Web Data Extractors 2018

A White Paper Link Compilation

By

Marcus P. Zillman, M.S., A.M.H.A.
Executive Director – Virtual Private Library
zillman@virtualprivatelibrary.com

Extracting data from the World Wide Web (WWW) has become an important issue in the last few years as the number of web pages available on the visible Internet has grown to billions of pages with trillions of pages available from the invisible web. Tools and protocols to extract all this information have now come in demand as researchers as well as web browsers and surfers want to discover new knowledge at an ever increasing rate! As robots (bots) and intelligent agents are at the heart of many extraction tools I decided to create a compilation of the latest sources and sites that extract information from the web.

Web Data Extractors 2018:

80legs - Powerful and Economical Service Platform for Crawling and Processing Web Content
http://www.80legs.com/

Agenty – Hosted Web Scraping Tool
https://www.agenty.com/

Anthracite
http://freecode.com/projects/anthracite

Apify – Web Scraping Platform for Coders
https://www.apify.com/

Aristo - Answer Questions with a Knowledgeable Machine
http://allenai.org/aristo/

artoo.js - The Client-Side Scraping Companion
http://medialab.github.io/artoo/
AutoMate - Automate Data Extraction
https://www.networkautomation.com/

Automated RSS Scraper Scripts
http://www.djeaux.com/rss/

Automated Information Solutions
http://www.automated-info-solutions.com/

Automatic Information Extraction From Semi-Structured Web Pages By Pattern Discovery
http://portal.acm.org/citation.cfm?id=640423&dl=ACM&coll=portal

Beautiful Soup
http://freecode.com/projects/beautifulsoup

Beautiful Soup - HTML/XML Parser for Quick Turnaround Screen Scraping and Web Data Extraction
http://www.crummy.com/software/BeautifulSoup/

Big Data: What Is Web Scraping and How To Use It
https://itsvit.com/blog/big-data-what-is-web-scraping-and-how-to-use-it-it-svit-blog/

BLIASoft Knowledge Discovery
http://www.bliasoft.com/Eindex.html

Bot Research 2018
http://www.BotResearch.info/

BYU Data Extraction Research Group
http://www.deg.byu.edu/

Captiva Software: Digital Information Capture Software
http://www.emc.com/enterprise-content-management/captiva/captiva.htm

ChartSearch Data Search Technology
http://www.ChartSearch.net/

Client-Side Deep Web Data Extraction
CloudScrape – Extract, Enrich and Connect
http://www.cloudscrape.com/

Cogitum Co-Citer
http://www.cogitum.com/co-tracker-text/more.shtml

Common Crawl
http://www.commoncrawl.org/

Connotate – Web Data Extraction and Monitoring
http://www.connotate.com/

Content Grabber – Extract Data from Websites
http://www.ContentGrabber.com/

ContextMiner - Tools to Collect Data, Metadata and Contextual Information
http://www.contextminer.org/

cQuery - Content Query Engine
http://cquery.com/

CrawlMonster
http://www.crawlmonster.com/

Crawly
http://crawly.diffbot.com/

Create a Crawler - Extract Data From an Entire Website
https://www.import.io/

cURL groks URLs - Command Line Tool for Transferring Data
http://curl.haxx.se/

Data Extraction Services
http://www.dataextractionservices.com/

DataHen – Advanced Web Scraping and Data Extraction Services
https://www.datahen.com/

Data Mining Resources 2018
http://www.DataMiningResources.info/
Data Miner – Extract Data From any Website in Seconds
https://data-miner.io/

Dataminr - Real-time Information Discovery
http://www.dataminr.com/

Data Scraper – East Web Scraping with Google Chrome

DataSift - Powerful Social Data Platform
http://datasift.com/

Data Toolbar – Web Data Extraction Software Made Simple
http://datatoolbar.com/

DataWatch Monarch – Self-Service Data Preparation
http://www.datawatch.com/

DataWrangler - Data Cleaning and Transformation Tool
http://vis.stanford.edu/wrangler/

Deep Web Research 2018
http://www.DeepWebResearch.info/

DEiXTo – Powerful Web Data Extraction Tool Based on W3C DOM
http://deixto.com/

dexi.io – Web Data Processing for Professionals – Extract, Enrich and Connect
https://dexi.io/

DiffBot – Web Data Extraction Using Artificial Intelligence
http://www.DiffBot.com/

Diggernaut – Cloud Based Extraction Tool
https://www.diggernaut.com/

Digital Footprints - Collect Facebook Data
http://digitalfootprints.dk/
DiscoverText - Import, Sort, Distribute and Analyze Electronic Content from eMail, Document Repositories, and Social Media
http://discovertext.com/

Easy PDF Cloud
https://www.easypdfcloud.com/

Easy Web Extract – Best Tool for Web Scraping
http://webextract.net/

eGrabber - Data Capture Tools
http://www.egrabber.com/

Facepager - Fetching Public Data From Facebook
https://github.com/strohne/Facepager

FeedsAPI - Extract Content from Web Pages Tool
http://www.feedsapi.com/

Ficstar Software - Web Data Extraction
http://www.ficstar.com/

File Information Tool Set (FITS)
https://projects.iq.harvard.edu/fits

FMiner – Web Scraping Software
http://www.fminer.com/

Fresh WebSuction
http://www.freshwebmaster.com/

GetData.io - Get Valuable Data from the Web in 3 Steps
https://getdata.io/

Grabby
https://grabby.io/

Grepsr – Web Scraping Made Simple, Fast and Manageable
https://www.grepsr.com/

Helium Scraper
http://www.heliumscraper.com/
Huginn - Your Agents Are Standing By
https://github.com/cantino/huginn

Hunter - Connect With Anyone
https://hunter.io/

HYPHE - Web Corpus Curation Tool Featuring A Research-Driven Web Crawler
http://hyphe.medialab.sciences-po.fr/

iMacros – Data Extraction
http://imacros.net/overview

Imagination Engines
http://www.Imagination-Engines.com/

Import.io - Turn the Web Into Data With Extractors, Crawlers and Connectors
https://import.io/

InfoExtractor - Extracts Relevant Information from Blogs, YouTube and Twitter
http://www.infoextractor.org/

Information Retrieval (IR) and Information Extraction (IE) on the Web
http://www.webir.org/

Introduction to Information Retrieval

Introduction to Web Scraping Using Python
https://github.com/qut-dmrc/web-scraping-intro-workshop

iOpus Internet Macros
http://www.iopus.com/imacros/

iRobotSoft – Visual Web Scraping and Web Automation
http://irobotsoft.com/

iWeb Scraping Services
http://www.iwebscraping.com/

Jaspersoft® ETL - The Open Source Data Integration Platform
https://community.jaspersoft.com/project/jaspersoft-ett
Junar - Discovering Data  
http://www.junar.com/

Karma - Data Integration Tool  
http://www.isi.edu/integration/karma/

Kimono - Turn Website Into Structured APIs From Your Browser In Seconds  
https://www.kimonolabs.com/

Knowledge Discovery Resources  
http://www.KnowledgeDiscovery.info/

Knowlesys® - Web Data Extraction, Web Grabber and Screen Scraper  
http://www.knowlesys.com/index.htm

Liberty Metrics – Web Scaping Services  
http://libertymetrics.com/

LingPipe – Information Extraction and Data Mining Tools  
http://alias-i.com/lingpipe/

Listly - Fully Automated Web Scraping Service  
http://listly.io/

Metadata Extraction Tool  
http://meta-extractor.sourceforge.net/

Mozenda – Comprehensive Web Data Gathering  
http://www.mozenda.com/

NCapture - Capture Web Content  

Netlytic - Making Sense of Online Conversations  
https://netlytic.org/home/

Newprosoft – Web Data Extraction Software  
http://newprosoft.com/

NewsClipper.com - Snip and Ship Dynamic News Content to Your Web Pages  
http://www.newsclipper.com/
Octoparse – Automated Web Scraping Software
http://www.octoparse.com/

Online Data Extractor Tool
http://www.onlinedataextractor.com/

Open Datasets
http://www.DataPortals.org/
https://github.com/caesar0301/awesome-public-datasets
https://www.kaggle.com/datasets
https://www.data.gov/
https://aws.amazon.com/public-datasets/
https://data.world/
http://data.worldbank.org/
http://www.OpenDataSets.info/

OutWit Hub - Harvest the Web With Your Own Web Collection Engine
http://www.outwit.com/

ParseHub – Web Crawling Using Machine Learning
http://www.ParseHub.com/

Pervasive Data Management and Integration Products
http://www.pervasive.com/

Priceonomics - Crawl Data From the Web
http://priceonomics.com/

Proxycrawl - Stay Anonymous While Crawling the Web
https://proxycrawl.com/

QL2 Software - Unstructured Data Management and Web Mining Software
http://www.ql2.com/

Quick Code
https://quickcode.io/

re3data.org - 2,000 Data Repositories
https://www.re3data.org/
REBOL Technologies
http://www.rebol.com/

ReVerb - Open Information Extraction Software
http://reverb.cs.washington.edu/

SalesTools.io
https://salestools.io/

Semantic Scholar - Free Scientific Literature Search and Discovery
http://allenai.org/semantic-scholar/

ScrapeForge
http://freecode.com/projects/scrapeforge

ScrapeHero
https://www.scrapehero.com/

Scrapper
http://freecode.com/projects/scraper

ScrapingHub – Cloud Based Data Extraction Tool
http://www.ScrapingHub.com/

Scraping Solutions – When the Solution You Seek Seems Impossible

Scrapy – Open Source Web Scraping Framework for Python
http://scrapy.org/

Screen-Scraper
http://freecode.com/projects/screenscraper

Screen-Scraper – Extracts Information From Web Sites
http://www.Screen-Scraper.com/

Screenscraping the Senate by Paul Ford

Search and Replace with TextPipe Pattern Matching
http://www.datamystic.com/textpipe.html
Sensible Code
http://sensiblecode.io/

Social Media Data Collection Tools
http://socialmediadata.wikidot.com/

Software for Web Scraping
http://scraping.pro/software-for-web-scraping/

Spinn3r - Indexing the Blogosphere
http://docs.spinn3r.com/#overview

SPSS Modeler
http://developer.ibm.com/predictiveanalytics

Squirro - Find, Remember, Organize and Share Important Information
https://squirro.com/

STACKS - Social Media Tracker, Analyzer, & Collector Toolkit at Syracuse
https://github.com/bitslabsyr/stack

TadaWeb - Clone and Amplify Human Intelligence for Web Data Collection and Analysis
https://www.tadaweb.com/

Texifter - Search, Sift, Sort, Classify and Analyze
http://texifter.com/

TextConverter 4
https://www.simx.com/

TextRazor - Text Analysis Infrastructure
https://www.textrazor.com/

Topicgrazer - Graze On Web Pages and Documents

UiPath – Web Data Extraction
https://www.uipath.com/guides/web-data-extraction

Unit Miner - Web Data Extraction Software
http://www.unitminer.com/
VietSpider
http://binhgiang.sourceforge.net/

VisualScraper – Web Data Extractor

Visual Web Ripper – Data Extraction Software

Visual Web Task
http://www.lencom.com/VisualWTSite.html

W3C Publishes Data Extraction Language (DEL) as W3C Note

Web Content Extractor
http://www.newprosoft.com/

Web Data Extraction
http://www.wintask.com/web-data-extraction.php

Web Data Extraction Software Data Toolbar
https://webdataextractionsoftwaredatatoolbar.en.softonic.com/

Web Data Extractor
http://www.rafasoft.com/

Web Data Extractor
http://www.webextractor.com/

Web Data Extractor
http://fivesmallq.github.io/web-data-extractor

Web Data Extractor

Web Data Guru – Web Data Extraction and Scraping Services
http://www.webdataguru.com/

Web-Harvest – Open Source Web Data Extraction Tool
http://web-harvest.sourceforge.net/index.php
WebHarvy – Intuitive Powerful Visual Web Scraper  
https://www.webharvy.com/index.html

Webhose.io – Web Data For Your Business  
http://www.webhose.io/

Web Robots – Web Scraping and Crawling  
https://webrobots.io/

Web Scraper  
http://www.webscraper.io/

https://en.wikipedia.org/wiki/Web_scraping

Website Data Extractor – Time to Rethink Web Scraping  
http://www.kofax.com/

Website Downloader  
https://websitedownloader.io/

Website Extractor – Offline Browser  

WebSunDew – Advanced Web Scraping Tool  
http://www.websundew.com/

Wikimedia Public Data Dumps  
http://meta.wikimedia.org/wiki/Data_dumps

WinAutomation  
http://www.winautomation.com/

XRay Web Scraping Tool  
http://freecode.com/projects/xrayguibasedwebscrapingtool

YaCy Web page Indexer  
http://freecode.com/projects/yacy
Subject Tracer™ Information Blogs

Subject Tracer™ Information Blogs created and developed by the Virtual Private Library™ combine the best of the latest tools on the Internet. Using bots, blogs and news aggregators the Subject Tracer™ Information blogs generate RSS feeds with the latest resources to create a current information resource flow through niched subject tracers. I am proud to be the creator of the Internet’s first Subject Tracer™ Information Blogs:

Virtual Private Library™
http://www.VirtualPrivateLibrary.com/

Accessibility Resources
http://www.AccessibilityResources.info/

Agriculture Resources
http://www.AgricultureResources.info/

AnswerSpot
http://www.AnswerSpot.us/

Artificial Intelligence Resources
http://www.AIResources.info/

Astronomy Resources
http://www.AstronomyResources.info/

Auction Resources
http://www.AuctionResources.info/

Biological Informatics
http://www.BiologicalInformatics.info/

Biotechnology Resources
http://www.BiotechnologyResources.info/

Bot Research
http://www.BotResearch.info/

Business Intelligence Resources
http://www.BIResources.info/
ChatterBots
http://www.ChatterBots.info/

Data Mining Resources
http://www.DataMiningResources.info/

Deep Web Research
http://www.DeepWebResearch.info/

Directory Resources
http://www.DirectoryResources.info/

eCommerce Resources
http://eCommerceResources.info/

Education and Academic Resources
http://www.EducationResources.info/

Elder Resources
http://www.ElderResources.info/

Employment Resources
http://www.EmploymentResources.info/

Entrepreneurial Resources
http://www.EntrepreneurialResources.info/

Fact Checkers Directory
http://www.FactCheckers.info/

Financial Sources
http://www.FinancialSources.info/

Finding People
http://www.FindingPeople.info/

Games Resources
http://www.GamesResources.info/

Genealogy Resources
http://www.GenealogyResources.info/
Grant Resources
http://www.GrantResources.info/

Green Files
http://www.GreenFiles.info/

Grid, Distributed and Cloud Computing Resources
http://www.GridResources.info/

Healthcare Resources
http://www.HealthcareResources.info/

Information Futures Markets
http://www.InformationFuturesMarkets.com/

Information Quality Resources
http://www.InformationQualityResources.info/

International Trade Resources
http://www.InternationalTradeResources.info/

Internet Alerts
http://www.InternetAlerts.info/

Internet Demographics
http://www.InternetDemographics.info/

Internet Experts
http://www.InternetExperts.info/

Internet Hoaxes
http://www.InternetHoaxes.info/

Intrapreneurial Resources
http://www.IntrapreneurialResources.info/

Journalism Resources
http://www.JournalismResources.info/

Knowledge Discovery
http://www.KnowledgeDiscovery.info/
Military Resources
http://www.MilitaryResources.info/

New Economy Analytics, Resources and Alerts
http://www.NewEconomyAnalytics.com/

Outsourcing/Offshoring Information and Resources
http://www.OutsourcingOffshore.us/

Privacy Resources
http://www.PrivacyResources.info/

Reference Resources
http://www.ReferenceResources.info/

Research Resources
http://www.ResearchResources.info/

RestStress™
http://www.RestStress.com/

Script Resources
http://www.ScriptResources.info/

ShoppingBots
http://www.ShoppingBots.info/

Social Informatics
http://www.SocialInformatics.info/

Statistics Resources and Big Data
http://www.StatisticsResources.info/

Student Research
http://www.StudentResearch.info/

Theology Resources
http://www.TheologyResources.info/

Tutorial Resources
http://www.TutorialResources.info/
World Wide Web Reference  
http://www.WWWReference.info/

Author Information: Marcus P. Zillman, M.S., A.M.H.A. Executive Director of the Virtual Private Library is an international Internet expert, author, keynote speaker and corporate consultant in the area of information retrieval, knowledge discovery, knowledge harvesting, artificial intelligence and bots/intelligent agents. He has created numerous world wide web sites including 54 Subject Tracer™ Information Portals and Blogs; written a number of internet miniguides, white papers, manuals and books; hosted over 160 weekly Internet television shows, writes a weekly and monthly column on Current Awareness on the Internet; writes a monthly newsletter Awareness Watch and delivers keynote presentations throughout the international marketplace. He also actively delivers one and two day workshops for key industry sectors displaying how the Internet can be used as a tool to maintain current awareness and professional competencies.

Additional websites by Marcus P. Zillman, M.S., A.M.H.A.:

Marcus P. Zillman's Blog (20,000+ Postings)  
http://www.zillman.us/

Marcus P. Zillman Abbreviated Bio  
http://www.zillman.info/
White Papers by Marcus P. Zillman
http://www.WhitePapers.us/

Internet MiniGuides™
http://www.InternetMiniGuide.com/

Awareness Watch™ Newsletter
http://www.AwarenessWatch.com/

Marcus P. Zillman's Columns
http://www.ZillmanColumns.com

LinkSeries Publications
http://www.LinkSeries.com/

Links By Marcus™
http://www.LinksByMarcus.com/

Workshops By Marcus™
http://www.WorkshopsByMarcus.com/

SourceSeries Internet Research Workshops
http://www.SourceSeries.com/

Watch Marcus™
http://www.WatchMarcus.com/

listen to marcus™
http://www.ListenToMarcus.com

Research White Papers, Articles, Lectures and Speeches by Marcus P. Zillman, M.S., A.M.H.A.:

2018 Directory of Directories

Academic and Scholar Search Engines and Sources
http://www.ScholarSearchEngines.com/

Bots, Blogs and News Aggregators 2018
http://www.BotsBlogs.com/
Business Intelligence Online Resources 2018
http://www.BIOnlineResources.info/

Cloud Computing Resources Primer 2018
http://www.zillman.us/white-papers/grid-distributed-and-cloud-computing-resources-primer/

Current Awareness Discovery Tools on the Internet 2018

Deep Web Research and Discovery Resources 2018 Online White Paper
http://DeepWeb.us/

eMarketing MiniGuide 2018
http://www.eMarketingMiniGuide.com/

eReference Library Link Toolkit 2018
http://www.eReferenceLibrary.com/

Finding Experts By Using the Internet 2018
http://www.FindingExperts.info/

Finding People Resources and Sites 2018
http://www.FindingPeople.info/

Healthcare Bots and Subject Directories 2018
http://www.HealthcareBots.info/

Knowledge Discovery Resources 2018
http://www.KDResources.info/

New Economy Resources 2018
http://www.NewEconomyResources.com/

Online Research Browsers 2018
http://www.zillman.us/white-papers/online-research-browsers/

Online Research Tools
http://www.OnlineResearchTools.info/

Online Social Networking 2018
http://www.OnlineSocialNetworking.info/
Searching the Internet 201 – The Primer
http://www.SearchingTheInternet.info/

Using the Internet As a Dynamic Resource Tool for Knowledge Discovery 2018

Web Data Extractors 2018
http://www.WebDataExtractors.com/

Web Guide for the New Economy 2018

White Papers By Marcus P. Zillman, M.S., A.M.H.A.
http://www.WhitePapers.us/

Internet Tutor by Marcus P. Zillman, M.S., A.M.H.A.
http://www.InternetTutor.info/
Visit this site to learn about the availability of Marcus P. Zillman to tutor you or your associate one on one in the privacy of your residence or office on the latest happenings of the Internet including Internet basics to advanced Internet searching using bots and creating your own personal blog.

Internet Speaking by Marcus P. Zillman, M.S., A.M.H.A.
http://www.InternetSpeaker.net
Visit this site to learn about Marcus P. Zillman’s speaking engagements for your organization meetings and events. View and listen to his previous presentations as well as his weekly television shows.

Internet Consulting by Marcus P. Zillman, M.S., A.M.H.A.
http://InternetConsultant.BlogSpot.com/
Visit this site to obtain information about obtaining the consultation services of Marcus P. Zillman for your company including eCommerce audits, utilization of bots, blogs and news aggregators or the creation of your own personal virtual private library powered by Subject Tracer™ Information bots!

Current Awareness Monitors, Alerts and Information Traps
http://www.ecurrentAwareness.com/
Marcus P. Zillman’s latest report Current Awareness Monitors, Alerts and Information Traps for is now for purchase online and for immediate download. This report is a comprehensive listing of the latest resources, sources and sites for current awareness on the Internet. This is a must read for anyone who must stay current in their profession
and/or business activity as the list of URLs will keep you at the leading edge of your career.

**Market Intelligence Resources**
http://www.MarketIntelligenceResources.com/
Marcus P. Zillman’s just released professional Internet MiniGuide is titled Market Intelligence Resources and is available for purchase online and immediate download. This 193 page digital miniguide represents a comprehensive listing of the latest resources, sources and sites to discover the latest Market Intelligence sources available on the Internet with many of them freely available! Designed specifically for today’s entrepreneur, professional and/or investor.

**Entrepreneurial Links 101**
http://www.EntrepreneurialLinks.com/
Marcus P. Zillman’s newly released 231 page eReference digital book for the up and coming entrepreneur. Entrepreneurial Links 101 gives an alphabetical listing of the very best Internet and World Wide Web sites covering Entrepreneur Resources, Business Intelligence Resources and an extremely comprehensive list of Online Research Tools. This is considered by many to be the entrepreneur’s bible for finding relevant and competent online resources!

**Internet Privacy and Security Resources**
http://www.InternetPrivacySecurity.net/
Marcus P. Zillman’s latest eReference digital publication is a selected comprehensive alphabetical listing of the latest resources and sites covering all aspects of privacy and security currently available over the Internet. From the board room to the family room, these resources and sites give you the information you need to maintain your privacy and security as you use the Internet in your business and personal life.

**Research Resources Online Guide**
http://www.ResearchResourcesOnline.net/
Marcus P. Zillman’s latest LinkSeries Publication is a 340 page digital guide of a selected comprehensive alphabetical listing of the latest and greatest resources and sites covering all areas of research that is currently available over the Internet. The guide covers online research resources and tools for the Newbie to research as well as the Seasoned researcher. Contents include: a) Research Resources, b) Research Tools, c) Student Research Resources Toolkit, d) Knowledge Discovery/Management and Data Mining Resources, e) Knowledge Discovery/Retrieval and the World Wide Web Resources, f) Business Intelligence Resources, g) Reference Resources, and h) Subject Tracer™ Information Blogs.
The Survivor’s Manual for The New Economy.
Marcus P. Zillman’s latest LinkSeries Publication is a 239 page digital read that gives excellent resources and annotated sources for the new economy analytics, alerts, ecommerce, financial sources, invisible and deep web resources, social and business networking sources along with new economy competitive and business intelligence resources and an extremely comprehensive listing of new economy online tools.