Haddaway *et al. Environ Evid (2017) 6:1*
DOI 10.1186/s13750-016-0079-2

Environmental Evidence

METHODOLOGY Open Access

# A rapid method to increase transparency and efficiency in web-based searches

Neal R. Haddaway[1*], Alexandra M. Collins[2,3], Deborah Coughlin[3,4] and Stuart Kirk[3,5]

## Abstract

**Background:** Many online search facilities allow searching for academic literature. The majority are bibliographic databases that catalogue published research in an iterative, semi-automated manner, e.g. Web of Science Core Collections, which indexes articles published in selected journals. Other resources, such as Google Scholar, identify academic articles by using search engines that crawl the internet for potentially relevant information. Often, systematic reviewers wish to document their searches for transparency or later screening. Indeed, such transparency is a cornerstone of systematic review methodology. Whilst bibliographic databases typically allow users to extract search results as citations in bulk, several other key resources, such as Google Scholar and organisation websites, do not allow this: citations must be extracted individually, which is often prohibitively time consuming.

**Methods:** Here, we describe novel methods for downloading results from searches of websites and web-based search engines into comprehensive databases as citations using free-to-use software. Citations from web-based search engines can then be integrated into review procedures along with those from traditional online bibliographic databases.

**Results and conclusions:** These methods substantially increase transparency and repeatability when searching online resources. They may also significantly reduce resource requirements for such searches and therefore represent a significant increase in efficiency.

**Keywords:** Grey literature, Systematic review, Repeatability, Literature, Searching, Evidence, Web searching

## Background

Researchers commonly perform searches using bibliographic databases, such as web of science. Bibliographic databases contain citations for articles published in academic publications and studies can generally be searched for using title, abstract or keyword search terms, and restrictions based on particular publication dates, authors, or journals can be included. Recently, free-to-use web-based search engines for academic literature, such as Google Scholar, have provided a potential alternative to subscription-based bibliographic databases [1–3]. These web-based search engines automatically catalogue new academic publications from across the internet and return results sorted by relevance according to an undisclosed algorithm, whereas bibliographic databases are populated via a systematic process based on newly published articles appearing in a selection of journals (e.g. https://oneentry.wordpress.com/faq-web-of-science), meaning individual bibliographic databases are less comprehensive than resources such as Google Scholar [4]. There are distinct advantages and disadvantages of each type of resource, as summarised in Table 1.

Academics often rely on resources such as Google Scholar when searching for academic information. One academic at Carleton University recently Tweeted: "I would be lost without Google Scholar—it is the single most important research tool that I use..." [Professor Steven Cooke @SJC_fishy, 2015]. Other researchers use internet searches to complement bibliographic databases when looking for articles.

In systematic reviews searches for grey literature are a critical means of mitigating publication bias. As a result,

*Correspondence: neal_haddaway@hotmail.com
[1] MISTRA EviEM, Stockholm Environment Institute, Box 24218, Stockholm, Sweden
Full list of author information is available at the end of the article

© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

**Table 1 Advantages and disadvantages of using bibliographic databases and web-based search engines**

| | Bibliographic databases | Academic search engines |
|---|---|---|
| Advantages | Citations are entered into individual databases according to selected journals so all entries are legitimate academic sources<br>Duplicates are uncommon [0.01% ± 0.0 (SD) (n = 10,000 records) (Haddaway et al. in press)]<br>Citations are classified according to predetermined subject categories<br>Full citation information is typically provided; including full abstracts (where available), author contact information and keywords<br>Some grey literature may be included (such as conference proceedings and theses) | Citations are identified automatically according to multiple criteria, including; the presence of a bibliography, a title followed by authorship (see http://scholar.google.co.uk/intl/en/scholar/inclusion.html for details)<br>Citations are collated on approximately a weekly basis so they are very up-to-date<br>All citations are included, not just those from specific journals or publishers<br>A wide range of grey literature is also included (such as organisational/government reports)<br>Access to the search facility is free-of-charge<br>Introduction of typographical mistakes in catalogued citations is avoided since citations are taken directly from their sources<br>Articles published on multiple sites are combined into single entries to minimise duplication |
| Disadvantages | Not all legitimate academic sources are included<br>Inclusion may take considerable time, sometimes several months or more since first appearing online<br>Typically cannot include 'online first' manuscripts that are published online but not in print<br>Subscription costs may be high<br>Typographical mistakes may be introduced as citations are transcribed manually (although many databases receive citations electronically) | Duplicates are relatively common [0.6% ± 0.6 (SD) (n = 6988 records) for full text searches and 3.1% ± 1.5 (SD) (n = 4194 records) (Haddaway et al. in press)] [duplicates are caused by errors during automatic text scanning and manual entry]<br>Mistakes at source are copied into collated citations, contributing to duplicate entries where citations are published on multiple sites<br>Citations are not full, since titles, abstracts and journal names are often shortened to fit into limited space<br>Citations may only be exported singly<br>Reproducibility of searches may be particularly low in comparison to bibliographic databases |

best practice is to combine searches of bibliographic databases with web-based academic search engines and organisation websites [5]. The relative strengths and weaknesses highlighted in Table 1 mean that a combined approach can be particularly productive in undertaking a comprehensive search for academic and grey literature (i.e. one that is able to detect a high proportion of the relevant literature). However, searching for grey literature in environmental reviews is often difficult and time consuming due the diverse sources in which it is found.

When synthesising evidence in a systematic review it is vital that processes are documented transparently to ensure a high level of repeatability [5, 6]. Transparency also allows verification of objectivity in the methods used by the review to identify and include/exclude studies. In turn, this allows the reader to make judgements regarding the robustness of the review's findings, increasing trust in the review's outputs. As environmental decision making is often dominated by multiple stakeholders with optimal solutions hard to identify, transparency is valued as it can help to increase confidence in the decisions that the review informs. Transparency can likewise improve the rigour and credibility of other review methods that involve searches for information, such as traditional literature reviews and meta-analyses, demonstrating that an unbiased approach to the collation of information has been used [7]. Whilst searching and screening (i.e. which articles are found, and which are included at what stage) of bibliographic databases can be documented readily; documenting searches of organisation websites and search engines (such as Google Scholar) may be much less transparent [8].

Given the ability to publish supplementary material in the majority of online academic journals, authors of reviews should make efforts to fully document what evidence has been assessed. Not only does this provide proof of activities, but it also allows the reader to examine in detail how the review was undertaken. The academic community is becoming increasingly aware of the *reproducibility crisis* [9]: the inability to reproduce the findings of a worrying volume of published studies. By fully documenting searching and screening activities, the analytical reproducibility (i.e. the ability to arrive at the same conclusions of a review given an identical set of search results) of reviews can be substantially improved. Furthermore, by transparently documenting the outputs of searching, review repeatability (the ability to replicate a study based on the methods provided) is enhanced. Since the internet is not a static entity (made worse by undisclosed changes to algorithms of key web-based search engines) searching via web-based tools can never be truly reproducible. However, reviewers can maximise repeatability and analytical reproducibility (i.e. screening onwards) by using the methods described herein.

Here, we describe methods for transparently documenting searches (i.e. both the search process and the search results) of various internet-based resources, including major academic search engines, using several novel software programs. Two approaches are described. The first approach ('Searches for grey literature', below) uses web-crawling software to collate data primarily from organisation website search results into one database in the form of detailed citations that can be readily updated, combined and modified. The second approach ('Extraction of full citations from search engines') uses download management software alongside web-crawling software to allow search results from academic search engines to be extracted as citations. We do not advocate the use of search engines such as Google Scholar as alternatives to bibliographic databases in systematic reviews: this has been demonstrated to lack comprehensiveness and repeatability [1]. Rather, we provide methods to transparently document the use of web-based search engines to maximise their effectiveness as supplementary sources of studies in a review. We detail two case studies that demonstrate the functionality of these methods.

Both methods described here make use of free-to-use web crawling software Import.io [10]. This was initially developed to allow detailed information on competitors and their pricing to be extracted and updated regularly by web-based businesses [11]. Here we describe methods that can apply the use of this software to web based searches and highlight the suitability of these methods using examples from each of the two approaches. We have found the software easy to use, and is supported by detailed help files, videos and personal assistance. Some of these methods require a moderate level of understanding of computing, but do not require a high degree of expertise, whilst other methods are simple to use.

## Searching for grey literature

Searches of organisation websites required to find grey literature can be simplified using web-crawling software to combine and transparently document searches, downloading citations into updatable databases. Extracted information may include: document/page titles, authorship, publication date, descriptions, and links to further descriptive information and full texts.

Web crawling software works by using Application Program Interfaces (APIs), which can be trained to detect repeated patterns in the text of source HTML files that indicate the presence of potentially useful data [12]. The method is not unlike text mining [13], but specifically mines web page code for patterns that meet certain predefined criteria. APIs allow an external program to interact with websites to extract information in a programmable way. These APIs can also be used to search organisation websites and allow results (see Fig. 1) to be downloaded as citations.

The main purpose of web crawling software is to search existing web-based search engines to extract data into a patterned database (i.e. downloading search results). Additionally, it can crawl across a specific domain (i.e. an entire website), extracting patterned data from all linked web pages. The former role can only examine indexed web pages, whilst the latter allows pages from the deep web (parts of the internet not indexed by search engines) to be examined.

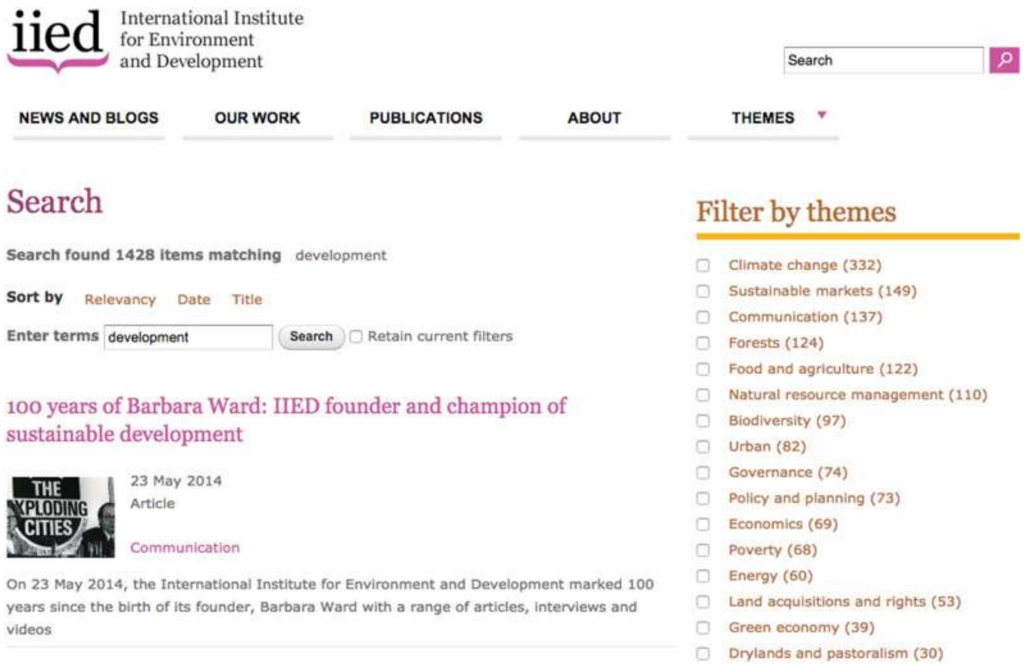
Web crawling software can work in three alternative ways, which enable searching where information is presented on web pages in differing formats.

1. Firstly, information on a web page that contains a table of data or any other relatively well-structured HTML code with a variety of columns (e.g. web-based search results) can be extracted into an offline database using a tool known within Import.io as an *extractor*.
2. Secondly, APIs, referred to within Import.io as *crawlers*, can be used to extract data from pages that are similar across a website. This works by training the crawler to perform a specified number of clicks so that information across web-pages with a similar layout can be obtained. This enables entire websites with multiple pages to be screened and information extracted into one database.
3. Thirdly, websites with built-in search facilities can be queried, using APIs known as *connectors*. Once the API has been trained, multiple websites can be searched simultaneously using key terms from within the web crawling software, with all search results extracted into one database. These queries can be refreshed and modified from within the web crawling software without having to revisit the individual pages and undertake multiple searches.

All of these APIs tools are refreshable, sharable and reusable, with minor editing necessary as and when websites are modified.

- Training APIs

Web crawling APIs are easily trained to recognise patterned data by the user highlighting which patterns in the data on a web-page constitute the required information and adding this to a 'row' within the final database of search results. Once the APIs have been trained to fill rows, individual columns must be created by highlighting, for example, a title, the authorship, and the publication date. Occasionally, several examples are needed so that the software can consistently recognise patterns; particularly where formatting across the pages of



Search

Search found 1428 items matching development

Sort by Relevancy Date Title

Enter terms development Search □ Retain current filters

100 years of Barbara Ward: IIED founder and champion of sustainable development

23 May 2014
Article

Communication

On 23 May 2014, the International Institute for Environment and Development marked 100 years since the birth of its founder, Barbara Ward with a range of articles, interviews and videos

Filter by themes

□ Climate change (332)
□ Sustainable markets (149)
□ Communication (137)
□ Forests (124)
□ Food and agriculture (122)
□ Natural resource management (110)
□ Biodiversity (97)
□ Urban (82)
□ Governance (74)
□ Policy and planning (73)
□ Economics (69)
□ Poverty (68)
□ Energy (60)
□ Land acquisitions and rights (53)
□ Green economy (39)
□ Drylands and pastoralism (30)

**Fig. 1** Results of an organisation website search (the International Institute for Environment and Development). Displayed are the extractable data, including: title and link to the document, publication date, format, topic area, and description

a website is variable. Along with text and images, URL links can be extracted by setting the type of data within a column to a 'link', allowing the user to proceed directly to the relevant page from within the downloaded database. Once the structure of the API has been established, a small number of example pages must be checked for consistency, but this is a quick process using the training already provided. Finally the API is checked by the software to ensure it can be repeatedly queried.

The most relevant APIs for academic purposes are likely to be connectors, since these can combine website search facilities into one tool that could save considerable time and will significantly increase transparency. For example, rather than having to visit each website and enter terms into their search facilities individually, one search could be performed for multiple websites and the results returned into one combined database. However, extractors and crawlers may also prove useful in academia. For example, regularly updated lists (e.g. a list species identified within a national park) could be extracted instantaneously as they are populated with new information. Similarly, websites could be crawled regularly to extract information from new pages that match a specific pattern, such as descriptions of reports or meetings.

Where a diverse range of websites must be searched and search terms cover a range of different topics, reviewers may not wish to use the same terms and strategies for every website. In such cases, connectors may not always be appropriate. However, in many cases reviewers will wish to use the same search terms across multiple websites, e.g. in systematic reviews. In these instances connectors may prove to be particularly resource efficient.

Some websites that use Google Custom Search as their website search facility are unable to be queried remotely via web crawling software as a result of restrictions put in place by Google. For these websites, search results can be extracted once as a snapshot of a particular search, but cannot be modified or combined with other APIs as described above (see "Searching for grey literature" section).

- Schema

In order to ensure consistency where several APIs are combined to produce one database, a standard column labelling system should be used (also known as a *schema*). As such, the column names used when training APIs should be identical across different websites to allow the web crawling software to stack matching columns together. The following column names may be useful templates: *title, publication_date, authors, format, url (as a link), full_text_url (as a link), description, organisational_label, subject_category, source/publisher.*

- Time requirements

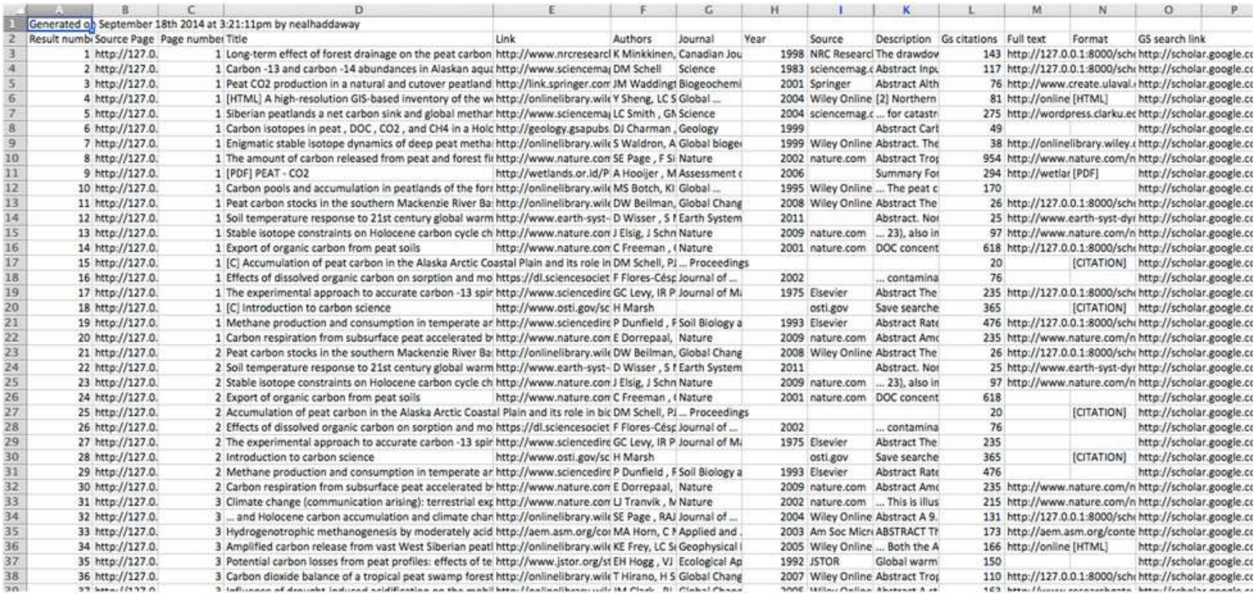
APIs can be established for websites as extractors, crawlers and connectors easily and rapidly in 5–10 min. Crawlers that extract data from multiple pages across a website may take longer if link depth (the number of pages through which all links are followed) is high (i.e. greater than 10) and websites are large. Once established, APIs can easily be updated to reflect changes in website design: web crawling software such as Import.io will regularly check the functioning of APIs, notifying the user when updates are necessary. Based on the authors' experiences of working with this software and organisation websites for systematic review searches for grey literature, minor updates are necessary a few times per year, but the frequency will vary hugely depending on the remit and resources of the organisation. Querying and refreshing existing APIs can be done instantaneously, approximately as fast as the original website can return search results.

- Outputs

The outputs of web crawling organisation websites are databases (e.g. spreadsheets) of search results separated into multiple columns as provided on each website. The outputs are downloadable in a variety of formats, including CSV and XLS files (Fig. 2). Along with the extracted data, the trained APIs can be shared amongst users, increasing time-efficiency as training time is negated. Users can also refresh their own APIs as and when a repeat search is necessary. The most notable beneficial output from this process is the entirely transparent documentation of searches (i.e. both search strategies and search results).

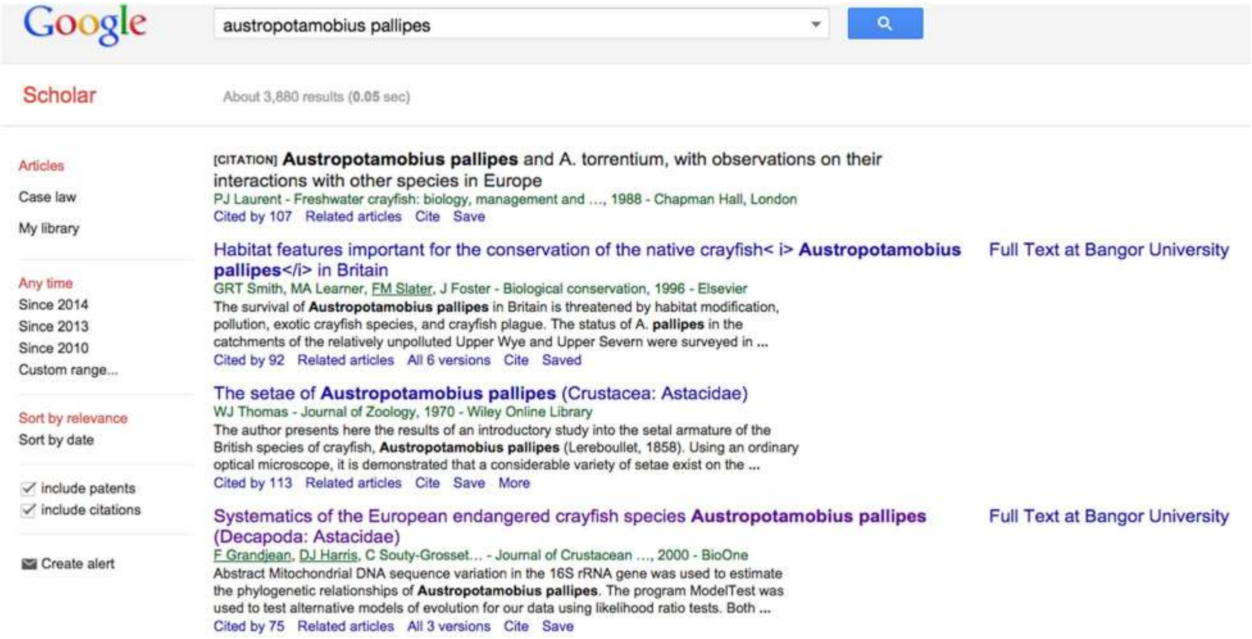
## Extraction of full citations from search engines

Search engines are powerful means of identifying potentially relevant information on the internet. This is particularly true for those with advanced settings, such as Google Scholar, which include author, journal, and date range options, along with Boolean operators and title versus full text search options [see 14]. Search engines often display multiple types of descriptive data for each search result (see Fig. 3), including: authorship (and links to author profiles), publication date, excerpts of summaries or abstracts, source publication and publisher, article formats, links to full descriptive information and full texts, and numbers of citations of and cited references within



| | A | B | C | D | E | F | G | H | I | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Generated on September 18th 2014 at 3:21:11pm by nealhaddaway | | | | | | | | | | | | | | |
| 2 | Result numb | Source Page | Page number | Title | Link | Authors | Journal | Year | Source | Description | Gs citations | Full text | Format | GS search link | |
| 3 | 1 | http://127.0. | 1 | Long-term effect of forest drainage on the peat carbon | http://www.nrcresearcl | K Minkkinen, | Canadian Jou | 1998 | NRC Researcl | The drawdov | 143 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 4 | 2 | http://127.0. | 1 | Carbon -13 and carbon -14 abundances in Alaskan aqua | http://www.sciencemag | DM Schell | Science | 1983 | sciencemag.c | Abstract Inpu | 117 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 5 | 3 | http://127.0. | 1 | Peat CO2 production in a natural and cutover peatland | http://link.springer.com | JM Waddingt | Biogeochemi | 2001 | Springer | Abstract Alth | 76 | http://www.create.ulaval. | | http://scholar.google.cc | |
| 6 | 4 | http://127.0. | 1 | [HTML] A high-resolution GIS-based inventory of the wo | http://onlinelibrary.wile | Y Sheng, LC S | Global ... | 2004 | Wiley Online | [2] Northern | 81 | http://online | [HTML] | http://scholar.google.cc | |
| 7 | 5 | http://127.0. | 1 | Siberian peatlands a net carbon sink and global methar | http://www.sciencemag | LC Smith , GM | Science | 2004 | sciencemag.c | ... for catastr | 275 | http://wordpress.clarku.ec | | http://scholar.google.cc | |
| 8 | 6 | http://127.0. | 1 | Carbon isotopes in peat , DOC , CO2 , and CH4 in a Holc | http://geology.gsapubs | DJ Charman , | Geology | 1999 | | Abstract Cart | 49 | | | http://scholar.google.cc | |
| 9 | 7 | http://127.0. | 1 | Enigmatic stable isotope dynamics of deep peat metha | http://onlinelibrary.wile | S Waldron, A | Global bioge | 1999 | Wiley Online | Abstract. The | 38 | http://onlinelibrary.wiley.c | | http://scholar.google.cc | |
| 10 | 8 | http://127.0. | 1 | The amount of carbon released from peat and forest fi | http://www.nature.com | SE Page , F Si | Nature | 2002 | nature.com | Abstract Trop | 954 | http://www.nature.com/n | | http://scholar.google.cc | |
| 11 | 9 | http://127.0. | 1 | [PDF] PEAT - CO2 | http://wetlands.or.id/P | A Hooijer , M | Assessment c | 2006 | | Summary For | 294 | http://wetlar | [PDF] | http://scholar.google.cc | |
| 12 | 10 | http://127.0. | 1 | Carbon pools and accumulation in peatlands of the for | http://onlinelibrary.wile | MS Botch, KI | Global ... | 1995 | Wiley Online | ... The peat c | 170 | | | http://scholar.google.cc | |
| 13 | 11 | http://127.0. | 1 | Peat carbon stocks in the southern Mackenzie River Ba | http://onlinelibrary.wile | DW Beilman, | Global Chang | 2008 | Wiley Online | Abstract The | 26 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 14 | 12 | http://127.0. | 1 | Soil temperature response to 21st century global warm | http://www.earth-syst- | D Wisser , S M | Earth System | 2011 | | Abstract. Nor | 25 | http://www.earth-syst-dyr | | http://scholar.google.cc | |
| 15 | 13 | http://127.0. | 1 | Stable isotope constraints on Holocene carbon cycle ch | http://www.nature.com | J Elsig, J Schn | Nature | 2009 | nature.com | ... 23), also in | 97 | http://www.nature.com/n | | http://scholar.google.cc | |
| 16 | 14 | http://127.0. | 1 | Export of organic carbon from peat soils | http://www.nature.com | C Freeman , C | Nature | 2001 | nature.com | DOC concent | 618 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 17 | 15 | http://127.0. | 1 | [C] Accumulation of peat carbon in the Alaska Arctic Coastal Plain and its role in | | DM Schell, PJ | ... Proceedings | | | | 20 | | [CITATION] | http://scholar.google.cc | |
| 18 | 16 | http://127.0. | 1 | Effects of dissolved organic carbon on sorption and mo | https://dl.sciencesociet | F Flores-Césp | Journal of ... | 2002 | | ... contamina | 76 | | | http://scholar.google.cc | |
| 19 | 17 | http://127.0. | 1 | The experimental approach to accurate carbon -13 spir | http://www.sciencedire | GC Levy, IR P | Journal of M: | 1975 | Elsevier | Abstract The | 235 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 20 | 18 | http://127.0. | 1 | [C] Introduction to carbon science | http://www.osti.gov/sc | H Marsh | | | osti.gov | Save searche | 365 | | [CITATION] | http://scholar.google.cc | |
| 21 | 19 | http://127.0. | 1 | Methane production and consumption in temperate ar | http://www.sciencedire | P Dunfield , F | Soil Biology a | 1993 | Elsevier | Abstract Rate | 476 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 22 | 20 | http://127.0. | 1 | Carbon respiration from subsurface peat accelerated b | http://www.nature.com | E Dorrepaal, | Nature | 2009 | nature.com | Abstract Amc | 235 | http://www.nature.com/n | | http://scholar.google.cc | |
| 23 | 21 | http://127.0. | 2 | Peat carbon stocks in the southern Mackenzie River Ba | http://onlinelibrary.wile | DW Beilman, | Global Chang | 2008 | Wiley Online | Abstract The | 26 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 24 | 22 | http://127.0. | 2 | Soil temperature response to 21st century global warm | http://www.earth-syst- | D Wisser , S M | Earth System | 2011 | | Abstract. Nor | 25 | http://www.earth-syst-dyr | | http://scholar.google.cc | |
| 25 | 23 | http://127.0. | 2 | Stable isotope constraints on Holocene carbon cycle ch | http://www.nature.com | J Elsig, J Schn | Nature | 2009 | nature.com | ... 23), also in | 97 | http://www.nature.com/n | | http://scholar.google.cc | |
| 26 | 24 | http://127.0. | 2 | Export of organic carbon from peat soils | http://www.nature.com | C Freeman , C | Nature | 2001 | nature.com | DOC concent | 618 | | | http://scholar.google.cc | |
| 27 | 25 | http://127.0. | 2 | Accumulation of peat carbon in the Alaska Arctic Coastal Plain and its role in bic | | DM Schell, PJ | ... Proceedings | | | | 20 | | [CITATION] | http://scholar.google.cc | |
| 28 | 26 | http://127.0. | 2 | Effects of dissolved organic carbon on sorption and mo | https://dl.sciencesociet | F Flores-Césp | Journal of ... | 2002 | | ... contamina | 76 | | | http://scholar.google.cc | |
| 29 | 27 | http://127.0. | 2 | The experimental approach to accurate carbon -13 spir | http://www.sciencedire | GC Levy, IR P | Journal of M: | 1975 | Elsevier | Abstract The | 235 | | | http://scholar.google.cc | |
| 30 | 28 | http://127.0. | 2 | Introduction to carbon science | http://www.osti.gov/sc | H Marsh | | | osti.gov | Save searche | 365 | | [CITATION] | http://scholar.google.cc | |
| 31 | 29 | http://127.0. | 2 | Methane production and consumption in temperate ar | http://www.sciencedire | P Dunfield , F | Soil Biology a | 1993 | Elsevier | Abstract Rate | 476 | | | http://scholar.google.cc | |
| 32 | 30 | http://127.0. | 2 | Carbon respiration from subsurface peat accelerated b | http://www.nature.com | E Dorrepaal, | Nature | 2009 | nature.com | Abstract Amc | 235 | http://www.nature.com/n | | http://scholar.google.cc | |
| 33 | 31 | http://127.0. | 3 | Climate change (communication arising): terrestrial exp | http://www.nature.com | LJ Tranvik , M | Nature | 2002 | nature.com | ... This is illus | 215 | http://www.nature.com/n | | http://scholar.google.cc | |
| 34 | 32 | http://127.0. | 3 | ... and Holocene carbon accumulation and climate char | http://onlinelibrary.wile | SE Page , RAJ | Journal of ... | 2004 | Wiley Online | Abstract A 9. | 131 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 35 | 33 | http://127.0. | 3 | Hydrogenotrophic methanogenesis by moderately acid | http://aem.asm.org/coi | MA Horn, C N | Applied and . | 2003 | Am Soc Micri | ABSTRACT Th | 173 | http://aem.asm.org/conte | | http://scholar.google.cc | |
| 36 | 34 | http://127.0. | 3 | Amplified carbon release from vast West Siberian peatl | http://onlinelibrary.wile | KE Frey, LC Si | Geophysical l | 2005 | Wiley Online | ... Both the A | 166 | http://online | [HTML] | http://scholar.google.cc | |
| 37 | 35 | http://127.0. | 3 | Potential carbon losses from peat profiles: effects of te | http://www.jstor.org/st | EH Hogg , VJ | Ecological Ap | 1992 | JSTOR | Global warm | 150 | | | http://scholar.google.cc | |
| 38 | 36 | http://127.0. | 3 | Carbon dioxide balance of a tropical peat swamp forest | http://onlinelibrary.wile | T Hirano, H S | Global Chang | 2007 | Wiley Online | Abstract Trop | 110 | http://127.0.0.1:8000/sch | | http://scholar.google.cc | |
| 39 | 37 | http://127.0 | 3 | Influence of drought-induced acidification on the mobil | http://onlinelibrary.wile | JM Clark, PJ | Global Chang | 2005 | Wiley Online | Abstract A st | 163 | http://www.researchgate | | http://scholar.google.cc | |

**Fig. 2** Downloaded database of search results from a search of Google Scholar. Search used Import.io, demonstrating the type of data that can be downloaded



About 3,880 results (0.05 sec)

[CITATION] **Austropotamobius pallipes** and A. torrentium, with observations on their interactions with other species in Europe
PJ Laurent - Freshwater crayfish: biology, management and ..., 1988 - Chapman Hall, London
Cited by 107 Related articles Cite Save

Habitat features important for the conservation of the native crayfish< i> **Austropotamobius pallipes**</i> in Britain — Full Text at Bangor University
GRT Smith, MA Learner, FM Slater, J Foster - Biological conservation, 1996 - Elsevier
The survival of **Austropotamobius pallipes** in Britain is threatened by habitat modification, pollution, exotic crayfish species, and crayfish plague. The status of A. **pallipes** in the catchments of the relatively unpolluted Upper Wye and Upper Severn were surveyed in ...
Cited by 92 Related articles All 6 versions Cite Saved

The setae of **Austropotamobius pallipes** (Crustacea: Astacidae)
WJ Thomas - Journal of Zoology, 1970 - Wiley Online Library
The author presents here the results of an introductory study into the setal armature of the British species of crayfish, **Austropotamobius pallipes** (Lereboullet, 1858). Using an ordinary optical microscope, it is demonstrated that a considerable variety of setae exist on the ...
Cited by 113 Related articles Cite Save More

Systematics of the European endangered crayfish species **Austropotamobius pallipes** (Decapoda: Astacidae) — Full Text at Bangor University
F Grandjean, DJ Harris, C Souty-Grosset... - Journal of Crustacean ..., 2000 - BioOne
Abstract Mitochondrial DNA sequence variation in the 16S rRNA gene was used to estimate the phylogenetic relationships of **Austropotamobius pallipes**. The program ModelTest was used to test alternative models of evolution for our data using likelihood ratio tests. Both ...
Cited by 75 Related articles All 3 versions Cite Save

**Fig. 3** Search results from an online search engine (Google Scholar). Figure shows information extractable as patterned data using web crawling APIs. Format (e.g. '[CITATION]'), title, title link, authorship, journal, publication year, publisher/source, description/abstract excerpt, citations (number/link), similar articles link, and full text links can all be extracted as separate columns for each result. Google and the Google logo are registered trademarks of Google Inc., used with permission

academic articles. Basic citations (i.e. authorship, publication year, title, journal, volume, and page) are manually extractable from some search engines, such as Google Scholar, but multiple citations often cannot be extracted at once, and automatic restriction facilities relating to fair use policies often limit the number of extractable citations within a given period. In the authors' experience, manual extraction of several hundred individual citations within a short period of time (i.e. several hours) is sufficient to result in a block on a particular IP address for c. 2 days. Websites typically restrict access to prevent their servers from being overloaded from malicious web traffic. This is problematic since it can reduce comprehensiveness of a search, significantly hampering efforts to conduct systematic searches using electronic search engines such as Google Scholar in a transparent manner.

We describe two easy-to-use methods that can extract search engine results in bulk. The first method builds on the web crawling method detailed above for organisation website searching allowing extraction of up to 1000 full citations from search engines such as Google Scholar. The second method uses citation management/analysis software to extract simple citations (lacking abstracts).

### Method 1—import.io method

This method can extract detailed citations including a short description (typically an extract from the abstract). It consists of a three-stage approach: preparing a list of URLs, downloading search results as HTML files, and scraping these locally saved HTML files for data. This approach is beneficial since a number of search engines prevent the use of APIs and block IP addresses in an attempt to prevent overloading by automated use of their websites. It should be noted that attempts to circumvent these restrictions may constitute a breach of the conditions of use, and users should establish whether this may be the case before proceeding.

- Step 1: Producing a list of URLs

The transparency of website searches can be rapidly improved by saving search results as HTML files, which typically preserve most information regarding the terms, dates and websites used. This also forms the first stage of the process for downloading citations from search engines. This is particularly easy in Google Scholar, where URLs of search results can be generated based on patterns within the URL itself (patterns are only evident from page 2 of the search results onwards, but can be back-generated for page 1). For example, a search for "crayfish" in Google Scholar yields the following URL for page 2: http://scholar.google.co.uk/scholar?start=10&q=crayfish&hl=en&as_sdt=0,5. This URL clearly contains the search term ('crayfish') and the starting record number ('10').

Advanced searches can also be created in this way. For example, the following Boolean search string can be used to generate an associated URL for page 3 of the search results:

*Boolean search string: "evidence-based" AND (conservation OR nature) NOT park [author: Smith, publication years: 2000 to 2014, publication name: Nature]*

*Google Scholar Search String: conservation OR nature "evidence based" -park author:smith [publication years: 2000 to 2014, publication name: nature]*

*Google Scholar URL: http://scholar.google.co.uk/scholar?start=0&q=conservation+OR+nature+%22evidence+based%22+-park+author:smith&hl=en&as_publication=nature&as_sdt=0,5&as_ylo=2000&as_yhi=2014*

Using the pattern for page numbers within the above URL it is possible to generate a sequence of up to 100 URLs for the first 1000 search results in Google Scholar. This can be done easily in spreadsheet software, such as Microsoft Excel, using a concatenate function (i.e. the function that joins cells contents together into one string; "=CONCATENATE (A1,B1,C1)") that splits the URL into two parts and replaces the start number ('0' in the Google Scholar URL above) with numbers from 0 to 990 in steps of 10, covering the first 1000 records. Google Scholar limits search results to the first 1000 records, limiting both the number of results visible online and the number of results extractable using the method detailed herein. Other search engines do not have the same limits. Systematic reviews published by the Collaboration for Environmental Evidence [5] typically screen the first 50 search results from Google Scholar [15]; the method described here facilitates the screening of substantially more results. The Additional file 1 provided allows a list of URLs for Google Scholar to be produced based on advanced search input boxes (see Additional file 2 for detailed instructions). Once a list of URLs has been produced, these URLs should be saved as an unformatted text file (TXT) for use in the next step.

- Step 2: Downloading HTML files as a snapshot of search results

In order to download search results as HTML files an add-on for the internet browser Mozilla Firefox, 'DownThemAll!', is available free of charge [16]. This software allows Firefox to download a list of URLs (using the TXT file detailed above) as HTML files, saving them locally as a snapshot of search results. Detailed instructions for this process are available as Additional file 3. The downloaded HTML files can then be scraped for data using Import.io in a similar method to that described above for organisation websites.

- Step 3: Extracting full citations from locally saved HTML files

Once HTML files have been saved locally as a snapshot of a particular search, Import.io can be used to extract full citations. An API for Google Scholar search results can be readily produced using the methods described for organisation websites above. Before data can be extracted from the downloaded HTML files, they must be made available on a local server to be accessible by Import.io. Detailed instructions for this process are provided via the Import.io website [10]. The process should take no more than a few minutes to set up.

One final step can help to preserve information regarding the page number of each HTML page of search results. This information aids ordering of search results so that the location of individual citations within the results can be preserved. A second scraper can be trained to recognise only the page numbers in the search results. A VLOOKUP formula (explained in further detail on the Microsoft Support website [17]) can then be used in Excel to look up page numbers for each saved HTML file from this second database of page numbers and HTML file names.

- Outputs

The output is similar to that of the organisation website searches described above: a database of up to 1000 search results split into a variety of columns representing aspects of the full citations. These citations can include links to further descriptions and full texts if required, which can be accessed directly from the downloaded databases.

### Method 2—citation management software method

Free to use software, namely Mendeley, 'Publish or Perish', and Zotero can be used individually to extract search engine results in bulk. Other software platforms may work in a similar way and the methods detailed here will likely apply to other programmes.

- Mendeley (https://www.mendeley.com/join/)

Mendeley is a reference management programme for research authors, which stores citations and full text articles in the cloud. The programme includes an add-on for internet browsers that allows the user to extract citations from within a web site, including from a page of up to 10 search results from search engines such as Google Scholar, one page at a time. Thus, users can extract results one page at a time. Citations do not include abstracts, but they do include full text PDFs where these are freely available. Search results are saved to the user's Mendeley account and can be accessed through the desktop programme. Google Scholar search results that are available as 'citations' only (i.e. references from within other articles with no active link to a publisher's site) are not extractable with this method.

- Publish or Perish (http://www.harzing.com/resources/publish-or-perish)

This free software is designed for authors to track citations to their papers, but the programme includes a facility for searching Google Scholar using the advanced search function (e.g. basic Boolean operators and title or full text searching). Using this facility, researchers can extract up to 1000 search results as basic citations. These citations include details of authorship, publication year, title, source journal, volume and pages, and include a link to the publisher's web site. No abstract is extracted. Users can extract the search results in a variety of formats, including excel files, comma separated value files, or reference management software files (e.g. RIS). Google Scholar search results that are available as 'citations' only (i.e. references from within other articles with no active link to a publisher's site) are not extractable with this method.

- Zotero (https://www.zotero.org/)

Zotero is a free programme for reference management, similar to Mendeley. Zotero also has an internet browser add-on that allows the user to download citations from a web site such as a page of results in Google Scholar, allowing the extraction of up to 20 citations from each page, one page at a time. These results are saved to the user's local (i.e. non-cloud-based) Zotero library. Abstracts are not extracted using this software but full texts are, where they are freely available. Google Scholar 'citations' are also extracted as citations.

### Comparison of methods

There are advantages and disadvantages of the methods described to extract full citations from search engines (i.e. citation management/analysis software versus web-crawling software). Reference management/analysis software allows more precise and complete extraction of citation titles, whilst the web-crawling method will extract only what is displayed by Google Scholar, which may result in incomplete records where titles are long. Some of the methods cannot extract results displayed as 'citations' (e.g. Mendeley), whilst others can (e.g. Zotero and web-crawling). Reference management software (e.g. Mendeley and Zotero) typically extract full text

PDFs where they are freely available, whilst web-crawlers and cannot. Neither of these methods is able to extract more than 1000 search results from Google Scholar. The Import.io method, whilst more laborious, allows the user to define precisely what information is extracted and is usable across a range of different search engines, whilst the other methods using citation management software are typically restricted to one or a few pre-specified search facilities (e.g. Google Scholar and PubMed).

## Case studies

We describe below two case study searches that were performed for organisation websites and for a web-based search engine (Google Scholar).

### Searches for grey literature

A total of 40 commonly cited websites were identified from an assessment of 57 recent systematic reviews published in the journal *Environmental Evidence* (Table 2). These were selected to cover a diverse range of organisation websites that belong to organisations that deal with environmental management topics for systematic reviews of a diverse range of subjects, for example "What are the human well-being impacts of terrestrial protected areas?" [18] and "What is the impact of land management of lowland peatlands on greenhouse gas fluxes and carbon cycling?" [19].

Updatable APIs were successfully established for 30 of the websites identified by using the methods outlined above, meaning that these sites could be searched together using a set of predefined search terms across all sites simultaneously, returning typically up to 200 results per search term for each site into one common database. In each case an API could be established in under 5 min. For these websites, up to 10 pages of search results (typically 10 results per page depending on the number of results displayed per page) were extracted using updatable search strings. Details of this process are outlined in Table 2 and outputs of the searching are available in Additional file 4.

For 10 websites, full search results could not be extracted since updatable APIs could not be constructed. The reasons for these failures are not fully clear, but in 4 cases this occurred because the websites employed a Google Custom Search facility, which does not permit updatable APIs to be created. However, between 20 and 50 search results were still extractable based on the search terms used during API training. For the remaining 6 websites, problems may have related to the use of complex site coding or page structuring that was not readily recognisable as patterned data. For the websites where updatable APIs could not be created, training of the first 2 pages of results allowed between 10 and 50 results to be extracted per search string and included in the results. The results of a combined, updatable search of these websites are provided in supplementary information, detailing the types of citation information extractable from the included websites (Additional file 4).

### Extraction of full citations from search engines

A search update was undertaken for a systematic review on the impacts of tillage intensity on soil organic carbon (SOC) in arable farmland (according to a protocol [20]), which followed on from a systematic map on the impacts of agricultural management on SOC [21]. The original systematic map employed hand searching of Google Scholar and reporting of any relevant results found within the first 100 hits. Conversely, Import.io was used to download up to the maximum of 1000 search results from Google Scholar in the search update of the tillage systematic review. These records were then screened at title, abstract and full text stages parallel to the search update for bibliographic databases. As part of this search update, full text and title searches of Google Scholar were undertaken using an adapted search string, as follows:

*Full Boolean search string*: soil* AND (arable OR agricult* OR farm* OR crop* OR cultivat*) AND (till* OR "no till*" OR "reduced till*" OR "direct drill*" OR "conservation till*" OR "minimum till*") AND ("soil organic carbon" OR "soil carbon" OR "soil C" OR "soil organic C" OR SOC OR "carbon pool" OR "carbon stock" OR "carbon storage" OR "soil organic matter" OR SOM OR "carbon sequestrat*" OR "C sequestrat*")

*Adapted Google Scholar search string*: soil AND carbon AND (till OR tillage OR "reduced tillage" OR "conservation tillage" OR "no tillage" OR "direct drill" OR "minimum till") [Google Scholar format: soil carbon till OR tillage OR "reduced tillage" OR "conservation tillage" OR "no tillage" OR "direct drill" OR "minimum till"]

Title and full text searches returned 163 and 23,200 results respectively (03/09/2015). All 17 pages of title search results and the first 100 pages of full text results were downloaded as HTML files using the DownThemAll! Add-on for Mozilla Firefox as HTML files. The two sets of HTML files were scraped for citations within Import.io on a Windows operating system using Mongoose Web Server [22] by training the software to recognise titles, URL links, authors, publication years, publications, short descriptions, source (publisher), number of Google Scholar citations and URL links to full texts in the citation information where available. Page numbers were extracted from each HTML search result page using a second crawler as described above. The resulting databases of citations are provided in Additional file 5 and a

**Table 2 Organisation websites commonly used within systematic reviews**

| Organisation | Crawler functioning (yes/no) | Notes | URL | Numbers of results extracted |
|---|---|---|---|---|
| Alterra (wageningen) | Yes | | www.wageningenur.nl/en/Expertise-Services/Research…/alterra.htm | 200 |
| BirdLife | No | Google custom search; extractable but no API, 20 results only per query | www.birdlife.org/ | 40 |
| CAFOD | Yes | | www.cafod.org.uk/ | 200 |
| Canadian Forestry Service | Yes | | www.nrcan.gc.ca/forests | 400 |
| CGIAR | Yes | | www.cgiar.org/ | 200 |
| CIFOR | Yes | | www.cifor.org/ | 200 |
| Ecosystem Marketplace | Yes | | http://www.ecosystemmarketplace.com/ | 600 |
| ELDIS | No | Extractable but no API, 20 results only per query | www.eldis.org/ | 40 |
| Environment Agency | Yes | | https://www.gov.uk/search?q=&filter_organisations%5B%5D=environment-agency | 400 |
| Environment Canada | Yes | | https://www.google.se/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&cad=rja&uact=8&sqi=2&ved=0CDsQFjAJahUKEwj_8Kj1wZ7IAhUC3SwKHZ9GCV4&url=http%3A%2F%2Fwww.ec.gc.ca%2F%3Flang%3DEn&usg=AFQjCNH1nH5cS8da11EdKH_6SLxXfle_0w&sig2=CdAV1RtbuTMKJgmPH6SI3Q&bvm=bv.103627116,d.bGg | 200 |
| Forestry Commission | No | Google custom search; extractable but no API, 20 results only per query | www.forestry.gov.uk/ | 40 |
| GEF | Yes | | https://www.thegef.org/ | 200 |
| Greenpeace | Yes | | www.greenpeace.org/ | 200 |
| Greenpeace Publication | Yes | | www.greenpeace.org/international/en/publications/ | 179 |
| ICIMOD | Yes | | www.icimod.org/ | 200 |
| ICIMOD Publications | Yes | | www.icimod.org/publications | 245 |
| IIED | Yes | | www.iied.org/ | 68 |
| IIED Publications | Yes | | pubs.iied.org/ | 200 |
| IUCN Publications | Yes | | www.iucn.org/ | 76 |
| Natural England | Yes | | https://www.google.se/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CB8QFjAAahUKEwi4iZ-Awp7IAhVHkCwKHS6WD4E&url=https%3A%2F%2Fwww.gov.uk%2Fgovernment%2Forganisations%2Fnatural-england&usg=AFQjCNFjk4k6mmZ29TcR2_j7VZ_GWh-7fg&sig2=v54ScjLDWQ1sWzpRhleoKg&bvm=bv.103627116,d.bGg | 341 |

**Table 2 continued**

| Organisation | Crawler functioning (yes/no) | Notes | URL | Numbers of results extracted |
|---|---|---|---|---|
| Natural Resources Canada | No | Extractable but no API, 10 results only per query | www.nrcan.gc.ca/home | 19 |
| Natural Resources Canada Publications | Yes | | www.nrcan.gc.ca/publications/1138 | 400 |
| Northern Ireland Environment Agency Publications | Yes | | www.doeni.gov.uk/niea/ | 200 |
| ODI | Yes | | www.odi.org/ | 200 |
| ODI Publications | Yes | | www.odi.org/publications | 200 |
| OECD | No | Google custom search; extractable but no API, 20 results only per query | www.oecd.org/ | 40 |
| RAMSAR | Yes | | www.ramsar.org/ | 400 |
| RSPB | Yes | | www.rspb.org.uk/ | 200 |
| SEPA | Yes | | http://www.sepa.org.uk/library/content-search/?q=&LibGo=Search&page=1 | 137 |
| TEAGASC | No | Google custom search; extractable but no API, 20 results only per query | www.teagasc.ie/ | 32 |
| The Nature Conservancy | No | Extractable but no API, 20 results only per query | www.nature.org/ | 40 |
| Tropenbos | Yes | | www.tropenbos.org/ | 200 |
| UNEP WCMC | No | Extractable but no API, 50 results only per query | http://www.unep-wcmc.org/resources-and-data | 73 |
| US EPA | Yes | | http://nlquery.epa.gov/epasearch/epasearch?querytext=&typeofsearch=epa&doctype=all&originalquerytext=water+quality+trading&areaname=&faq=true&filter=sample4filt.hts&fld=&sessionid=F49B4A4D12A56F4638CAE8FB3DB62382&referer=&prevtype=epa&result_template=2col.ftl&stylesheet= | 2561 |
| USAID | Yes | | http://www.usaid.gov/gsearch/ | 200 |
| USFWS | Yes | | www.fws.gov/ | 400 |
| Wetlands International | Yes | | www.wetlands.org/ | 200 |
| Wildlife Conservation Society | No | Extractable but no API, 10 results only per query | www.wcs.org/ | 20 |
| World Bank | No | Extractable but no API, 20 results only per query | http://search.worldbank.org/all?qterm=&title=&filetype= | 40 |
| WWT | Yes | | www.wwt.org.uk/ | 140 |

Listed are reviews published by *Environmental Evidence* that were searched for literature to demonstrate the applicability of web crawling to academic searching sample of the title search database is displayed in Fig. 4. All citations were extracted successfully, along with page numbers and result location (result row within pages).

The search update from Google Scholar resulted in one new publication being used in meta-analyses that would otherwise not have been discovered through bibliographic database and organisation website searches (Haddaway, unpublished data).

In addition to this example, scraping of Google Scholar search results was used extensively in a study of the use of Google Scholar in evidence reviews. This study examined the nature and frequency of grey literature within the first 1000 results from full text and title searches of Google Scholar based on seven systematic review case study search strings. Details of this study can be found in Haddaway et al. [23].

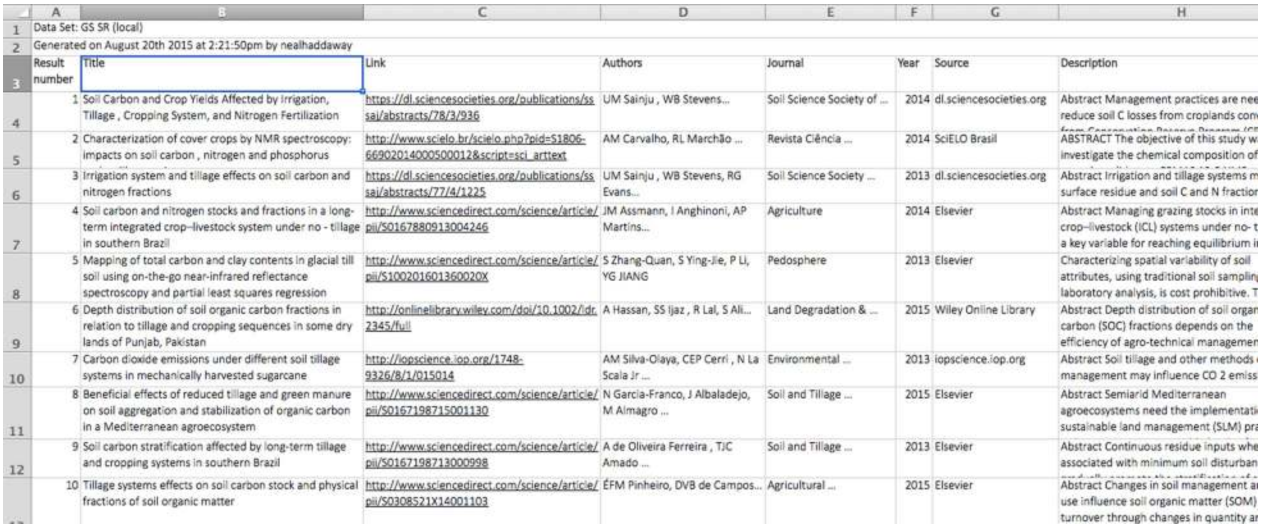
## Conclusions

The methods described herein to search for and record the details of outputs from web searches provide a high degree of transparency, either for personal reference or for those wishing to record their activities as part of a formal review, such as a systematic review, providing clear benefits to researchers. Systematic reviewers, for example, can dramatically improve the transparency of their non-bibliographic (i.e. grey literature) searches, producing databases of citations from academic search engines and organisation websites that can be subsequently screened for relevance and possible inclusion in the review.

Along with increasing transparency, these methods can lead to significant improvements in resource efficiency, since web-crawler APIs can be established and shared in efficient ways, allowing multiple resources to be searched simultaneously. Many freely available full texts are also extractable together with search results from academic search engines. In addition, lists of relevant information can be updated instantaneously.

Systematic review methods are based on a requirement for detailed reporting of methods used [24], and the same level of detail should be employed to describe both bibliographic database and other web-based searching. The methods described herein add to this high level of transparency, repeatability and reproducibility by allowing reviewers to produce a comprehensive record of the searches performed at one point in time across organisation websites and web-based search engines. This documentation could be used as freely accessible supplementary information for a highly detailed systematic review report, or it could be maintained privately to facilitate updating of the review.

We have provided methods that can assist with the recording of searches of individual websites and search engines. Whilst we do not advocate any one approach, the methods described can assist in improving the transparency, repeatability and reproducibility of systematic and other evidence reviews, thus optimising review reliability and improving the confidence end-users have in their findings.



|  | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Data Set: GS SR (local) |  |  |  |  |  |  |  |
| 2 | Generated on August 20th 2015 at 2:21:50pm by nealhaddaway |  |  |  |  |  |  |  |
| 3 | Result number | Title | Link | Authors | Journal | Year | Source | Description |
| 4 | 1 | Soil Carbon and Crop Yields Affected by Irrigation, Tillage , Cropping System, and Nitrogen Fertilization | https://dl.sciencesocieties.org/publications/sssaj/abstracts/78/3/936 | UM Sainju , WB Stevens... | Soil Science Society of ... | 2014 | dl.sciencesocieties.org | Abstract Management practices are nee reduce soil C losses from croplands con |
| 5 | 2 | Characterization of cover crops by NMR spectroscopy: impacts on soil carbon , nitrogen and phosphorus | http://www.scielo.br/scielo.php?pid=S1806-66902014000500012&script=sci_arttext | AM Carvalho, RL Marchão ... | Revista Ciência ... | 2014 | SciELO Brasil | ABSTRACT The objective of this study w investigate the chemical composition of |
| 6 | 3 | Irrigation system and tillage effects on soil carbon and nitrogen fractions | https://dl.sciencesocieties.org/publications/sssaj/abstracts/77/4/1225 | UM Sainju , WB Stevens, RG Evans... | Soil Science Society ... | 2013 | dl.sciencesocieties.org | Abstract Irrigation and tillage systems m surface residue and soil C and N fractior |
| 7 | 4 | Soil carbon and nitrogen stocks and fractions in a long-term integrated crop–livestock system under no - tillage in southern Brazil | http://www.sciencedirect.com/science/article/pii/S0167880913004246 | JM Assmann, I Anghinoni, AP Martins... | Agriculture | 2014 | Elsevier | Abstract Managing grazing stocks in inte crop–livestock (ICL) systems under no- t a key variable for reaching equilibrium i |
| 8 | 5 | Mapping of total carbon and clay contents in glacial till soil using on-the-go near-infrared reflectance spectroscopy and partial least squares regression | http://www.sciencedirect.com/science/article/pii/S100201601360020X | S Zhang-Quan, S Ying-Jie, P Li, YG JIANG | Pedosphere | 2013 | Elsevier | Characterizing spatial variability of soil attributes, using traditional soil samplin laboratory analysis, is cost prohibitive. T |
| 9 | 6 | Depth distribution of soil organic carbon fractions in relation to tillage and cropping sequences in some dry lands of Punjab, Pakistan | http://onlinelibrary.wiley.com/doi/10.1002/ldr.2345/full | A Hassan, SS Ijaz , R Lal, S Ali... | Land Degradation & ... | 2015 | Wiley Online Library | Abstract Depth distribution of soil orgar carbon (SOC) fractions depends on the efficiency of agro-technical managemen |
| 10 | 7 | Carbon dioxide emissions under different soil tillage systems in mechanically harvested sugarcane | http://iopscience.iop.org/1748-9326/8/1/015014 | AM Silva-Olaya, CEP Cerri , N La Scala Jr ... | Environmental ... | 2013 | iopscience.iop.org | Abstract Soil tillage and other methods management may influence CO 2 emiss |
| 11 | 8 | Beneficial effects of reduced tillage and green manure on soil aggregation and stabilization of organic carbon in a Mediterranean agroecosystem | http://www.sciencedirect.com/science/article/pii/S0167198715001130 | N Garcia-Franco, J Albaladejo, M Almagro ... | Soil and Tillage ... | 2015 | Elsevier | Abstract Semiarid Mediterranean agroecosystems need the implementati sustainable land management (SLM) pra |
| 12 | 9 | Soil carbon stratification affected by long-term tillage and cropping systems in southern Brazil | http://www.sciencedirect.com/science/article/pii/S0167198713000998 | A de Oliveira Ferreira , TJC Amado ... | Soil and Tillage ... | 2013 | Elsevier | Abstract Continuous residue inputs whe associated with minimum soil disturban |
|  | 10 | Tillage systems effects on soil carbon stock and physical fractions of soil organic matter | http://www.sciencedirect.com/science/article/pii/S0308521X14001103 | ÉFM Pinheiro, DVB de Campos... | Agricultural ... | 2015 | Elsevier | Abstract Changes in soil management a use influence soil organic matter (SOM) turnover through changes in quantity ar |

**Fig. 4** Database of search results from title search in Google Scholar. Advanced search was used: (allintitle: soil carbon till OR tillage OR "reduced tillage" OR "conservation tillage" OR "no tillage" OR "direct drill" OR "minimum till"; 03/09/2015) using methods described herein

The methods described here may also prove particularly useful for rapid review methods, which restrict the scope of reviews in order to allow them to be conducted within reduced timeframes, to all reviews meet time-sensitive policy needs, for example. These methods enable more rapid searching so that strict requirement of systematic reviews would not need to be compromised. By transparently documenting all search results identified, such rapid reviews may also lend themselves to more efficient updating into a full systematic review if required, since a rapid review may use a sampling approach to examine a subset of the identified results.

There are other benefits of the approaches detailed herein. Along with searching for citations, researchers often turn to the internet to find or update information from a variety of sources, for example finding up-to-date information relating to species lists, gene sequences, laboratory protocols and chemical concentrations that are listed on single websites. Rather than copying and pasting such detailed lists, web crawling software such as the tool described above allow such data to be extracted and regularly updated as necessary.

Reviews typically become out-of-date within several years of publication (i.e. approximately 1–6 years according to [25]), and transparent documentation that includes web based searches can drastically improve resource efficiency of updates, since work need not be repeated for the years covered by the original review. In a similar way, novel reviews that share overlapping evidence bases with published reviews can benefit from the inclusion of evidence from these previous reviews [26], potentially saving time and increasing comprehensiveness.

In addition to transparently recording the outputs of web-based searches, we emphasise that reviewers should also record the searching activities that generated those outputs. Such detailed information should include the following information:

- What resource was searched: the URL for the search page, the organisation's name and the access point if using a third party platform
- When it was searched: the date searches were undertaken
- How was it searched: the search terms used along with any other optional settings, such as 'title only' or document type

Web-scraping software, such as Import.io, facilitates this level of reporting by including search terms, searching dates and URLs directly into the search result database, ensuring that each search result can be traced back to its origin. For transparency and ease of interrogation, it is advisable to summarise this information in systematic review methods text or additional files describing methods used.

The two examples detailed above and evidenced by the supplementary information demonstrate that the methods described, using organisation website searching and a web-based search engine, can significantly increase transparency in online searching.

Substantial effort has been invested in making information, such as research data, freely available on the internet. As this data is updated, increased and improved, emerging methods such as the web-crawling methodology described herein can help researchers to make the most of available data. By learning from a relatively nascent commercial innovation workloads can be reduced and efficiency maximised. Those dealing with 'big data', such as systematic reviewers, may particularly benefit from these methods. Furthermore, evidence-informed decisions in any discipline should be based on transparent, objective and repeatable methods [5]. The methods described above provide an immediate, low-cost method for transparently documenting web-based searches for evidence.

## Additional files

**Additional file 1.** URL generator based on advanced search facility in Google Scholar.

**Additional file 2.** Instructions on how to use the URL generator for Google Scholar.

**Additional file 3.** Instructions on how to download web-based search engine search results using DownThemAll add-on.

**Additional file 4.** Database of search results from organisation website searches for the terms "research" and "report".

**Additional file 5.** Database of search results from Google Scholar title and full text searches for tillage systematic review case study.

### Abbreviations

API: Application Program Interfaces (file format); HTML: Hyper Text Markup Language (internet coding language); CSV: comma separated values (spreadsheet/text file format); XLS: Excel binary file format (spreadsheet file format); PDF: Portable Document Format (document file format); RIS: research information systems (citation file format); URL: Uniform Resource Identifier (World Wide Web address format); TXT: text file (document/text file format); SOC: soil organic carbon.

### Authors' contributions

NH investigated and refined the methodology. NH and AC tested the methodology. NH drafted the manuscript. NH, AC, DC and SK edited the manuscript.

### Author details

[1] MISTRA EviEM, Stockholm Environment Institute, Box 24218, Stockholm, Sweden. [2] Centre for Environmental Policy, Imperial College, London, UK. [3] Department for Environmental, Food and Rural Affairs, London, UK. [4] Department for Civil and Environmental Engineering, Imperial College, London, UK. [5] European Commission, DG Research & Innovation, Scientific Advice Mechanism, 8 Square Frère Orban, 1049 Brussels, Belgium.

**Acknowledgements**

NH was hosted at Bangor University. AC acknowledges a Policy.

Placement Fellowship funded by the Natural Environment Research Council, the UK Department for Environment Food and Rural Affairs and the Environment Agency.

**Competing interests**

The authors declare they have no competing interests.

**Funding**

The work described herein was funded as part of a project by the Department for Environment, Food and Rural Affairs under contract WT1552, project managed by the Joint Water Evidence Group.

Received: 3 August 2016 Accepted: 7 December 2016
Published online: 03 January 2017

**References**

1. Boeker M, Vach W, Motschall E. Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. BMC Med Res Methodol. 2013;13(1):1.
2. De Winter JC, Zadpoor AA, Dodou D. The expansion of Google Scholar versus web of science: a longitudinal study. Scientometrics. 2014;98(2):1547–65.
3. Gehanno J-F, Rollin L, Darmoni S. Is the coverage of Google Scholar enough to be used alone for systematic reviews. BMC Med Inform Decis Mak. 2013;13(1):1.
4. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. FASEB J. 2008;22(2):338–42.
5. CEE. Guidelines for systematic review and evidence synthesis in environmental management. Bangor: The Collaboration for Environmental Evidence. 2013.
6. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley; 2011.
7. Haddaway N, Woodcock P, Macura B, Collins A. Making literature reviews more reliable through application of lessons from systematic reviews. Conserv Biol. 2015;29(6):1596–605.
8. Briscoe S. Web searching for systematic reviews: a case study of reporting standards in the UK Health Technology Assessment programme. BMC Res Notes. 2015;8(1):1.
9. Churchill GA. When are results too good to be true? Genetics. 2014;198(2):447–8.
10. Import.io. Import.io. 2016. http://www.import.io. Accessed 22 Nov 2016.
11. Pant G, Menczer F, editors. Topical crawling for business intelligence. International conference on theory and practice of digital libraries. Berlin:Springer; 2003.
12. Pant G, Srinivasan P, Menczer F. Crawling the web in web dynamics. Berlin: Springer; 2004. p. 153–77.
13. Toon E, Timmermann C, Worboys M. Text-mining and the history of medicine: big data, big questions? Med Hist. 2016;60(02):294–6.
14. Scholar G. Google Scholar. 2016. http://scholar.google.co.uk.
15. Roe D, Fancourt M, Sandbrook C, Sibanda M, Giuliani A, Gordon-Maclean A. Which components or attributes of biodiversity influence which dimensions of poverty? Env Evid. 2014;3(1):1.
16. DownThemAll. DownThemAll. 2016. http://www.downthemall.net. Accessed 22 Nov 2016.
17. Support M. VLOOKUP function. 2016. https://support.office.com/en-us/article/VLOOKUP-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1. Accessed 22 Nov 2016.
18. Pullin AS, Bangpan M, Dalrymple S, Dickson K, Haddaway NR, Healey JR, et al. Human well-being impacts of terrestrial protected areas. Env Evid. 2013;2(1):1.
19. Haddaway NR, Burden A, Evans CD, Healey JR, Jones DL, Dalrymple SE, et al. Evaluating effects of land management on greenhouse gas fluxes and carbon balances in boreo-temperate lowland peatland systems. Env Evid. 2014;3(1):1.
20. Haddaway NR, Hedlund K, Jackson LE, Kätterer T, Lugato E, Thomsen IK, et al. How does tillage intensity affect soil organic carbon? A systematic review protocol. Env Evid. 2016;5(1):1.
21. Haddaway NR, Hedlund K, Jackson LE, Kätterer T, Lugato E, Thomsen IK, et al. What are the effects of agricultural management on soil organic carbon in boreo-temperate systems? Env Evid. 2015;4(1):1.
22. Mongoose. Mongoose Web Server 2016. https://code.google.com/p/mongoose. Accessed 22 Nov 2016.
23. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. Plos ONE. 2015;10(9):e0138237.
24. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336–41.
25. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007;147(4):224–33.
26. Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, et al. A systematic review identified few methods and strategies describing when and how to update systematic reviews. J Clin Epidemiol. 2007;60(11):1095–104.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central