 2 non-programmers to manage the ongoing harvesting.

Connotate offers a robust website content harvesting platform specifically geared to handle the extreme volume, velocity, and complexity of the big data universe. The platform can extract data buried deep within a site or on multiple sites, whether information is hidden in an overlooked PDF or is an amalgamation of trends from hundreds of pages.

One of Connotate's customers runs over 500,000 Agents. Companies find that they can extract increasing volumes of relevant data with no increase in staff. Typically a very small team of non-programmers use Connotate's point-and-click interface to create all the Agents needed by a company.

Agents are robust; they continue to perform through many site changes that would stymie a script. Further, they return precisely the data that's requested and leverage change detection to return only what's changed on a site. Users can automate the schedule on which Agents run, so the task of returning the latest data is seamlessly folded into workflows.

Connotate pays particular attention to data quality, as evidenced by it 2014 patent for quality assurance on data flows. It watches data as it flows to see what it looks like and how it is related to other data. When a deviation in this pattern occurs, it can inform the extraction engine to compensate for the anomaly or to raise an exception to be reviewed by a user.

A key differentiator for Connotate lies in the area of operationalizing data. Getting the data is half the battle; leveraging it to create data products or incorporating it into workflows is the other half. The platform not only harvests website content; it also performs pre-processing normalization at runtime that transforms it into the format users need to put the data to work.

Companies find that they can extract increasing volumes of relevant data with no increase in staff

Connotate offers two deployment options. Its platform can be licensed for on-premise use. If companies prefer, Connotate or one of its trusted partners will provide managed data services, building and managing a custom datapipe and delivering just the data stream.

Conclusion

Valuable information is locked in websites in unstructured formats, making it difficult to extract at scale and transform into normalized, machine-ingestible data. To gain competitive advantage or create data products from website content, companies must employ harvesting platforms that work at scale, can be leveraged by non-programmers, and ensure easy integration of clean, normalized data into downstream operations and applications. The platform must be resilient so that when websites change, content continues to be extracted. It must be efficient, identifying changes to extracted content. The deployment model should be flexible, allowing companies to choose whether they need an on-premise solution or managed data services. Before adopting any approach, companies should ensure the solution has these benefits.

Connotate meets all of these requirements, with advanced, sophisticated, and patented machine learning and pattern recognition. It is specifically designed for data-driven processes and businesses and the complex ways they extract, ingest, and analyze large data flows. It not only enables companies to effectively get all the content they need, but it also delivers efficiencies and lower complexity throughout downstream operations.

In the world of high-scale harvesting of website content, many advances and innovations are just around the bend. Soon, machine learning platforms like Connotate will be able to learn in-depth from sites so that when they are pointed at similar sites, they can auto-create a new Agent. Companies should get started now in strategizing how they will capitalize on all the power website content can provide and how they can make such data a key part of their operations and data monetization efforts.

This paper was created by CITO Research and sponsored by Connotate

To learn more about Connotate's technology and managed data services for enterprise scale web harvesting platform, visit **www.connotate.com**, or call **732 296 8844**.

Follow: in 🐦

CITO Research

CITO Research is a source of news, analysis, research and knowledge for CIOs, CTOs and other IT and business professionals. CITO Research engages in a dialogue with its audience to capture technology trends that are harvested, analyzed and communicated in a sophisticated way to help practitioners solve difficult business problems.

Visit us at http://www.citoresearch.com

Connotate

www.connotate.com

info@connotate.com

732 296 8844