The Use of Web-scraping Software in Searching for Grey Literature

Article *in* Grey Journal · October 2015

CITATIONS
5

READS
1,089

1 author:

Neal Haddaway
Stockholm Environment Institute
53 PUBLICATIONS **476** CITATIONS

SEE PROFILE

The Use of Web-scraping Software in Searching for Grey Literature

Neal R. Haddaway *(Sweden)*

Abstract

Searches for grey literature can require substantial resources to undertake but their inclusion is vital for research activities such as systematic reviews. Web scraping, the extraction of patterned data from web pages on the internet, has been developed in the private sector for business purposes, but it offers substantial benefits to those searching for grey literature. By building and sharing protocols that extract search results and other data from web pages, those looking for grey literature can drastically increase their transparency and resource efficiency. Various options exist in terms of web-scraping software and they are introduced herein.

The Challenge of Searching for Grey Literature

The editorial scrutiny and peer-review that form integral parts of commercial academic publishing are useful in assuring reliability and standardised reporting in published research. However, publication bias can cause an overestimation of effect sizes in syntheses of the (commercially) published literature (Gurevitch and Hedges 1999; Lortie et al. 2007). In a recent study by Kicinski et al. (2015), the largest analysis of publication bias in meta-analyses to-date, publication bias was detected across the Cochrane Library of systematic reviews, although there was evidence that more recent research suffered to a lesser degree, thanks to mitigation measures applied in medical research in recent decades.

Some applied subject areas, such as conservation biology, are particularly likely to be reported in sources other than academic journals, so called *practitioner-held data* (Haddaway and Bayliss in press); for example reports of the activities of non-governmental organisations. Such grey literature is vital for a range of research, policy and practical applications, particularly informing policy decision-making. Documents produced by governments, business, non-governmental organisation and academics can provide a range of useful information, but are often overlooked in traditional meta-analyses and literature reviews.

Systematic reviews were established in the medical sciences to collate and synthesise research on particular clinical interventions in a reliable, transparent and objective manner, and were a response to the susceptibility to bias common to traditional literature reviews (Allen and Richmond 2011). In the last decade systematic review methodology has been translated into a range of other subjects, including social science (Walker et al. 2013) and environmental management (CEE 2013). A key aspect of systematic review methodology is that searches are undertaken for grey literature to mitigate possible publication bias and to include practitioner-held data. These searches may fail to find any research that is ultimately included (e.g. Haddaway et al. 2014), but it is important for the reliability and transparency of the review to demonstrate that this is the case: other reviews have demonstrated significant proportions of grey literature in the synthesised evidence base (Bernes et al. 2015).

Systematic review searches for grey literature can be particularly challenging and time-consuming. No comprehensive database resources exist in the environmental sciences for grey literature, as in many other disciplines, and so searches must include web-based search engines, specialist databases such as repositories for theses, organisational web sites such as non-governmental organisations, governmental databases and university repositories. Typically between 30 (Pullin et al. 2013) and 70 (Haddaway et al. 2014) individual web sites are searched. Systematic reviews often complement these manual searches using web-based search engines; both general (e.g. Google) and academic (e.g. Google Scholar). Not only are searches of this number of resources time-consuming, but they are also typically undertaken in a very non-transparent manner: excluded articles are rarely recorded and searches are not readily updatable or repeatable. Furthermore, included resources must be listed individually by hand in any documentation of search activities, whilst search results from academic databases, such as Web of Science, can be downloaded as full citations.

Web Scraping Software: a potential solution

Data scraping is a term used to describe the extraction of data from an electronic file using a computer program. *Web scraping* describes the use of a program to extract data from HTML files on the internet. Typically this data is in the form of patterned data, particularly lists or tables. Programs that interact with web pages and extract data use sets of commands known as *application programming interfaces* (APIs). These APIs can be 'taught' to extract patterned data from single web pages or from all similar pages across an entire web site. Alternatively, automated interactions with websites can be built into APIs, such that links within a page can be 'clicked' and data extracted from subsequent pages. This is particularly useful for extracting data from multiple pages of search results. Furthermore, this interactivity allows users to automate the use of websites' search facilities, extracting data from multiple pages of search results and only requiring users to input search terms rather than having to navigate to and search each web site first.

One major current use for web scraping is for businesses to track pricing activities of their competitors: pricing can be established across an entire site in relatively short time scales and with minimal manual effort. Various other commercial drivers have caused a large number and variety of web scraping programs to have been developed in recent years (see Table 1). Some of these programs are free, whilst others are purely commercial and charge a one off or regular subscription fee.

These web scraping tools are equally as useful in the research realm. Specifically, they can provide valuable opportunities in the search for grey literature, by: i) making searches of multiple websites more resource-efficient; ii) drastically increasing transparency in search activities; and iii) allowing researches to share trained APIs for specific websites, further increasing resource-efficiency.

A further benefit of web scraping APIs relates to their use with traditional academic databases, such as Web of Science. Whilst citations, including abstracts, are readily extractable from most academic databases, many databases hold more useful information that is not readily exportable, for example corresponding author information. Web scraping tools can be used to extract this information from search results, allowing researchers to assemble contact lists that may prove particularly useful in requests for additional data, calls for submission of evidence, or invitations to take part in surveys, for example.

Table 1. List of Major web scraping tools
(adapted from http://scraping.pro/software-for-web-scraping/). Prices were accurate at the time of publication.

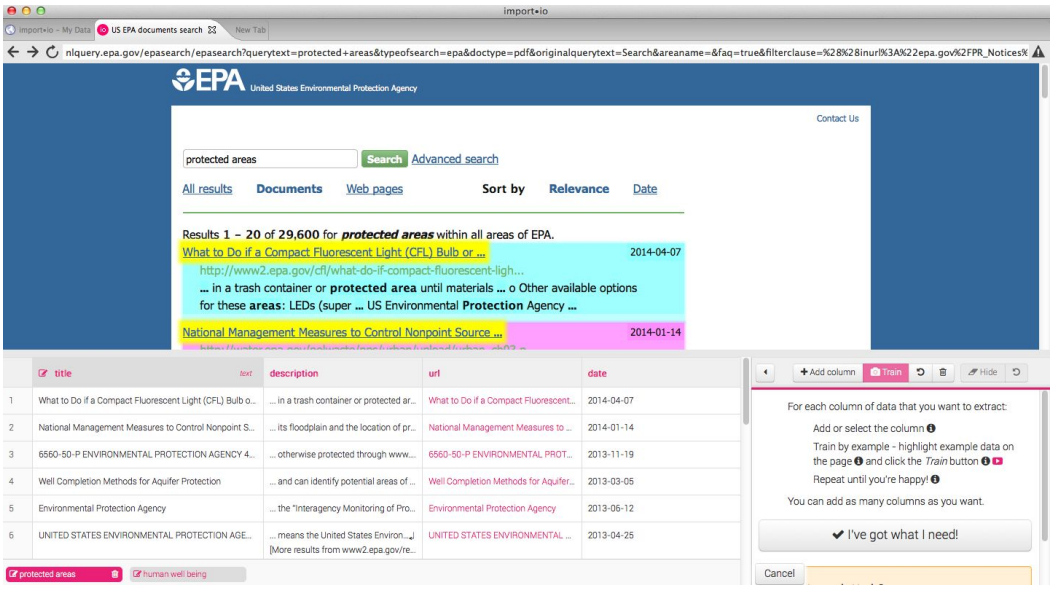
Platform	Description[a]	Cost[b]	URL
Import.io	"Instantly Turn Web Pages into Data. No Plugin, No Training, No Setup. Create custom APIs or crawl entire websites using our desktop app - no coding required!"	Free (charge for premium service)	www.import.io
DataToolBar	" The Data Toolbar is an intuitive web scraping tool that automates web data extraction process for your browser. Simply point to the data fields you want to collect and the tool does the rest for you."	Free to try (no export facility), $24 for full version	www.datatoolbar.com
Visual Web Ripper	"Visual Web Ripper is a powerful visual tool used for automated web scraping, web harvesting and content extraction from the web. Our data extraction software can automatically walk through whole web sites and collect complete content structures such as product catalogs or search results."	$299 (including 1 year of maintenance)	www.visualwebripper.com
Helium Scraper	"Extract data from any website. Choose what to extract with a few clicks. Create your own actions. Export extracted data to a variery [sic] of file formats. "	Basic version $99 to Enterprise version $699	www.heliumscraper.com
OutWit Hub	"OutWit Hub breaks down Web pages into their different constituents. Navigating from page to page automatically, it extracts information elements and organizes them into usable collections."	Lite version free, Pro version at $89.90	www.outwit.com
Screen Scraper	"Screen Scraper automates copying text from a web page, clicking links, entering data into forms and submitting them, iterating through search results pages, downloading files (PDF, MS Word, images, etc.)."	Basic edition free, commercial versions from $549 to $2,799	www.screen-scraper.com
Web Content Extractor	"Web Context Extractor is a professional web data extraction software designed not only to perform the most of [sic] dull operations automatically, but also to greatly	$89	www.newprosoft.com/web-content-extractor.htm

	increase productivity and effectiveness of the web data scraping process. Web Content Extractor is highly accurate and efficient for extracting data from websites."		
Kimono	"Kimono lets you turn websites into APIs in seconds. You don't need to write any code or install any software to extract data with Kimono. The easiest way to use Kimono is to add our bookmarklet to your browser's bookmark bar. Then go to the website you want to get data from and click the bookmarklet. Select the data you want and Kimono does the rest."	Free with additional features costing up to $180	www.kimonolabs.com
FMiner	"FMiner is a software for web scraping, web data extraction, screen scraping, web harvesting, web crawling and web macro support for windows and Mac OS X. It is an easy to use web data extraction tool that combines best-in-class features with an intuitive visual project design tool, to make your next data mining project a breeze."	Free 15 day trial, $168-£248	www.fminer.com
Data Extractor by Mozenda	"The Mozenda data scraper tool is very basic; all you have to do is use the program to scrape up information you need off of [sic] websites without all the tiring work of searching websites one by one. Whether you are working for the government such as a police officer or a detective, in the medical field, or even a large business or entrepreneur, website scraping is fast, easy and affordable, plus it saves you or your employees a ton of stressful work and time; use Mozenda's data scraper and let the program do all the hard work for you."	Free trial (500 page credits), $99 to $199 per month	www.mozenda.com/data-extractor
WebHarvy Data Extractor Tool	"WebHarvy is a visual web scraper. There is absolutely no need to write any scripts or code to scrape data. You will be using WebHarvy's in-built browser to navigate web pages. You can select the data to be scraped with mouse clicks."	$99 - $399	www.webharvy.com
Web Data Extractor	"Web Data Extractor [is] a powerful and easy-to-use application which helps you automatically extract specific information from web pages which is necessary in your day-to-day internet / email marketing or SEO activities. Extract targeted company contact data (email, phone, fax) from web for responsible b2b communication. Extract url, meta tag (title, desc, keyword) for website promotion, search directory creation, web research."	$89 - $199	www.webextractor.com
Easy Web Extractor	"An easy-to-use tool for web scrape solutions (web data extracting, screen scraping) to scrape desired web content (text, url, image, html) from web pages just by few screen clicks. No programing required."	$69.99 (with in-app upgrades)	www.webextract.net
WebSundew	"WebSundew is a powerful web scraping tool that extracts data from the web pages with high productivity and speed. WebSundew enables users to automate the whole process of extracting and storing information from the web sites. You can capture large quantities of bad-structured data in minutes at any time in any place and save results in any format. Our customers use WebSundew to collect and analyze the wide range of data that exists on the Internet related to their industry."	$69 - $2,495	www.websundew.com
Handy Web Extractor	" Handy Web Extractor is a simple tool for everyday web content monitoring. It will periodically download the web page, extract the necessary content and display it in the window on your desktop. One may consider it as the data extraction software, taking its own nitch [sic] in the scraping software and plugins."	Free	www.scraping.pro/handy-web-extractor

[a] Descriptions are taken from product web sites (31/01/2015)
[b] Costs correct at time of publication (2015)

Figure 1 shows a screen shot of one web scraping program being used to establish an API for an automated search for grey literature from the website of the US Environmental Protection Agency. This particular web-scraping platform is in the form of a downloadable, desktop-based program; a web browser. The browser is then used to visit and train APIs by identifying *rows* and *columns* in the patterned data: in practice rows will typically be search records, whilst columns will be different aspects of the patterned data, such as titles, authors, sources, publication dates, descriptions, etc. Detailed methods for the use of web scrapers are available elsewhere (Haddaway et al. in press). In this way, citation-like information can be extracted for search results according to the level of detail provided by the website. In addition to extracting search results, as described above, static lists and individual, similarly patterned pages can also be extracted. Furthermore, active links can be maintained, allowing the user to examine linked information directly from the extracted database.

Figure 1. Screenshot of web scraping software being used to train an API for searching for grey literature on the Environmental Protection Agency website. Program used is Import.io.



Just as search results from organisational websites can be extracted as citations, as described above, search results from web-based search engines can be extracted and downloaded into databases of quasi-citations. Microsoft Academic Search (http://academic.research.microsoft.com) results can be extracted in this way, and in fact a pre-trained API is available from Microsoft for extracting data from search results automatically (http://academic.research.microsoft.com/about/Microsoft%20Academic%20Search%20API%20User%20Manual.pdf).

Perhaps a more comprehensive alternative to Microsoft Academic Search is Google Scholar (http://scholar.google.com). Google Scholar, however, does not support the use of *bots* (automated attempts to access the Google Scholar server), and repeated querying of the server by a single IP address (approximately 180 queries or citation extractions in succession) can results in a an IP address being blocked for an extended period (approximately 48-72 hours) (personal observation)[1]. Whilst it is understandable that automated traffic could be a substantial problem for Google Scholar, automation of activities that would otherwise be laboriously undertaken by hand is arguably of great value to researchers with limited resources. Thus, a potential work-around involves the scraping of locally saved search results HTML pages after they had been downloaded individually or in bulk (this may still constitute an infringement of the Google Scholar conditions of use, however). A further cautionary note relates to demands on the servers that host the web sites being scraped. Scraping a significant volume of pages from one site or scraping multiple pages in a short period of time can put significant strain on smaller servers. However, the level of scraping necessary to extract 100s to 1,000s of search results is unlikely to have detrimental impacts on server functionality.

[1] Details of Google Scholar's acceptable use policy are available from the following web page: https://scholar.google.co.uk/intl/en/scholar/about.html.

Systematic reviewers must download hundreds or thousands of search results for later screening from a suite of different databases. At present, Google Scholar is only cursorily searched in most reviews (i.e. by examining the first 50 search results). The addition of Google Scholar as a resource for finding additional academic and grey literature has been demonstrated to be useful for systematic reviews (Haddaway et al. in press). Automating searches and transparency documenting the results would increase transparency and comprehensiveness of the reviews with a highly resource-efficient activity at little additional effort for reviewers. These implications apply equally to other situations where web-based searching is beneficial but potentially time-consuming.

Web scrapers are an attractive technological development in the field of grey literature. The availability of a wide range of free and low-cost web scraping software provides an opportunity for significant benefits to those with limited resources, particularly researchers working alone or small organisations. Future developments will make use of the software even easier; for example the one click, automatic training provided by Import.io (https://magic.import.io). Web scrapers can increase resource-efficiency and drastically improve transparency, and existing networks can benefit through readily sharable trained APIs. Furthermore, many programs can be easily used by those with minimal or no skill or prior knowledge of this form of information technology. Researchers could benefit substantially by investigating the applicability of web scraping to their own work.

Acknowledgments

The author wishes to thank MISTRA EviEM for support during the preparation of this manuscript.

References

Allen C, Richmond K (2011) The Cochrane Collaboration: International activity within Cochrane Review Groups in the first decade of the twenty-first century. Journal of Evidence-Based Medicine 4(1):2–7

Bernes, C., Carpenter, S.R., Gårdmark, A., Larsson, P., Persson, L., Skov, C., Speed, J.D.M., Van Donk, E. (2015). What is the influence of a reduction of planktivorous and benthivorous fish on water quality in temperate eutrophic lakes? A systematic review. Environmental Evidence

Collaboration for Environmental Evidence. 2013. Guidelines for systematic review and evidence synthesis in environmental management. Version 4.2. Environmental Evidence. Available from www.environmentalevidence.org/Documents/Guidelines/Guidelines4.2.pdf (accessed January 2014).

Gurevitch, J. and Hedges, L.V. 1999. Statistical issues in ecological meta-analyses. – *Ecol.* 80: 1142-1149.

Haddaway, N.R. and Bayliss, H.R. 2015. Shades of grey: two forms of grey literature important for conservation reviews. In press

Haddaway, N. R., Burden, A., Evans, C., Healey, J. R., Jones, D. L., Dalrymple, S. E., Pullin, A. S. (2014) Evaluating effects of land management on greenhouse gas fluxes and carbon balances in boreo-temperate lowland peatland systems. Environmental Evidence, 3:5.

Haddaway, N.R., Collins, A.M., Coughlin, D., Kirk, S. (2015) The role of Google Scholar in academic searching and its applicability to grey literature searching. PLOS ONE, in press.

Kicinski, M., Springate, D. A., Kontopantelis, E. (2015). Publication bias in meta‐analyses from the Cochrane Database of Systematic Reviews. Statistics in Medicine, 34: 2781-2793.

Lortie, C.J. 2014. Formalized synthesis opportunities for ecology: systematic reviews and meta-analyses. – *Oikos* 123: 897-902.

Pullin, A. S., Bangpan, M., Dalrymple, S. E., Dickson, K., Haddaway, N. R., Healey, J. R., Hauari, H., Hockley, N., Jones, J. P. G, Knight, T., Vigurs, C., Oliver, S. (2013) Human well-being impacts of terrestrial protected areas. Environmental Evidence, 2:19.

Walker, D., G. Bergh, E. Page, and M. Duvendack. 2013. Adapting a Systematic Review for Social Research in International Development: A Case Study from the Child Protection Sector. London: ODI.