# Resources for Extracting Information from the World Wide Web

By

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

The June 2015 Zillman Column features **Resources for Extracting Information from the World Wide Web** by Marcus P. Zillman, M.S., A.M.H.A. and is a comprehensive listing of data extraction resources currently available on the Internet. 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 over 20 billion pages with over 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. There are a number of eMail extraction tools still available through the Internet and I have decided not to list these as they aid to the on-going and increasing problem of SPAM except for a readily available DMOZ Directory listing. The below compilation of sources is taken from my white paper titled **Web Data Extractors** and is constantly updated with Subject Tracer™ bots at the following URL:

http://www.WebDataExtractors.com/

These resources and sources will guide you through the many pathways available on the Internet to find the latest information extraction sources and sites.

## Resources for Extracting Information from the World Wide Web:

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

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

artoo.js - The Client-Side Scraping Companion
http://medialab.github.io/artoo/

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

Automation Anywhere – Web Data Extraction Software
http://www.automationanywhere.com/solutions/webDataExt.htm

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/

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

Bot Research
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






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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

Client-Side Deep Web Data Extraction
http://www.tic.udc.es/~mad/publications/ceceast2004.pdf

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

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

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

Crowbar – Open Source Web Scraping Environment
http://simile.mit.edu/wiki/Crowbar

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

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

Data Extractors and Pass-Through Systems – A Selected List
http://datalib.chass.utoronto.ca/misc/dli/extracts.htm

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

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

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

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






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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

DiffBot – Get Data From Web Pages Automatically
http://www.DiffBot.com/

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/

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

ExtractData Technologies - SearchExtract Software
http://www.extradata.com/

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

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

FerretSoft
http://www.webferret.com/

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

File Information Tool Set (FITS)
http://code.google.com/p/fits/

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

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






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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
http://www-nlp.stanford.edu/IR-book/

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

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

iWebMiner – Web Data Mining and Content Scraping
http://www.iwebminer.com/

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

jSEO - Web Crawler For Search Engine Optimization
http://codecanyon.net/item/jseo-web-crawler-for-search-engine-optimization/8770392

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/






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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

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

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

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

NCapture - Capture Web Content
http://www.qsrinternational.com/products_nvivo_add-ons.aspx

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/

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

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

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

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

Realtime Products - The Social Media Data You Need, The Moment You Need It
http://gnip.com/products/realtime/






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

REBOL Technologies
http://www.rebol.com/

ScissorsFly - Your Web Clipper and Scrapbook
https://www.scissorsfly.com/

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

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

ScraperWiki - Community of Programmers Sifting Information To Give You the Edge
https://scraperwiki.com/

ScrapeShield - Monitor and Track Misuse of Your Content
https://www.cloudflare.com/apps/scrapeshield

Scrapple - A Framework For Creating Web Scrapers and Web Crawlers
http://scrappleapp.github.io/scrapple/

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
http://www.xml.com/pub/a/2004/09/01/hack-congress.html

Screen Snarfs - Darkspell: Perl Code for Screen Scrapers
http://www.darkspell.com/gadgets/snarfs/

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






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

Spinn3r - Indexing the Blogosphere
http://www.spinn3r.com/

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

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

Text Mining and Web-Based Information Retrieval Reference
http://filebox.vt.edu/users/wfan/text_mining.html

Topicgrazer - Graze On Web Pages and Documents
http://www.topicscape.com/Topicgrazer/help.php

Topsy - Twitter Search, Monitoring and Analytics
http://topsy.com/

Unit Miner - Web Data Extraction Software
http://www.unitminer.com/

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

W3C Publishes Data Extraction Language (DEL) as W3C Note
http://xml.coverpages.org/ni2001-11-06-a.html

Web Data Extraction Software
http://www.automationanywhere.com/solutions/webDataExt.htm

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

Web-Harvest – Open Source Web Data Extraction Tool
http://web-harvest.sourceforge.net/index.php






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

WebQL
http://www.ig.com.au

Website Extractor – Offline Browser
http://www.internet-soft.com/extractor.htm

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

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

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

YaCy Web page Indexer
http://freecode.com/projects/yacy






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

## 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/

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

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 Checker Directory
http://www.FactCheckers.us/

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

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

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

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






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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/





zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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/






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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

Tutorial Resources
http://www.TutorialResources.info/

World Wide Web Reference
http://www.WWWReference.info/

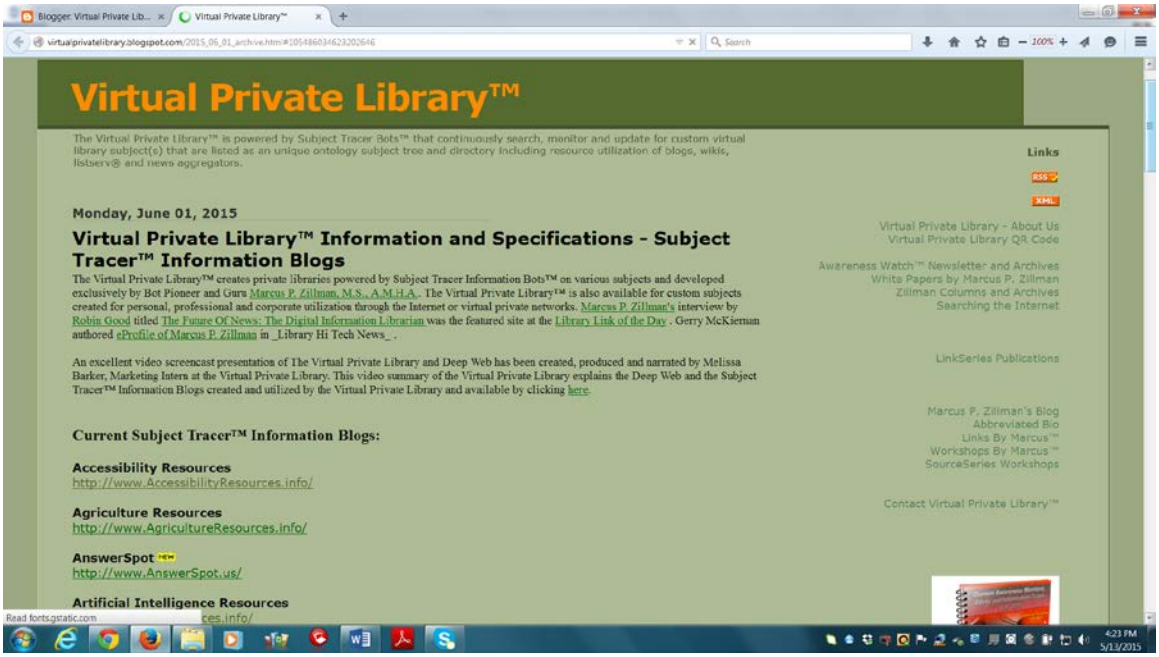

**Figure 1: Virtual Private Library™**

**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




zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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
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/






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

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

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

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

Bots, Blogs and News Aggregators
http://www.BotsBlogs.com/

Business Intelligence Online Resources
http://www.BIOnlineResources.info/

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

Current Awareness Discovery Tools on the Internet
http://www.zillman.us/white-papers/current-awareness-discovery-tools-on-the-internet/

Deep Web Research and Discovery Resources 2015 Article - LLRX and Online White Paper
http://zillman.blogspot.com/2015/01/llrx-deep-web-research-and-discovery.html
http://DeepWeb.us/

eMarketing miniGuide 2015
http://www.eMarketingMiniGuide.com/

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

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

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

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

Knowledge Discovery Resources 2015
http://www.KDResources.info/

New Economy Resources 2015
http://www.NewEconomyResources.com/

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

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

Online Social Networking
http://www.OnlineSocialNetworking.info/

Searching the Internet
http://www.SearchingTheInternet.info/

Using the Internet As a Dynamic Resource Tool for Knowledge Discovery
http://www.zillman.us/white-papers/using-the-internet-as-a-dynamic-resource-tool-for-knowledge-discovery/

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

Web Guide for the New Economy
http://www.WebGuideNewEconomy.com/

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






zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462

**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 available 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, and f) Subject Tracer™ Information Blogs.

**The Survivor's Manual for The New Economy**.
http://www.NewEconomyManual.com/
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.




zillman@VirtualPrivateLibrary.com
eVoice: (800) 858-1462