Web Data Extractors

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 over eight billion pages with over 800 billion 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 discovery 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 Google™ Directory listing:

Web Data Extractors:

Advanced Information Extractor (AIE)
http://poorva.com/aie/

Anomic HTTP Proxy
http://freshmeat.net/releases/173068/

Anthracite
http://freshmeat.net/projects/anthracite/

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://freshmeat.net/projects/beautifulsoup/

BotMakers
http://www.BotMakers.com/

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

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

Captiva Software: Digital Information Capture Software
http://www.captivasoftware.com/index.asp

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

Compaq's Web Language
http://research.compaq.com/SRC/WebL/

Dapper - Extract and Use Website Information for Mashups
http://www.dappit.com/index.php

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

Data Extractors – Special Report
http://snipurl.com/8flw

Data Extraction Web Scraping and Web Mining Using Screen Scraping Technology
http://www.connotate.com/

Data Locators on the Web
http://www.nd.edu/~lsrteach/archive_old%20site/01207all/presentation_final.htm

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

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

Effective Web Data Extraction with Standard XML Technologies
http://www10.org/cdrom/papers/102/

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

Extracting Knowledge
http://www.intelligentkm.com/feature/010507/feat1.jhtml?_requestid=185742

FerretSoft
http://www.FerretSoft.com/

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

Google™ Directory – Extractors
http://directory.google.com/Top/Computers/Software/Shareware/Windows/Internet/Email/Extractors/

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

Information Foraging and Extraction Techniques for Internet-Based Literature and Data
http://www.asceditor.usm.edu/ASC%202006%20CD/2006pro/2006/CEGE03_Hannon06_7900.htm

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

Intelliseek
http://www.Intelliseek.com/

iOpus® Internet Macros™
http://www.iopus.com/iim.htm

Kapow RoboSuite Platform - Solutions for Data Collection
http://www.kapowtech.com/solutions_datacollection.htm

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

zillman@VirtualPrivateLibrary.com

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

Lingway
http://www.lingway.fr/en/

Mastering Data Extraction
http://www.dbmsmag.com/9606d05.html

mnot: xpath2rss - HTML->RSS scraper
http://www.mnot.net/xpath2rss/

NewsClipper.com - Snip and Ship Dynamic News Content to Your Web Pages
http://www.newsclipper.com/

NQL Technologies
http://www.nqltech.com/

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

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

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

ScrapeForge
http://freshmeat.net/projects/scrapeforge/

ScrapeGoat
http://www.ScrapeGoat.com/

Scraper
http://freshmeat.net/projects/scraper/

Screen-Scraper
http://freshmeat.net/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

zillman@VirtualPrivateLibrary.com

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

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

Sitescooper: Scoop Websites Onto Your PalmPilot
http://sitescooper.org/

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

That Robot Site – Smart Internet Solutions
http://www.thatrobotsite.com/

Unit Miner - Web Data Extraction Software
http://www.qualityunit.com/unitminer/web-extraction-tool.html

URL Link Extractor
http://www.fuddyduddy.connectfree.co.uk/urlgen.htm

Visual Web Spider
http://www.newprosoft.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 Extractor
http://www.webextractor.com/

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

Web Data Mining
http://www.blossom.com/web_mining.html

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

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

WebScraper Plus +
http://www.velocityscape.com/

Website Extractor
http://www.hot-shareware.com/internet-tools/website-extractor/

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

Web Spider, Link Extraction, And Other Extractor Products
http://www.pjltechnology.com/

WebWrappers - Agents for E-Cataloging, Automation, Data Aggregation and Integration
http://www.webwrappers.com/

wisosoft - Intelligent Agents for Web Aggregation and Deep Web Processing
http://www.oejhw.or.at/rand.htm

WWW::Extractor
http://freshmeat.net/projects/ade/

XRay Web Scraping Tool
http://freshmeat.net/projects/xrayguibasedwebscrapingtool/

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/

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/

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

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

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

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

zillman@VirtualPrivateLibrary.com

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

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

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

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

Grid Resources
http://www.GridResources.info/

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

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

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

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

Military Resources
http://www.MilitaryResources.info/

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

zillman@VirtualPrivateLibrary.com

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.WcriptResources.info/

ShoppingBots
http://www.ShoppingBots.info/

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

Statistics Resources
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/

zillman@VirtualPrivateLibrary.com

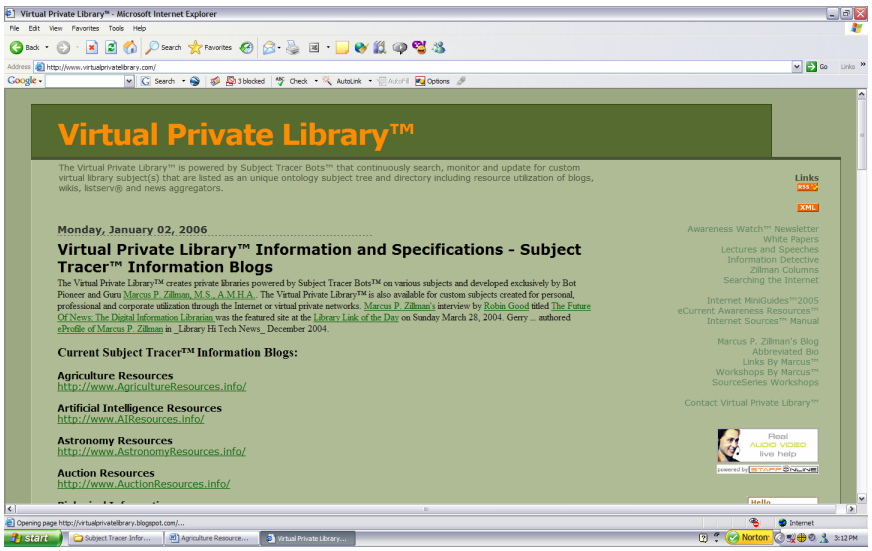

Figure 2 Virtual Private Library™

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 46 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
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™ 2005
http://www.InternetMiniguide.com/

zillman@VirtualPrivateLibrary.com

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

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

eCurrent Awareness Resources™ 2005 Business Intelligence Report
http://www.eCurrentAwareness.com/

Internet Sources™ Manual
http://www.InternetSources.info/

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.:

Academic and Scholar Search Engines and Sources
http://zillman.blogspot.com/2004/12/academic-and-scholar-search-engines.html

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

Business Intelligence Online Resources
http://zillman.blogspot.com/2005/04/business-intelligence-online-resources.html

Current Awareness Discovery Tools on the Internet
http://zillman.blogspot.com/2004/09/current-awareness-discovery-tools-on.html

Deep Web Research 2006 Article - LLRX
http://zillman.blogspot.com/2006/01/llrx-january-2006-issue-deep-web.html

Healthcare Bots and Subject Directories
http://zillman.blogspot.com/2005/05/healthcare-bots-and-subject.html

Information Detective – Online Streaming Tutorial Videos
http://www.InformationDetective.com/

Knowledge Discovery Resources 2006
http://zillman.blogspot.com/2005/03/knowledge-discovery-resources-2006.html

Lectures and Speeches by Marcus P. Zillman, M.S., A.M.H.A.
http://snipurl.com/57jp

Online Research Browsers
http://zillman.blogspot.com/2004/10/online-research-browsers-internet.html

Online Research Tools
http://zillman.blogspot.com/2004/09/online-research-tools.html

Online Social Networking
http://zillman.blogspot.com/2004/09/online-social-networking-internet.html

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

Using the Internet As a Dynamic Resource Tool for Knowledge Discovery
http://zillman.blogspot.com/2004/09/using-internet-as-dynamic-resource.html

Web Data Extractors
http://zillman.blogspot.com/2004/09/web-data-extractors.html

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!

Marcus P. Zillman's latest 378 page manual **Internet Sources™** is now available for purchase online and for immediate download. This book makes a great reference resource for the "newbie" to the Internet as well as the seasoned veteran "Internaut". Visit the following site for additional information and online ordering fulfillment:

Internet Sources™ Manual
http://www.InternetSources.info

Marcus P. Zillman's latest report eCurrent Awareness Resources 2005 is now 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. Visit the following site for additional information and online ordering fulfillment:

eCurrent Awareness Resources 2005
http://www.ecurrentAwareness.com/

zillman@VirtualPrivateLibrary.com