Web Data Integration –
Leveraging the Ultimate Dataset

Anna Griem
Octavio Marenzi

January 2019

opimas

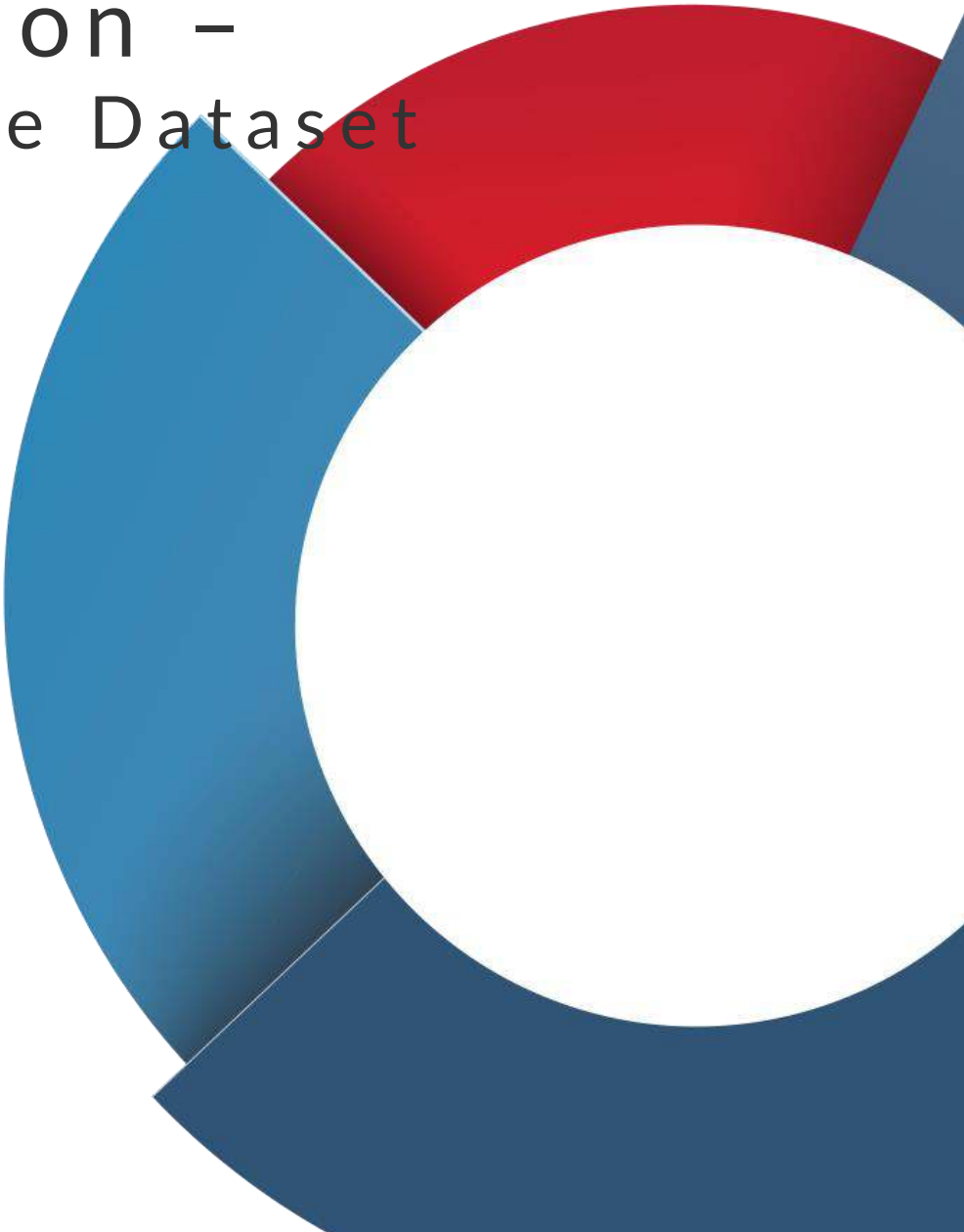
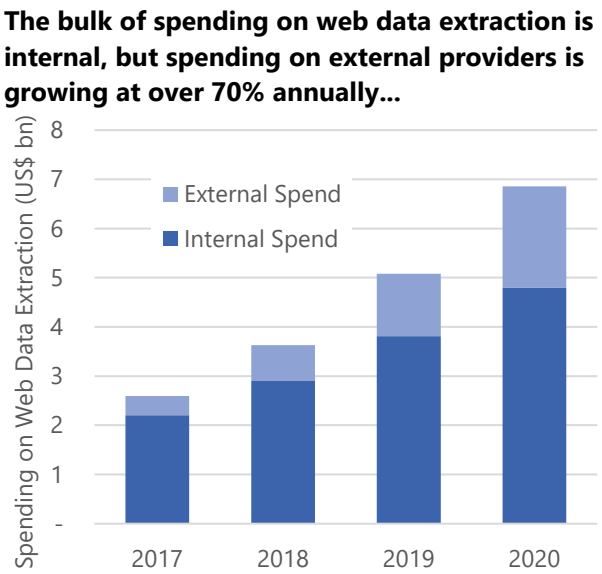


TABLE OF CONTENTS

EXECUTIVE SUMMARY

FIGURE 1. INTERNAL AND EXTERNAL SPENDING ON EXTRACTING WEB DATA



Source: Opimas Analysis

The web is the ultimate dataset. A rapidly increasing amount of data on the web, ever-more connected devices, and growing social media usage are creating more sources from which valuable insights can be gleaned. Web data comprises a valuable portion of alternative datasets revolutionizing decision making for corporations. Extracting these data from the web is a complex endeavor, with the information required to perform analyses spread across multiple sites, in different, often unstructured, formats.

The range of use cases for web data integration is rapidly increasing, and with it the necessary investment. While spending on this area amounted to about US$2.5bn in 2017, we expect that by 2020 the market will reach almost US$7bn. The bulk of this spending is currently weighted towards internal, home-grown systems. However, with the complexity of creating and maintaining web data extractors and preparing that data for consumption by applications and analytics platforms, spending is increasingly shifting to specialized external technology and service providers. External spending is set to increase from about US$400mn in 2017 to over US$2bn in 2020 (see Figure 1).

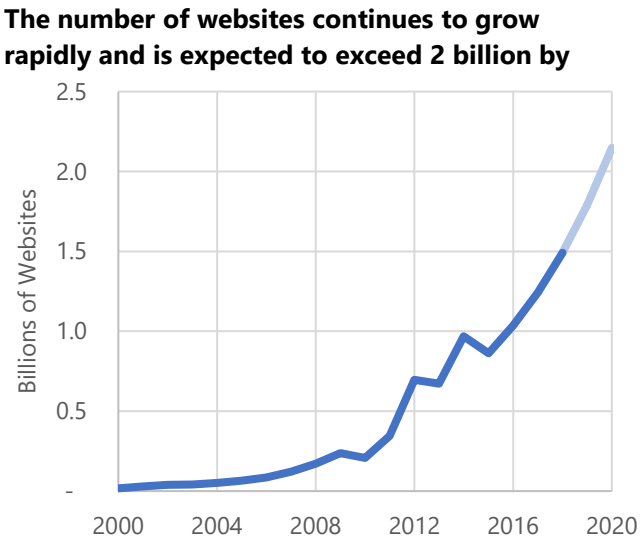
While several technical (anti-webscraping defenses) and business risks exist, these appear to be mostly manageable.

Over the coming years, making sense of big data and gleaning value will be a priority. We will see a rapid growth in spending and value extracted from the web using external tools, especially in investment decision-making, e-commerce, and manufacturing.

INTRODUCTION

SIZING THE INTERNET

FIGURE 2. NUMBER OF WEBSITES

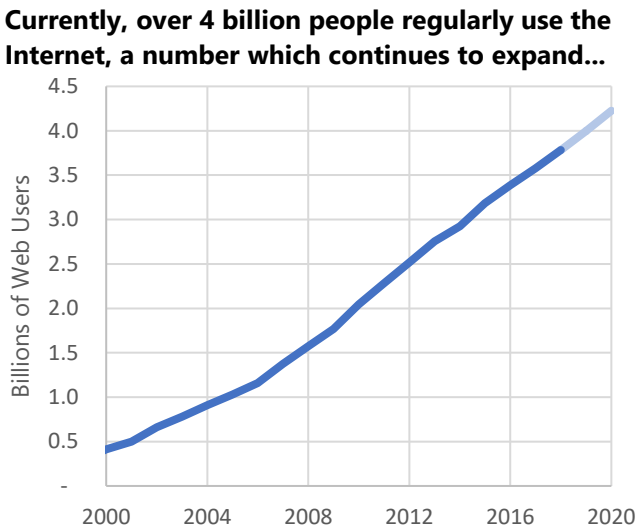


Source: Internet Live Stats, Opimas Analysis

Firms are increasingly turning to the web as a source of data to support their decision-making processes and strengthen their services. Total web content has grown at a breakneck rate since the late 1990s and will continue this pace and direction for the foreseeable future – by 2020 we expect that there will be over two billion web sites (Figure 2). There has also been a steady growth in the number of web users (Figure 3), accompanied by a proliferation of e-commerce sites and other web-based services. Connected devices, the much-heralded Internet-of-things, continue to thrive and beget a wide variety of applications and services.

Machine learning systems, increasingly used as firms look to automate processes where humans are not essential or best equipped, are hungry for datasets for their algorithms to be appropriately trained and tuned.

FIGURE 3. INTERNET USERS

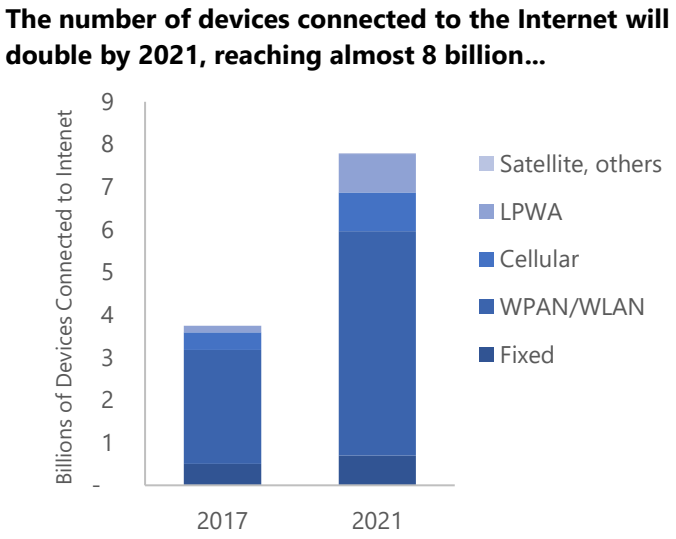


Source: Internet Live Stats, Opimas Analysis

The total amount of data lurking on the web will continue to mushroom, and firms who do not successfully harness this source will quickly be left behind by savvier competitors. Firms can also leverage their web data integration capabilities to reduce risk, using the web for better identification and screening of customers, suppliers, and distributers.

At the enterprise level, automation and quality control are integral to even unsophisticated data extraction efforts. High-quality web data integration enables the speedy and repeatable automation of website data capture and aggregation – essential for enterprises looking to employ data from the web at scale or for critical business functions.

FIGURE 4. INTERNET-OF-THINGS



Source: Statista

OVERVIEW OF WEB DATA INTEGRATION

Extracting and utilizing data from web pages is not a simple task, and often requires specialized expertise. An overview of the web data integration process is shown in Figure 5. Significant resources have been applied to extracting unique data from the web, but this endeavor has long been heavily reliant on home-grown, often piecemeal methods. Do-it-yourself approaches are costly to maintain at scale, are complex, and are revealing themselves to be insufficient at handling the demand to harness meaningful web content and data. Enterprise-level web data integration projects vary widely in purpose, scope, and appropriate technological approach. Free or low-cost tools are likely sufficient only for sporadic, non-mission critical one-time projects run by an engineer or technical business analyst.

FIGURE 5. WEB DATA INTEGRATION PROCESS OVERVIEW

Identify
- Identification of web sites and pages
- Analysis of page layout
- Definition of data to be extracted and its structure
- Determination of update frequency

Extract
- Field definitions
- Creation of extractors
- Download of HTML and other file formats
- Circumvention of anti-web scraping defences
- Identification of extractors not working due to web page layout changes

Prepare
- Monitoring of successful and unsuccessful downloads
- De-duplication of data
- Verification of outliers
- Derivation of data
- Data quality reports

Integrate
Delivery of data via
- Alerts
- Structured datasets
- API
- CSV, Excel or other formats
- JSON or SQL table updates

Consume
- Input into statistical models or machine learning
- Reports, graphs
- Alert management
- Feed into business intelligence tools
- Visualization

Source: Opimas Analysis

For more ongoing and strategic projects, enterprise-level web data integration providers help their clients run web data integration projects where update frequency, number of websites, level of anti-scraping navigation, regulatory compliance, quality and uptime requirements, and the ultimate delivery of the data is more secure and complex. The importance of uptime cannot be underestimated for integration relating to critical business functions. Firms must determine how many hours they can function without their web data integrations operating correctly before it causes serious problems. For e-commerce sites, even a few hours without appropriate adjustment of pricing could result in significant losses.

As select solutions are now delivering a true enterprise service where security, reliability, regulatory compliance, content coverage, and ease of use are promised, more firms are aiming to complete strategic web data integration projects (

Figure 6). Many firms run several concurrent and very complex web data integration projects across their divisions. The largest financial institutions, for example, commonly spend well in excess of several million dollars annually on web data integration across various lines of business and activities. Much of this spending is shifting towards external solution providers which, in many cases, is enabling more complex web datasets to be integrated at lower build and maintenance costs.

Especially across organizations active in retailing, financial services, manufacturing, and market research, applications of large-scale web data integration projects abound. We consider some of the most common use cases in a separate chapter.

Everywhere one looks, the demand for high quality data is increasing. Robust commercial web data integration solutions geared to the needs of enterprises are positioning themselves to help deliver this data in an actionable format.

FIGURE 6. LEAST TO MOST COMPLEX USE CASE OF WEB DATA INTEGRATION TECHNOLOGY

	Simplest case	Most complex case
Web page structure	Static web pages Web page URLs known ahead of time Simple HTML and CSS	Dynamic website Content is spread across multiple pages Web page URLs cannot be known ahead of time Asynchronous loading of content using JavaScript or other techniques ASP sites, files on sites, images, PDFs
Update frequency of the structure of the target web site	Annual or more	Weekly
Update frequency of the content of the target web site	Monthly or less	Real time
Number of web pages	Under 10	Low millions
Number of web sites	1	Hundreds or more
Number of pages per month	< 100	100s to 100,000,000s
Anti-scraping defenses	None	CAPTCHA, password, IP address blocking, user agent blocking, region blocking, others
Delivery	CSV file	Structured tables, APIs, custom integration, SQL updates, XML, insert into business intelligence, statistical models, and machine learning applications
Quality control	Little or none	Timely completion of run Content verified Extraction is corrected based on web site changes Real-time monitoring of extraction success metrics Quality reporting and tracking
Typical external spend	$0	> $100,000
Percentage of web pages available	50%	99%
Appropriate tool	Self-service tools, home-grown development projects	Enterprise-class solution provider
User	Engineer with coding experience	Business analyst or manager - technical skills not required
Data format	Same format as on the web site	Transformation - e.g. common schema, convert data, currency conversion, split/concatenate fields, perform calculations
Localization of browser	Single location	Geographic specific- web sites often show different information by location
Data security and privacy	Not important	Very important
Avoid collection of PII data	Not important	Very important
Time sensitive extraction	Not important	Very important e.g. competitive pricing
Data storage and IT infrastructure	User firm provides	Enterprise-class solution provides

Source: Opimas Analysis

BEYOND WEB SCRAPING

The term "web scraping" is often used to describe the process of collecting data from the web, which we have outlined in the previous chapters. However, there is an important distinction to be made between *web scraping* and *web data integration*. Web scraping activities tend to revolve around in-house attempts to extract data from the internet, typically relying on free or low-cost tools, combined with extensive proprietary system development in Python or other programming languages.

Web data integration, on the other hand, involves the use of specialized external service providers. Frequently, firms embarking on web scraping initiatives quickly find that the complexity and cost of creating and maintaining their systems overwhelms the benefits derived. As a result, we are increasingly seeing situations where internal development has been tried, with some initial success, only to see that internal web scraping teams start to struggle with the complexity of extracting and transforming data, maintaining and ensuring data quality, and reacting to growing demands from business users and data analysts.

In Figure 7, a comparison of the most salient differences between web scraping and web data integration can be seen.

FIGURE 7. WEB DATA INTEGRATION VS. WEB SCRAPING

	Web scraping	Web data integration
Data quality	Poor quality, brittle connections	High quality and reliable
Resource requirements	Instensive, with specialized skills	Low, primarily requirements definition
Business risk	High	Low: suppliers anonymize & indemnify
Data integration	Typically flat files, need for manual integration or an ETL tool	Direct integration, ready for use
Time to value	Slow	Fast
Maintenance	High cost and time consuming	Automatically integrated into solution
Scalability	Poor and fragile	Highly scalable

USE CASES

FIGURE 8. COMMON USE CASES

Source: Opimas Analysis

Data extracted from web pages can help companies give context to why an event occurred, predict what is going to happen, and to be prescriptive about how to best react to future events. Motivations for web data integration at the enterprise level are plentiful, and some of the most popular use cases are shown in Figure 8. It must be pointed out, that this is far from being an exhaustive list, and hundreds, if not thousands, of uses for web-extracted data exist. In the following section, we explain and describe some of the most common applications of web data integration.

SENTIMENT ANALYSIS

Firms often extract information from social media platforms and news feeds to monitor their firm's or products' reputations, keep track of competitors, and stay abreast of any breaking stories that may be of importance.

After this information is harvested, natural language processing enables detection of correct entities – e.g. "apple" the fruit versus "Apple" the company – and measures general feeling of the public towards this

entity. This type of analysis is regularly used in investment decision-making. Recall the oft-cited correlation between the actress Anne Hathaway receiving positive press and Warren Buffett's Berkshire Hathaway's shares increasing in value. It is entirely possible that sentiment analysis of Hathaway played a role.

PRICE COMPARATORS

Competitor pricing and inventory monitoring has become especially important with e-commerce sites competing for clients by changing their prices several times per day, perhaps even by the hour or minute. The extraction of rivals' pricing helps companies appropriately adjust the going rate of their own products. Making a mistake on this front, for even just half a day, could result in significant losses.

Websites like Amazon.com have upped the ante for selling consumer goods online, and firms selling products via the Amazon marketplace must constantly monitor the website to ensure their products are prominently featured or make adjustments.

Wholesalers also use web data integration to monitor their retail distributors' prices to ensure compliance with their pricing policies.

SUPPLY CHAIN MANAGEMENT

Web data integration is frequently used by corporations looking to manage logistics, procurement, and to appropriately forecast demand. For example, when an airline is planning routes, they must consider travel demand and competing airlines servicing those routes, including capacity and pricing. Accessing these data using a web data integration solution is effectively the only way that an airline can monitor the market and adapt its pricing and routes.

In manufacturing, a firm may also benefit from speedy tracking of raw goods and weather patterns, depending on the product they create. A coffee seller may decide to increase an order of beans from Uganda if a drought looks likely in Mexico and Guatemala.

ACCOUNT AGGREGATION

Firms, and even retail customers, frequently have accounts with multiple service providers, such as

financial institutions. Obtaining a consolidated view requires the aggregation of multiple accounts. While this is occasionally possible through APIs, web data integration is applicable in a broader set of scenarios.

Services like Mint.com provide its clients with overviews of their spending behaviors by creating visualizations of their clients' financial data, extracted with permission from their clients' bank accounts. This service helps individuals understand their financial health and improve spending habits.

MARKETING & SALES

Web data integration enables more informed marketing and sales planning and decision making. Data extraction helps to better target campaigns and monitor the effectiveness of marketing efforts. Gathering information about competitors, optimizing search engine result placement, and identifying sales leads are all critical to a company's survival.

Analyzing information harvested from the news, social media, and reviews can also draw attention to issues requiring customer service and support, as well as influence product development and market entry strategies.

Web data extractors can also be used in lead generation, identifying individuals or companies that could be targets for specific services or products.

ECONOMIC & INVESTMENT RESEARCH

Alternative data has become a widespread buzzword, but was born in capital markets. Large quantitative hedge funds began feeding signals from web datasets into their statistical models to improve fund performance several years ago. Reliable web data integration that produces data fit for normalization and analysis, especially when delivered via API directly into the firm's models, can power investment decisions near real-time.

Investment firms use datasets measuring mobile locations, traffic patterns, weather, satellite imagery, financial statements, macro-economic indicators, among many others to help shape investment decisions.

Market data aggregation is also employed widely outside of direct investment decision-making. For industries under-served by market research firms, accessing their own indicators from publicly available web sources may be their only opportunity to gain critical insights impacting future planning.

RISK MANAGEMENT

Beyond reputation monitoring, many risk management measures can be enhanced with web data integration. Web data can be used to complete background checks on employees and suppliers, and to manage counterparty risks. These tools can also be used to gather information for know-your-customer purposes.

Web data can also be used to monitor for geopolitical risks and to detect fraud. An insurance company may be able to detect, for instance, if a person receiving workers compensation has just spent the weekend kitesurfing, and shared photos with their social media followers. Compliance monitoring and reporting is aided substantially by the ability to automate the gathering of dispersed data across the web.

OTHERS

Other potential use cases abound, and are often very creative, relying on unusual and innovative applications of web data. We expect the range and scope of use cases to rapidly increase in coming years, as firms become more adept with the technology and more sources of data appear and are discovered on the web.

MARKET SIZE

In this chapter we examine the overall size of the market for web data integration. Total spending on web data integration will rapidly increase over the next three years, from US$2.6bn in 2017 to nearly US$7bn in 2020 (Figure 9). We break this spending down along several variables, including internal versus external spending, regional variations, as well as by industry and use case.

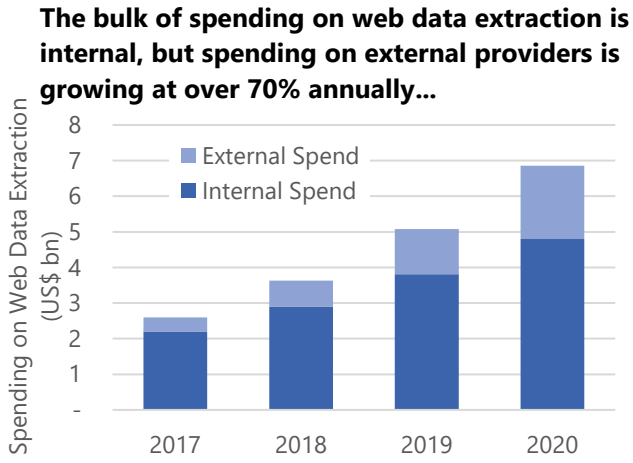
INTERNAL VERSUS EXTERNAL

Currently, the bulk of spending is internally focused, with firms using a combination of manual processes and their own IT solutions. In the coming years, we believe firms will increasingly seek to outsource this activity to improve automation and reliability of their web data integration efforts.

There are several factors driving the outsourcing of this function. First, many firms who have attempted to build complex web data integration systems have found that they do not have the necessary expertise to create industrial-strength systems. This is particularly true of the more complex projects outlined in Figure 6. Here we frequently see home grown solutions that are unreliable, and unable to access all the desired data sources. Secondly, the maintenance and quality control costs for homegrown systems are often higher than the original build costs, putting a strain on budgets.

As a result of these drivers, we expect internal spending, while still growing at an annual rate of over 30%, to lag considerably behind the growth of external spending, which will exceed 70% annually for the next few years.

FIGURE 9. INTERNAL AND EXTERNAL SPENDING ON WEB DATA EXTRACTION AND INTEGRATION

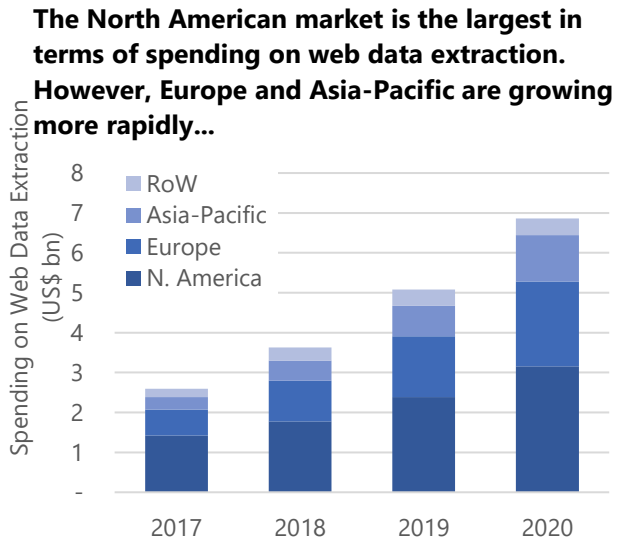


Source: Opimas Analysis

REGIONAL BREAKDOWN

The North American market is the largest currently, accounting for more than half of overall spending in 2017 (see Figure 10). In e-commerce, many of the largest users of price comparators using web data are US-based, while in capital markets New York City's large quantitative hedge funds, have led the charge in terms of using web data for investment decision making. However, spending levels in Europe and Asia-Pacific will increase more rapidly in coming years, as the business value of web data integration is better understood. Germany, France, and Belgium have led Europe's spending on web data integration for last few years. Japan and China are leading Asia-Pacific.

FIGURE 10. REGIONAL SPENDING ON WEB DATA EXTRACTION AND INTEGRATION



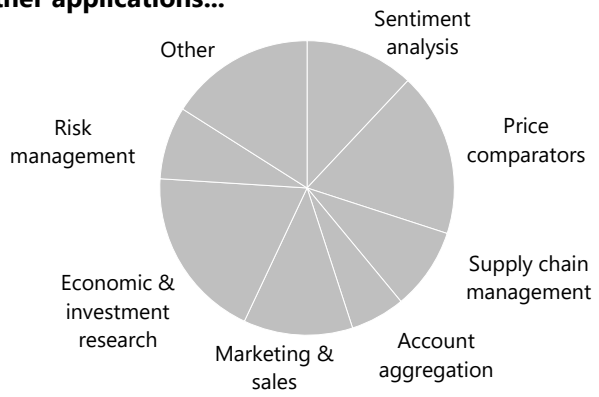
Source: Opimas Analysis

INDUSTRY AND USE CASE

In Figure 11 and Figure 12 the spending on web data integration can be seen broken down by use case and by industry. The key take-away from these figures is that the market for web data integration technology is underpinned by broad usage across many industries and use cases, without a particular concentration.

FIGURE 11. SPENDING BY USE CASE

While price comparisons and investment research are the most common use cases, there are numerous other applications...



Total Spending on Web data extraction and integration 2017: US$2.6bn

Source: Opimas Analysis

FIGURE 12. SPENDING BY INDUSTRY

Finance and services are the largest users of web data integration, but spending is broadly spread across other industries...

Total Spending on Web data extraction and integration 2017: US$2.6bn

Source: Opimas Analysis

POTENTIAL IMPEDIMENTS

Web data integration is becoming more and more widespread, but several challenges must be piloted carefully. Anti-webscraping technologies, policies, and regulations must all be heeded with caution.

REGULATION

A variety of regulatory obstacles must be navigated in web data integration. Most notably, the enforcement of the European Union's General Data Protection Regulation (GDPR) began in May 2018. This requires organizations to follow strict rules in their handling of personally identifiable information (PII). Personal data, in the context of GDPR, includes name, address, birth date, national identity numbers, and financial information, but potentially also genetic information, social media posts, photographs, preferences, transaction histories, and IP addresses. This personal information can only be used on an opt-in basis - meaning that any firm using PII must receive permission, for each desired use, from the person in question.

Firms under the jurisdiction of GDPR need to avoid mishandling PII. Sophisticated enterprise solutions do offer protections on this front by ensuring avoidance of collecting PII in the first place, or by masking the data when it is collected. Beyond the threat of lawsuits from individuals whose PII is mishandled, firms should consider the steep fines GDPR threatens. For a single infraction, firms may face fines up to €20 million or 4% of annual global turnover, whichever is greater.

JUDICIAL RULINGS

Recent court cases examining web data integration shed light on the tide of opinion. In August 2017's *hiQ v. LinkedIn*, a judge ruled in favor of hiQ Labs over LinkedIn (who had been acquired by Microsoft the previous year). hiQ Labs' service offers its corporate clients "people insights" to aid in managing their workforces, including analysis of employees at risk of turnover based on their LinkedIn activity.

The core question in the *hiQ v LinkedIn* case is whether LinkedIn could claim that hiQ was "hacking" LinkedIn's computer systems by accessing LinkedIn's publicly available web pages using an automated web browser and without LinkedIn's consent. LinkedIn claimed hiQ Labs was violating the United States Computer Fraud and Abuse Act (CFAA) of 1986. The judge dismissed LinkedIn's claim that it could revoke a user's "authorization" to view its public website regardless of how the website was accessed, "a user does not 'access' a computer 'without authorization' by using bots, even in the face of technical countermeasures, when the data it accesses is otherwise open to the public".

Oracle v. Rimini centered on a related issue. Oracle sued Rimini Street, a software support firm that services Oracle customers amongst many others, for automating the extraction of Oracle technical support documents from the Oracle website. They also accused Rimini Street of copyright infringement. The terms of use on Oracle's website prohibited the automated downloading of these files. While a 2012 ruling initially held in Oracle's favor on both counts, Rimini Street came out on top regarding the extraction issue in a 2017 appeal. The appeals court decided that, as Rimini Street did have permission to access the files in question, the manner in which accessing occurs is not material, even though automated access was explicitly prohibited by the Oracle website's terms of use.

In both cases, the recent court decisions similarly decided that web content providers, in this case LinkedIn and Oracle, do not have the right to allow access to materials and then seek relief when materials are accessed in a manner they do not agree with. Oracle and LinkedIn faced rival services by Rimini Street and hiQ respectively, and both judges implied in their rulings that Oracle and LinkedIn were pursuing Rimini Street and hiQ Labs to stamp out competition, rather than the way they accessed the information in question. It should be pointed out that Rimini Street is still facing steep fines for copyright issues related to this case.

While these cases signal successes for web data integration, caution is advised. In the United States, the Better Online Ticket Sales Act of 2016 restricting web extraction of ticket sales platforms, ruled that "violations shall be treated as unfair or deceptive acts or practices under the Federal Trade Commission Act." The legality of specific actions in this space is still being decided.

Overall, recent lawsuits in the United States related to web data integration have fallen in favor of the practice rather than against it.

ANTI-WEBSCRAPING DEFENSES

Beyond terms of service prohibiting automated browsing of website – which US courts have largely rejected – some companies use technical defenses to make their web pages difficult to automatically browse. These defenses include password protection, CAPTCHAs, misdirection via delivery of faulty data, user agent blocking, IP address blocking, and IP region blocking. These defenses are put in place for a variety of reasons: putting web pages behind password protection means that those pages are only available to registered users; the use of CAPTCHAs and IP restrictions are often used in order to combat deliberate Denial of Service (DoS) attacks. Many of the defenses described above can be legitimately navigated by sophisticated web data integration solutions.

Another way in which data extractors can be thrown off course is when even simple changes are made to the display of content on a website. Enterprise-level web data integration solutions constantly monitor for such changes and update their processes accordingly. Firms looking to engage with such a solution provider should ensure that it is equipped to navigate these anti-scraping technologies and deterrents. They should also feel confident that their solution provider is able to detect changes to website content that throw a scraper agent off course and fix it within an appropriate timeframe.

Anti-scraping technologies like ShieldSquare and Distil Networks sell software designed to block or misdirect web data extractors. Leading web data integration providers claim to be able to navigate around most of these.

LOOKING FORWARD

Although some impediments to web data integration exist, the hunger for actionable data is immense, and the overall picture for this industry looks very positive. The world is teeming with alternative datasets that can be created from geolocations, satellite imagery, and other esoteric sources. The web however, is the ultimate dataset. Web data integration technology enables the organized collation of these dispersed points of interest. High quality, enterprise-level web data integration platforms take their clients one step further: analysis.

Market research firms have long sold datasets that firms and certain industries had difficulty creating on their own. Large quantitative hedge funds have invested significantly in their ability to feed algorithmic trading by hiring large teams of quants. Companies selling unique datasets are flooding the market, as are aggregators of such datasets. Select web data extractors have also thrown their hats into the ring in order to take their clients beyond the collection of data. Only a handful of firms like Import.io, Sequentum's Content Grabber, and Competitive Analytics provide additional analytical capabilities. Beyond normalizing data and delivering it in the desired format, they are also able to provide their clients with insight regarding what the data means for them.

The market for web data integration is growing quickly, driven by the hunger of the market for customized, actionable data.

The market for web data integration is growing quickly, driven by the hunger for customized, actionable datasets. Large quantitative hedge funds will want to complete the lion's share of analysis in-house in order to keep their recipes secret, but the long tail of firms with less quantitative capabilities will still want quality data that produces actionable insights. They will be willing to outsource analysis. As the spending on web data integration for the building of customized datasets shifts to outside solution providers, those web data integration solutions that are *also* able to provide analytics will be particularly well positioned to meet this need.

About the Report Authors

Octavio Marenzi, founder and CEO of Opimas LLC, leads the firm's asset management, and equities trading practices and advises many of the world's major financial institutions on their business, financial, operations and technology strategies.

Anna Griem is a senior research analyst focused on financial regulation and regulatory technology at Opimas.

About Opimas

Opimas is a management consultancy focused on capital markets, serving leading financial institutions around the world. Our specialisation allows us to bring our expertise to bear from the very beginning of projects, to provide insight and craft strategies for our clients more quickly, without sacrificing quality. In addition, Opimas continuously invests considerable resources, representing about one third of our revenues, in market research. This investment allows us to create a pool of intellectual capital on issues of strategic importance to our clients.