www.gwern.net /Search

Internet Search Tips

Gwern Branwen

70-89 minutes

> Over time, I developed a certain google-fu and expertise in finding references, papers, and books online. Some of these tricks are not well-known, like checking the Internet Archive (IA) for books. I try to write down my search workflow, and give general advice about finding and hosting documents.

Google-fu search skill is something I've prided myself ever since elementary school, when the librarian challenged the class to find things in the almanac; not infrequently, I'd win. The Internet is the greatest almanac of all, and to the curious, a never-ending cornucopia, so I am sad to see many fail to find things after a cursory search—or not look at all. For most people, if it's not the first hit in Google/Google Scholar, it doesn't exist. Below, I reveal my best Internet search tricks and try to provide a rough flowchart of how to go about an online search, explaining the subtle tricks and intuition of search-fu.

Search

Preparation

Do or do not; there is no try. The first thing you must do is develop a habit of searching when you have a question: ""Google is your friend."" Your only search guaranteed to fail is the one you never run. (Beware trivial inconveniences!)

1. Query syntax knowledge

   Know your basic Boolean operators & the key G search operators: double quotes for exact matches, hyphens for negation/exclusion, and `site:` for search a specific website or specific directory of that website (eg `foo site:gwern.net/docs/genetics/`). You may also want to play with Advanced Search to understand what is possible. (There are many more G search operators (Russell description) but they aren't necessarily worth learning, because they implement esoteric functionality and most seem to be buggy[1].)

2. Hotkey shortcuts (*strongly recommended*)

   Enable some kind of hotkey search with both prompt and copy-paste selection buffer, to turn searching Google (G)/Google Scholar (GS)/Wikipedia (WP) into a reflex.[2] You should be able to search instinctively within a split second of becoming curious, with a few keystrokes. (If you can't use it while IRCing without the other person noting your pauses, it's not fast enough.)

   Example tools: AutoHotkey (Windows), Quicksilver (Mac), xclip+Surfraw/StumpWM's `search-engines`/XMonad's `Actions.Search`/`Prompt.Shell` (Linux). DuckDuckGo offers 'bangs', within-engine special searches (most are equivalent to a kind of Google `site:` search), which can be used similarly or combined with prompts/macros/hotkeys.

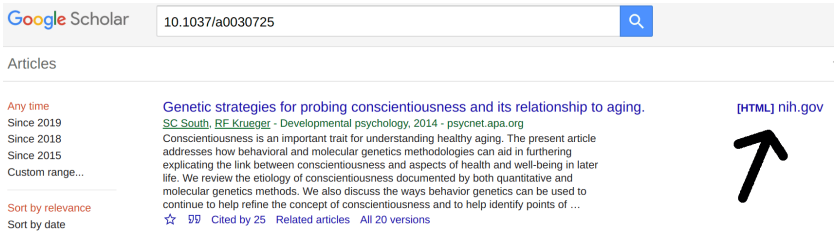
   I make heavy use of the XMonad hotkeys, which I wrote, and which gives me window manager shortcuts: while using any program, I can highlight a title string, and press `Super-shift-y` to open the current selection as a GS search in a new Firefox tab within an instant; if I want to edit the title (perhaps to add an author surname, year, or keyword), I can instead open a prompt, `Super-y`, paste with `C-y`, and edit it before a `\n` launches the search. As can be imagined, this is extremely helpful for searching for many papers or for searching. (There are in-browser equivalents to these shortcuts but I disfavor them because they only work if you are in the browser, typically require more keystrokes or mouse use, and don't usually support hotkeys or searching the copy-paste selection buffer: Firefox, Chrome)

3. Web browser hotkeys

   For navigating between sets of results and entries, you should have good command of your tabbed web browser. You should be able to go to the address bar, move left/right in tabs, close tabs, open new blank tabs, go to a specific tab, etc. (In Firefox, respectively: `C-l`, `C-PgUp`, `C-PgDwn`, `C-w`, `C-t`, `M-[1-9]`.)

Searching

Having launched your search in, presumably, Google Scholar, you must navigate the GS results. For GS, it is often as simple as clicking on the `[PDF]` or `[HTML]` link in the top right which denotes (what GS believes to be) a fulltext link, eg:



An example of a hit in Google Scholar: note the `[HTML]` link indicating there is a fulltext Pubmed version of this paper (often overlooked by newbies).

GS: if no fulltext in upper right, look for soft walls. ☛ **WARNING**: In GS, remember that a fulltext link is *not* always denoted by a "[PDF]" link! Check the top hits by hand: there are often 'soft walls' which block web spiders but still let you download fulltext (perhaps after substantial hassle, like SSRN).

Note that GS supports other useful features like alerts for search queries, alerts for anything citing a specific paper, and reverse citation searches (to followup on a paper to look for failures-to-replicate or criticisms of it).

By Title

Title searches: if a paper fulltext doesn't turn up on the first page, start tweaking (hard rules cannot be given for this, it requires development of "'mechanical sympathy'" and asking a mixture of "how would a machine think to classify this" and "how would other people think to write this"):

- The golden mean: Keep mind when searching, you want some but not too many or too few results. A few hundred hits in GS is around the sweet spot. If you have less than a page of hits, you have made your search too specific.

  If nothing is turning up, try trimming the title. Titles tend to have more errors towards the end than the beginning, and people often drop So start cutting words off the end of the title to broad the search. Think about what kinds of errors you make when you recall titles: you drop punctuation or subtitles, substitute in more familiar synonyms, or otherwise simplify it. (How might OCR software screw up a title?)

  Pay attention to technical terms that pop up in association with your own query terms, particularly in the snippets or full abstracts. Which ones look like they might be more popular than yours, or indicate yours are usually used slightly different from you think they mean? You may need to switch terms.

  If deleting a few terms then yields way too many hits, try to filter out large classes of hits with a negation `foo -bar`, adding as many as necessary; also useful is using OR clauses to open up the search in a more restricted way by adding in possible synonyms, with parentheses for group. This can get quite elaborate—I have on occasion resorted to search queries as baroque as `(foo OR baz) AND (qux OR quux) -bar -garply -waldo -fred` to the point where I hit search query length limits. (By that point, it is time to consider alternate attacks.)

- Tweak the title: quote the title; delete any subtitle; try the subtitle instead; be suspicious of any character which is not alphanumeric and if there are colons, split it into two title quotes (instead of searching `Foo bar: baz quux`, or `"Foo bar: baz quux"`, search `"Foo bar" "baz quux"`); swap their order.

- Tweak the metadata:

  - Add/remove the year.
  - Add/remove the first author's surname. Try searching GS for *just* the author (`author:foo`).

- Delete unusual characters/punctuation:

  Libgen had trouble with colons for a long time, and many websites still do (eg GoodReads); I don't know why colons in particular are such trouble, although hyphens/em-dashes and any kind of quote or apostrophe or period are problematic too.

- Tweak spelling: Try alternate spellings of British/American terms. This shouldn't be necessary, but then, deleting colons or punctuation shouldn't be necessary either.

- Use URLs: if you have a URL, try searching chunks of it, typically towards the end, stripping out dates and domain names.

- Date search:

  Use a search engine (eg G/GS)'s date range to search ±4 years: metadata can be wrong, publishing conventions can be odd, publishers can be *extremely* slow. This is particularly useful if you add a date constraint & simultaneously loosen the search query to turn up the most temporally-relevant of what would otherwise be far too many hits. If this doesn't turn up the relevant target, it might turn up related discussions or fixed citations, since most things are cited most shortly after publication and then vanish into obscurity.

  If a year is not specified, try to guess from the medium: popular media has heavy recentist bias & prefers only contemporary research which is 'news', while academic publications go back a few more years; the style of the reference can give a hint as to how relatively old some mentioned

research or writings is. Frequently, given the author surname and a reasonable guess at some research being a year or two old, the name + date-range + keyword in GS will be enough to find the paper.

- Add jargon: Add technical terminology which *might* be used by relevant papers; for example, if you are looking for an article on college admissions statistics, any such analysis would probably be using logistic regression and, even if they do not say ""logistic regression"" (in favor of some more precise yet unguessable term) would express their effects in terms of ""odds"".

  If you don't know what jargon might be used, you may need to back off and look for a review article or textbook or WP page and spend some quality time reading. If you're using the wrong term, period, nothing will help you; you can spend hours going through countless pages, but that won't make the wrong term work. You may need to read through overviews until you finally recognize the skeleton of what you want under a completely different (and often rather obtuse) name. Nothing is more frustrating that *knowing* there must be a large literature on a topic (""Cowen's Law"") but being unable to *find* it because it's named something completely different from expected—and many fields have different names for the same concept or tool. (Occasionally people compile "Rosetta stones" to translate between fields: eg Baez & Stay 2009, Bertsekas 2018, Metz et al 2018's Table 1. These are invaluable.)

- Even the humble have a tale to tell: Beware hastily dismissing 'bibliographic' websites as useless —they may have more than you think.

  While a bibliographic-focused library site like `elibrary.ru` is (almost) always useless & clutters up search results by hosting only the citation metadata but not fulltext, every so often I run into a peculiar foreign website (often Indian or Chinese) which happens to have a scan of a book or paper. (eg Darlington 1954, which eluded me for well over half an hour until, taking the alternate approach of hunting its volume, I out of desperation clicked on an Indian index/library website which… had it. Go figure.) Sometimes you have to check every hit, just in case.

- Search the Internet Archive:

  The Internet Archive (IA) deserves special mention as a target because it has a remarkable assortment of scans & uploads from all sorts of sources, including the aforementioned Indian/Chinese libraries with more laissez-faire approaches. It also exposes OCR of them all. So not infrequently, a book may be available, or a paper exists in the middle of a scan of an entire journal volume, but the IA will be ranked very low in search queries and the snippet will be misleading due to bad OCR. A good search strategy is to drop the quotes around titles or excerpts and focus down to `site:archive.org` and check the first few hits by hand.

Hard Cases

If the basic tricks aren't giving any hints of working, you will have to get serious. The title may be completely wrong, or it may be indexed under a different author, or not directly indexed at all, or hidden inside a database. Here are some indirect approaches to finding articles:

- Reverse citations: Take a look in GS's ""related articles"" or ""cited by"" to find similar articles such as later versions of a paper which may be useful. (These are also good features to know about if you want to check things like "has this ever been replicated?" or are still figuring out the right jargon to search.)

- Anomalous hits: Look for hints of hidden bibliographic connections. Does a paper pop up high in the search results which doesn't *seem* to make sense, such as not containing your keywords in the displayed snippet? GS generally penalizes items which exist as simply bibliographic entries, so if one is ranked high in a sea of fulltexts, that should make you wonder why it is being prioritized. Similarly, for Google Books (GB): a book might be forbidden from even snippets but rank high; that might be for a good reason. It may actually contain the fulltext hidden inside it, or something else relevant.

- Compilation files: Some papers can be found by searching for the volume or book title to find it indirectly, especially conference proceedings or anthologies; many papers *appear* to not be available online but are merely buried inside a 500-page PDF, and the G snippet listing is misleading.

  Conferences are particularly complex bibliographically, so you may need to apply the same tricks as for page titles: drop parts, don't fixate on the numbers, know that the authors or ISBN or ordering of "title:subtitle" can differ between sources, etc.

- Search by issue: Another approach is to look up the listing for a journal issue, and find the paper by hand; sometimes papers are listed in the journal issue's online Table of Contents, but just don't appear in search engines (⁉). In particularly insidious cases, a paper may be digitized & available—but lumped in with another paper due to error, or only as part of a catch-all file which contains the last 20 miscellaneous pages of an issue. Page range citations are particularly helpful here because they show where the overlap is, so you can download the suspicious overlapping 'papers' to see what they *really* contain.

  Esoteric as this may sound, this has been a problem on multiple occasions. (A particularly epic example was Shockley 1966 where after an hour of hunting, all I had was bibliographic echoes despite apparently being published in a high profile, easily obtained, & definitely digitized journal, *Science*—leaving me thoroughly baffled. I eventually looked up the ToC and inferred it had been hidden in a set of abstracts![3] Or a number of SMPY papers turned out to be split or merged with neighboring items in journal issues, and I had to fix them by hand.)

- Masters/PhD theses: sorry. It may be hopeless if it's pre-2000. You may well find the citation and even an abstract, but actual fulltext…?

  If you have a university proxy, you may be able to get a copy off ProQuest. Otherwise, you need full university ILL services[4], and even that might not be enough (a surprising number of universities appear to restrict access only to the university students/faculty, with the complicating factor of most theses being stored on microfilm).

- Reverse Image Search: If images are involved, a reverse image search in Google Images or TinEye or Yandex Search can turn up important leads.

  Bellingcat has a good guide by Aric Toller: "'Guide To Using Reverse Image Search For Investigations'".

- Enemy action: Is a page or topic not turning up in Google/IA that you *know* ought to be there? Check the website's `robots.txt` & sitemap. While not as relevant as they used to be (due to increasing use of dynamic pages & entities ignoring it), `robots.txt` can sometimes be relevant: key URLs may be excluded from search results, and overly-restrictive `robots.txt` can cause enormous holes in IA coverage, which may be impossible to fix (but at least you'll know why).

- Patience: not every paywall can be bypassed immediately, and papers may be embargoed or proxies not immediately available.

  If something is not available at the moment, it may become available in a few months. Use calendar reminders to check back in to see if an embargoed paper is available or if LG/SH have obtained it, and whether to proceed to additional search steps like manual requests.

- Domain knowledge-specific tips:

  - *US federal courts*: US federal court documents can be downloaded off PACER after registration; it is pay-per-page ($0.10/page) but users under a certain level each quarter (currently $15) have their fees waived, so if you are careful, you may not need to pay anything at all. There is a public mirror, called RECAP, which can be searched & downloaded from for free. If you fail to find a case in RECAP and must use PACER (as often happens for obscure cases), please install the Firefox/Chrome RECAP browser extension, which will copy anything you download into RECAP. (This can be handy if you realize later that you should've kept a long PDF you downloaded or want to double-check a docket.)

    Navigating PACER can be difficult because it is an old & highly specialized computer system which assumes you are a lawyer, or at least very familiar with PACER & the American federal court system. As a rule of thumb, if you are looking up a particular case, what you want to do is to search for the first name & surname (even if you have the case ID) for either criminal or civil cases as relevant, and pull up all cases which might pertain to an individual; there can be multiple cases, cases can hibernate for years, be closed, reopened as a different case number, etc. Once you have found the most active or relevant case, you want to look at the "'docket'", and check the options to see *all* documents in the case. This will pull up a list of many documents as the case unfolds over time; most of these documents are legal bureaucracy, like rescheduling hearings or notifications of changed lawyers. You want the *longest* documents, as those are most likely to be useful. In particular, you want the "indictment", the "criminal complaint"[5], and any transcripts of trial testimony.[6] Shorter documents, like 1–2pg entries in the docket, *can* be useful, but are much less likely to be useful unless you are interested in the exact details of how things like pre-trial negotiations unfold. So carelessly choosing the 'download all' option on PACER may blow through your quarterly budget without getting you anything interesting (and also may interfere with RECAP uploading documents).

    There is no equivalent for state or county court systems, which are balkanized and use a thousand different systems (often privatized & charging far more than PACER); those must be handled on a case by case basis. (Interesting trivia point: according to Nick Bilton's account of the Silk Road 1 case, the FBI and other federal agencies in the SR1 investigation would deliberately steer cases into state rather than federal courts in order to hide them from the relative transparency of the PACER system. The use of multiple court systems can backfire on them, however, as in the case of SR2's DoctorClu (see the DNM arrest census for details), where the local police filings revealed the use of hacking techniques to deanonymize SR2 Tor users, implicating CMU's CERT center—details which were belatedly scrubbed from the PACER filings.)

  - *charity financials*: for USA charity financial filings, do `Form 990 site:charity.com` and then check GuideStar (eg looking at Girl Scouts filings or "'Case Study: Reading Edge's financial filings'"). For UK charities, the Charity Commission for England and Wales may be helpful.

  - *education research*: for anything related to education, do a site search of ERIC, which is similar to IA in that it will often have fulltext which is buried in the usual search results

  - *Wellcome Library*: the Wellcome Library has many old journals or books digitized which are impossible to find elsewhere; unfortunately, their SEO is awful & their PDFs are unnecessarily hidden behind click-through EULAs, so they will not show up normally in Google Scholar or elsewhere. If you see the Wellcome Library in your Google hits, check it out carefully.

    - *magazines* (as opposed to scholarly or trade journals) are hard to get in general. They are not covered in Libgen/Sci-Hub, which outsource that to MagzDB; coverage is poor, however. An alternative is pdf-giant. Particularly for pre-2000 magazines, one may have to resort to looking for old used copies on eBay.

- newspapers: like theses, tricky. I don't know of any general solutions short of a LexisNexis subscription.[7] An interesting resource for American papers is Chronicling America's ""Historic American Newspaper"" scans.

By Quote or Description

For quote/description searches: if you don't have a title and are falling back on searching quotes, try varying your search similarly to titles:

- Novel sentences: Try the easy search first—whatever looks most memorable or unique.
- Short quotes are unique: Don't search too long a quote, a sentence or two is usually enough to be near-unique, and can be helpful in turning up other sources quoting different chunks which may have better citations.
    - *Break up quotes*: Because even phrases can be unique, try multiple sub-quotes from a big quote, especially from the beginning and end, which are likely to overlap with quotes which have prior or subsequent passages.
    - *Odd idiosyncratic wording*: Search for oddly-specific phrases or words, especially numbers. 3 or 4 keywords is usually enough.
    - *Paraphrasing*: Look for passages in the original text which seem like they might be based on the same source, particularly if they are simply dropped in without any hint at sourcing and don't sound like the author; authors typically don't cite every time they draw on a source, usually only the first time, and during editing the 'first' appearance of a source could easily have been moved to later in the text. All of these additional uses are something to add to your searches.
- Robust quotes: You are fighting a game of Chinese whispers, so look for unique-sounding sentences and terms which can survive garbling in the repeated transmissions. Avoid phrases which could be easily reworded in multiple equivalent ways, as people usually will reword them when quoting from memory, screwing up literal searches.
- Tweak spelling: Watch out for punctuation and spelling differences hiding hits.
- Gradient Descent: Longer, less witty versions are usually closer to the original and a sign you are on the right trail. The worse, the better. (Authors all too often fail to write what they were supposed to write—as Yogi Berra remarked, "I really didn't say everything I said.")
- Search books: Switch to GB and hope someone paraphrases or quotes it, and includes a real citation; if you can't see the full passage or the reference section, look up the *book* in Libgen.

Dealing With Paywalls

Use Sci-Hub/Libgen for books/papers. A paywall can usually be bypassed by using Libgen (LG)/Sci-Hub (SH): papers can be searched directly (ideally with the DOI, but title+author with no quotes will usually work), or an easier way may be to prepend[8] `sci-hub.tw` (or whatever SH mirror you prefer) to the URL of a paywall. Some paywalls can be bypassed by looking for accounts/passwords in Google by searching "$NAME password" etc; libraries & schools will often list credentials on a page for their patrons' convenience (a good way for getting access to the OED & the New Yorker, among others).

Use university Internet. If those don't work and you do not have a university proxy or alumni access, many university libraries have IP-based access rules and also open WiFi or Internet-capable computers with public logins inside the library, which can be used, if you are willing to take the time to visit a university in person, for using their databases (probably a good idea to keep a list of needed items before paying a visit).

If that doesn't work, there is a more opaque ecosystem of filesharing services: booksc/bookfi/bookzz, private torrent trackers like Bibliotik, IRC channels with XDCC bots like `#bookz`/`#ebooks`, old P2P networks like eMule, private DC++ hubs…

Site-specific notes:

- Elsevier/`sciencedirect.com`: easy, always available via SH/LG

  Note that many Elsevier journal websites do not work with the SH proxy, although their `sciencedirect.com` version *does* and/or the paper is already in LG. If you see a link to `sciencedirect.com` on a paywall, try it if SH fails on the journal website itself.
- PsycNET: one of the worst sites; SH/LG never work with the URL method, rarely work with paper titles/DOIs, and with my university library proxy, combined searches don't usually work (frequently failing to pull up even bibliographic entries), and only DOI or manual title searches in the EBSCOhost database have a chance of fulltext. (EBSCOhost itself is a fragile search engine which is difficult to query reliably in the absence of a DOI.) Try to find the paper anywhere else besides PsycNET!

Request

Human flesh search engine. Last resort: if none of this works, there are a few places online you can request a copy (however, they will usually fail if you have exhausted all previous avenues):

- /r/scholar
- `#icanhazpdf`
- Wikipedia Resource Request
- LW help desk

Finally, you can always try to contact the author. This only occasionally works for the papers I have the hardest time with, since they tend to be old ones where the author is dead or unreachable—any author publishing a paper since 1990 will usually have been digitized *somewhere*—but it's easy to try.

Post-finding

After finding a fulltext copy, you should find a reliable long-term link/place to store it and make it more findable (remember—if it's not in Google/Google Scholar, it doesn't exist!):

- never link unreliable hosts:
  - *LG/SH*: Always operate under the assumption they could be gone tomorrow. (As my uncle found out with Library.nu shortly after paying for a lifetime membership!) There are no guarantees either one will be around for long under their legal assaults, and no guarantee that they are being properly mirrored or will be restored elsewhere. Download anything you need and keep a copy of it yourself and, ideally, host it publicly.
  - *NBER*: never rely on a `papers.nber.org/tmp/` or `psycnet.apa.org` URL, as they are temporary. (SSRN is also undesirable due to making it increasingly difficult to download, but it is at least reliable.)
  - *Scribd*: never link Scribd—they are a scummy website which impede downloads, and anything on Scribd usually first appeared elsewhere anyway.
  - *RG*: avoid linking to ResearchGate (compromised by new ownership & PDFs get deleted routinely, apparently often by authors) or `Academia.edu` (the URLs are one-time and break)
  - *high-impact journals*: be careful linking to Nature.com or *Cell* (if a paper is not *explicitly* marked as Open Access, even if it's available, it may disappear in a few months!); similarly, watch out for `wiley.com`, `tandfonline.com`, `jstor.org`, `springer.com`, `springerlink.com`, & `mendeley.com`, who pull similar shenanigans.
  - `~/`: be careful linking to academic personal directories on university websites (often noticeable by the Unix convention `.edu/~user/`); they have short half-lives.
- check & improve metadata.

  Adding metadata to papers/books is a good idea because it makes the file findable in G/GS (if it's not online, does it really exist?) and helps you if you decide to use bibliographic software like Zotero in the future. Many academic publishers & LG are terrible about metadata, and will not include even title/author/DOI/year.

  PDFs can be easily annotated with metadata using ExifTool: : `exiftool -All` prints all metadata, and the metadata can be set individually using similar fields.

  For papers hidden inside volumes or other files, you should extract the relevant page range to create a single relevant file. (For extraction of PDF page-ranges, I use `pdftk`, eg: `pdftk 2010-davidson-wellplayed10-videogamesvaluemeaning.pdf cat 180-196 output 2009-fortugno.pdf`.)

  I try to set at least title/author/DOI/year/subject, and stuff any additional topics & bibliographic information into the ""Keywords"" field. Example of setting metadata:

  ```
  exiftool -Author="Frank P. Ramsey" -Date=1930 -Title="On a Problem of
  Formal Logic" -DOI="10.1112/plms/s2-30.1.264" \
      -Subject="mathematics" -Keywords="Ramsey theory, Ramsey's
  theorem, combinatorics, mathematical logic, decidability, \
      first-order logic,  Bernays-Schönfinkel-Ramsey class of first-
  order logic, _Proceedings of the London Mathematical \
      Society_, Volume s2-30, Issue 1, 1 January 1930, pg264-286" 1930-
  ramsey.pdf
  ```

- PDF editing: if a scan, it may be worth editing the PDF to crop the edges, threshold to binarize it (which, for a bad grayscale or color scan, can drastically reduce filesize while increasing readability), and OCRing it. I use gscan2pdf but there are alternatives worth checking out.
- public hosting: if possible, host a public copy; especially if it was very difficult to find, even if it was useless, it should be hosted. The life you save may be your own.
- link on WP/social media: for bonus points, link it in appropriate places on Wikipedia or Reddit or Twitter; this makes people aware of the copy being available, and also supercharges visibility in search engines.

Advanced

Aside from the (highly-recommended) use of hotkeys and Booleans for searches, there are a few useful tools for the researcher, which while expensive initially, can pay off in the long-term:

- `archiver-bot`: automatically archive your web browsing and/or links from arbitrary websites to forestall linkrot; particularly useful for detecting & recovering from dead PDF links

- subscriptions like PubMed & GS search alerts: set up alerts for a specific search query, or for new citations of a specific paper. (Google Alerts is not as useful as it seems.)

    1. *PubMed* has straightforward conversion of search queries into alerts: “"Create alert"” below the search bar. (Given the volume of PubMed indexing, I recommend carefully tailoring your search to be as narrow as possible, or else your alerts may overwhelm you.)
    2. To create generic *GS* search query alert, simply use the “"Create alert"” on the sidebar for any search. To follow citations of a key paper, you must: 1. bring up the paper in GS; 2. click on “"Cited by X"”; 3. *then* use “"Create alert"” on the sidebar.

- GCSE: a Google Custom Search Engines is a specialized search queries limited to whitelisted pages/domains etc (eg my Wikipedia-focused anime/manga CSE).

    A GCSE can be thought of as a saved search query on steroids. If you find yourself regularly including scores of the same domains in multiple searches search, or constantly blacklisting domains with `-site:` or using many negations to filter out common false positives, it may be time to set up a GCSE which does all that by default.

- clippings: note-taking services like Evernote/Microsoft OneNote: regularly making and keeping excerpts creates a personalized search engine, in effect.

    This can be vital for refinding old things you read where the search terms are hopelessly generic or you can't remember an *exact* quote or reference; it is one thing to search a keyword like “autism” in a few score thousand clippings, and another thing to search that in the entire Internet! (One can also reorganize or edit the notes to add in the keywords one is thinking of, to help with refinding.) I make heavy use of Evernote clipping and it is key to refinding my references.

- Crawling websites: sometimes having copies of whole websites might be useful, either for more flexible searching or for ensuring you have anything you might need in the future. (example: “"Darknet Market Archives (2013–2015)"”).

    Useful tools to know about: wget, cURL, HTTrack; Firefox plugins: NoScript, uBlock origin, Live HTTP Headers, Bypass Paywalls, cookie exporting.

    Short of downloading a website, it might also be useful to pre-emptively archive it by using `linkchecker` to crawl it, compile a list of all external & internal links, and store them for processing by another archival program (see Archiving URLs for examples). In certain rare circumstances, security tools like `nmap` can be useful to examine a mysterious server in more detail: what web server and services does it run, what else might be on it (sometimes interesting things like old anonymous FTP servers turn up), has a website moved between IPs or servers, etc.

With proper use of pre-emptive archiving tools like `archiver-bot`, fixing linkrot in one's own pages is much easier, but that leaves other references. Searching for lost web pages is similar to searching for papers:

- just search the title: if the page title is given, search for the title.

    It is a good idea to include page titles in one's own pages, as well as the URL, to help with future searches, since the URL may be meaningless gibberish on its own, and pre-emptive archiving can fail. HTML supports both `alt` and `title` parameters in link tags, and, in cases where displaying a title is not desirable (because the link is being used inline as part of normal hypertextual writing), titles can be included cleanly in Markdown documents like this: `[inline text description](URL "Title")`.

- clean URLs: check the URL for weirdness or trailing garbage like `?rss=1` or `?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+blogspot%2FgJZg+%28Google+AI+Blog%` Or a variant domain, like a `mobile.foo.com/m.foo.com/foo.com/amp/` URL? Those are all less likely to be findable or archived than the canonical version.

- domain search: restrict G search to the original domain with `site:`, or to related domains

- time search: restrict G search to the original date-range/years

- switch engines: try a different search engine: corpuses can vary, and in some cases G tries to be too smart for its own good when you need a literal search; DuckDuckGo and Bing are usable alternatives (especially if one of DuckDuckGo's 'bang' special searches is what one needs)

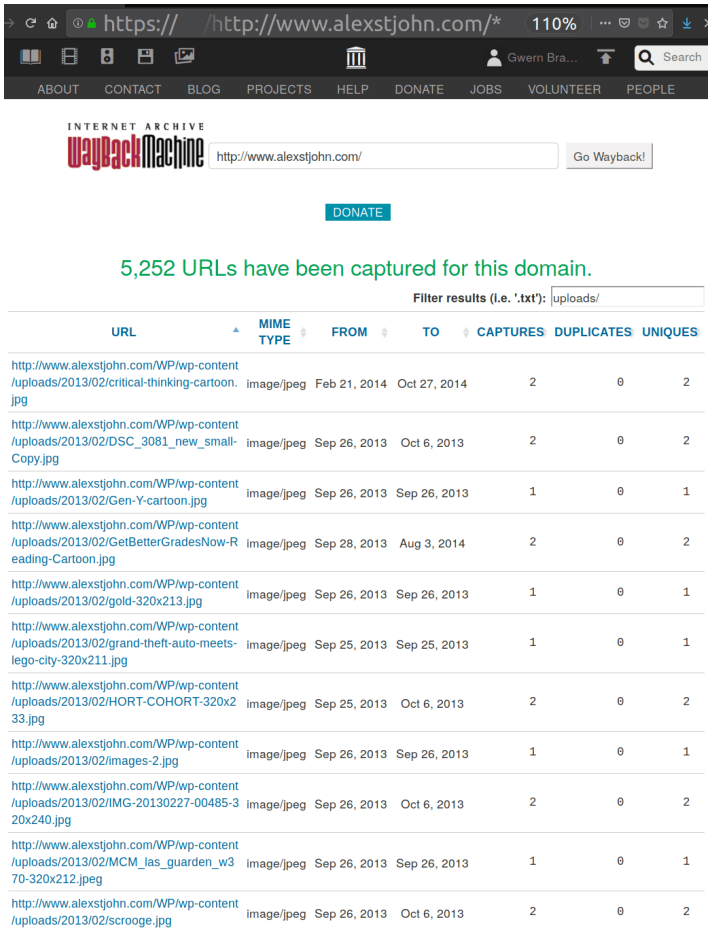
- check archives: if nowhere on the clearnet, try the Internet Archive (IA) or the Memento meta-archive search engine:

    IA is the default backup for a dead URL. If IA doesn't Just Work, there may be other versions in it:

    - *misleading redirects*: did the IA 'helpfully' redirect you to a much-later-in-time error page? Kill the redirect and check the earliest stored version for the exact URL rather than the redirect. Did the page initially load but then error out/redirect? Disable JS with NoScript and reload.

    - within-domain archives: IA lets you list all URLs with any archived versions, by searching for `URL/*`; the list of available URLs may reveal an alternate newer/older URL. It can also be useful to filter by filetype or substring.

For example, one might list all URLs in a domain, and if the list is too long and filled with garbage URLs, then using the ‘"'Filter results"'" incremental-search widget to search for ‘"'uploads/"'" on a WordPress blog.[9]



URL	MIME TYPE	FROM	TO	CAPTURES	DUPLICATES	UNIQUES
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/critical-thinking-cartoon.jpg	image/jpeg	Feb 21, 2014	Oct 27, 2014	2	0	2
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/DSC_3081_new_small-Copy.jpg	image/jpeg	Sep 26, 2013	Oct 6, 2013	2	0	2
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/Gen-Y-cartoon.jpg	image/jpeg	Sep 26, 2013	Sep 26, 2013	1	0	1
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/GetBetterGradesNow-Reading-Cartoon.jpg	image/jpeg	Sep 28, 2013	Aug 3, 2014	2	0	2
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/gold-320x213.jpg	image/jpeg	Sep 26, 2013	Sep 26, 2013	1	0	1
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/grand-theft-auto-meets-lego-city-320x211.jpg	image/jpeg	Sep 25, 2013	Sep 25, 2013	1	0	1
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/HORT-COHORT-320x233.jpg	image/jpeg	Sep 25, 2013	Oct 6, 2013	2	0	2
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/images-2.jpg	image/jpeg	Sep 26, 2013	Sep 26, 2013	1	0	1
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/IMG-20130227-00485-320x240.jpg	image/jpeg	Sep 26, 2013	Oct 6, 2013	2	0	2
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/MCM_las_guarden_w370-320x212.jpeg	image/jpeg	Sep 26, 2013	Sep 26, 2013	1	0	1
http://www.alexstjohn.com/WP/wp-content/uploads/2013/02/scrooge.jpg	image/jpeg	Sep 26, 2013	Oct 6, 2013	2	0	2

Screenshot of an oft-overlooked feature of the Internet Archive: displaying all available/archived URLs for a specific domain, filtered down to a subset matching a string like `*uploads/*`.

- `wayback_machine_downloader` (not to be confused with the `internetarchive` Python package which provides a CLI interface to uploading files) is a Ruby tool which lets you download whole domains from IA, which can be useful for running a local fulltext search using regexps (a good `grep` query is often enough), in cases where just looking at the URLs via `URL/*` is not helpful. (An alternative which might work is `websitedownloader.io`.)

  Example:

  ```
  gem install --user-install wayback_machine_downloader
  ~/.gem/ruby/2.5.0/bin/wayback_machine_downloader
  wayback_machine_downloader --all-timestamps
  'https://blog.okcupid.com'
  ```

- did *the domain change*, eg from `www.foo.com` to `foo.com` or `www.foo.org`? Entirely different as far as IA is concerned.
- is this a *Blogspot blog*? Blogspot is uniquely horrible in that it has versions of each blog for every country domain: a `foo.blogspot.com` blog could be under any of `foo.blogspot.de`, `foo.blogspot.au`, `foo.blogspot.hk`, `foo.blogspot.jp`…[10]
- did the website provide *RSS feeds*?

  A little known fact is that Google Reader (GR; October 2005–July 2013) stored all RSS items it crawled, so if a website's RSS feed was configured to include full items, the RSS feed history was an alternate mirror of the whole website, and since GR never removed RSS items, it was possible to retrieve pages or whole websites from it. GR has since closed down, sadly, but before it closed, Archive Team downloaded a large fraction of GR's historical RSS feeds, and *those* archives are now hosted on IA. The catch is that they are stored in mega-WARCs, which, for all their archival virtues, are not the most user-friendly format. The raw GR mega-WARCs are difficult enough to work with that I defer an example to the appendix.

- `archive.today`: an IA-like mirror

- any *local archives*, such as those made with my `archiver-bot`
- *Google Cache* (GC): GC works, sometimes, but the copies are usually the worst around, ephemeral & cannot be relied upon. Google also appears to have been steadily deprecating GC over the years, as GC shows up less & less in search results. A last resort.

Digital

E-books are rarer and harder to get than papers, although the situation has improved vastly since the early 2000s. To search for books online:

- More straightforward: book searches tend to be faster and simpler than paper searches, and to require less cleverness in search query formulation, perhaps because they are rarer online, much larger, and have simpler titles, making it easier for search engines.

  Search G for title.

  ☛ **WARNING**: book fulltexts usually don't show up in G*S* (for unknown reasons). You need to check both when searching for books.

  To double-check, you can try a `filetype:pdf` search; then check LG. Typically, if the main title + author doesn't turn it up, it's not online. (In some cases, the author order is reversed, or the title:subtitle are reversed, and you can find a copy by tweaking your search, but these are rare.).

- IA: the Internet Archive has many books scanned which do not appear easily in search results (poor SEO?).

  - If an IA hit pops up in a search, *always check it*; the OCR may offer hints as to where to find it. If you don't find anything or the provided, try doing an IA site search in G (*not* the IA built-in search engine), eg `book title site:archive.org`.
  - *DRM workarounds*: if it *is* on IA but the IA version is DRMed and is only available for "'checkout'", you can jailbreak it.

    Download the PDF version to Adobe Digital Elements ≤4.0, which can be run in Wine, and then import it to Calibre with the De-DRM plugin, which will produce a DRM-free PDF inside Calibre's library. (Getting De-DRM running can be tricky, especially under Linux. I wound up having to edit some of the paths in the Python files to make them work with Wine.) You can then add metadata to the PDF & upload it to LG[11]. (LG's versions of books are usually better than the IA scans, but if they don't exist, IA's is better than nothing.)

- Google Play: use the same PDF DRM as IA, can be broken same way
- HathiTrust also hosts many book scans, which can be searched for clues or hints or jailbroken.

  HathiTrust blocks whole-book downloads but it's easy to download each page in a loop and stitch them together, for example:

  ```
  for i in {0 .. 151}
  do if [[ ! -s "$i.pdf" ]]; then
      wget "https://babel.hathitrust.org/cgi/imgsrv/download/pdf?
  id=mdp.39015050609067;orient=0;size=100;seq=$i;attachment=0" \
           -O "$i.pdf"
      sleep 10s
   fi
  done

  pdftk *.pdf cat output 1957-super-
  scientificcareersandvocationaldevelopmenttheory.pdf

  exiftool -Title="Scientific Careers and Vocational Development
  Theory: A review, a critique and some recommendations" \
      -Date=1957 -Author="Donald E. Super, Paul B. Bachrach" -
  Subject="psychology" \
      -Keywords="Bureau Of Publications (Teachers College Columbia
  University), LCCCN: 57-12336, National Science Foundation, public
  domain, \
      https://babel.hathitrust.org/cgi/pt?
  id=mdp.39015050609067;view=1up;seq=1
  http://psycnet.apa.org/record/1959-04098-000" \
      1957-super-scientificcareersandvocationaldevelopmenttheory.pdf
  ```

  Another example of this would be the Wellcome Library; while looking for *An investigation into the relation between intelligence and inheritance*, Lawrence 1931, I came up dry until I checked one of the last search results, a "'Wellcome Digital Library'" hit, on the slim off-chance that, like the occasional Chinese/Indian library website, it just might have fulltext. As it happens, it did—good news? Yes, but with a caveat: it provides *no* way to download the book! It provides OCR, metadata, and individual page-image downloads all under CC-BY-NC-SA (so no legal problems), but… not the book. (The OCR is also unnecessarily zipped, so that is why Google ranked the page so low and did not show any revealing excerpts from the OCR transcript: because it's

hidden in an opaque archive to save a few kilobytes while destroying SEO.) Examining the download URLs for the highest-resolution images, they follow an unfortunate schema:

1. `https://dlcs.io/iiif-img/wellcome/1/5c27d7de-6d55-473c-b3b2-6c74ac7a04c6/full/2212,/0/default.jpg`
2. `https://dlcs.io/iiif-img/wellcome/1/d514271c-b290-4ae8-bed7-fd30fb14d59e/full/2212,/0/default.jpg`
3. etc

Instead of being sequentially numbered 1–90 or whatever, they all live under a unique hash or ID. Fortunately, one of the metadata files, the 'manifest' file, provides all of the hashes/IDs (but not the high-quality download URLs). Extracting the IDs from the manifest can be done with some quick `sed` & `tr` string processing, and fed into another short `wget` loop for download

```
fgrep '@id' manifest\?
manifest\=https\:%2F%2Fwellcomelibrary.org%2Fiiif%2Fb18032217%2Fmanifest
 | \
   sed -e 's/.*imageanno\/\(.*\)/\1/' | egrep -v '^ .*' | tr -d ',' |
tr -d '"' # "
# bf23642e-e89b-43a0-8736-f5c6c77c03c3
# 334faf27-3ee1-4a63-92d9-b40d55ab72ad
# 5c27d7de-6d55-473c-b3b2-6c74ac7a04c6
# d514271c-b290-4ae8-bed7-fd30fb14d59e
# f85ef645-ec96-4d5a-be4e-0a781f87b5e2
# a2e1af25-5576-4101-abee-96bd7c237a4d
# 6580e767-0d03-40a1-ab8b-e6a37abe849c
# ca178578-81c9-4829-b912-97c957b668a3
# 2bd8959d-5540-4f36-82d9-49658f67cff6
# ...etc
I=1
for HASH in $HASHES; do
    wget "https://dlcs.io/iiif-
img/wellcome/1/$HASH/full/2212,/0/default.jpg" -O $I.jpg
    I=$((I+1))
done
```

And then the 59MB of JPGs can be cleaned up as usual with `gscan2pdf` (empty pages deleted, tables rotated, cover page cropped, all other pages binarized), compressed/OCRed with `ocrmypdf`, and metadata set with `exiftool`, producing a readable, downloadable, highly-search-engine-friendly 1.8MB PDF.

- ebook.farm is a Kindle pirate website which takes Amazon gift-cards as currency; it has many recent e-books which are DRM-free and can be uploaded to LG.
- remember the analog hole works for papers/books too:

  if you can find a copy to *read*, but cannot figure out how to *download* it directly because the site uses JS or complicated cookie authentication or other tricks, you can always exploit the 'analogue hole'—fullscreen the book in high resolution & take screenshots of every page; then crop, OCR etc. This is tedious but it works. And if you take screenshots at sufficiently high resolution, there will be relatively little quality loss. (This works better for books that are scans than ones born-digital.)

Physical

Expensive but feasible. Books are something of a double-edged sword compared to papers/theses. On the one hand, books are much more often unavailable online, and must be bought offline, but at least you almost always *can* buy used books offline without much trouble (and often for <$10 total); on the other hand, while paper/theses are often available online, when one is not unavailable, it's usually *very* unavailable, and you're stuck (unless you have a university ILL department backing you up or are willing to travel to the few or only universities with paper or microfilm copies).

Purchasing from used book sellers:

- Sellers:
  - used book search engines: Google Books/find-more-books.com: a good starting point for seller links; if buying from a marketplace like AbeBooks/Amazon/Barnes & Noble, it's worth searching the seller to see if they have their own website, which is potentially much cheaper. They may also have multiple editions in stock.
  - bad: eBay & Amazon are often bad, due to high-minimum-order+S&H and sellers on Amazon seem to assume Amazon buyers are easily rooked; but can be useful in providing metadata like page count or ISBN or variations on the title
  - good: AbeBooks, Thrift Books, Better World Books, B&N, Discover Books.

    Note: on AbeBooks, international orders can be useful (especially for behavioral genetics or psychology books) but be careful of international orders with your credit card—many debit/credit cards will fail on international orders and trigger a fraud alert, and PayPal is not accepted.

- subscriptions: if a book is not available or too expensive, set price watches: AbeBooks supports email alerts on stored searches, and Amazon can be monitored via CamelCamelCamel (remember the CCC price alert you want is on the *used third-party* category, as new books are more expensive, less available, and unnecessary).

Scanning:

- destructive vs non-destructive: the fundamental dilemma of book scanning—destructively debinding books with a razor or guillotine cutter works much better & is much less time-consuming than spreading them on a flatbed scanner to scan one-by-one[12], because it allows use of a sheet-fed scanner instead, which is easily 5x faster and will give higher-quality scans (because the sheets will be flat, scanned edge-to-edge, and much more closely aligned), but does, of course, require effectively destroying the book.

- Tools:

    - *cutting*: For simple debinding of a few books a year, an X-acto knife/razor is good (avoid the 'triangle' blades, get curved blades intended for large cuts instead of detail work).

      Once you start doing more than one a month, it's time to upgrade to a guillotine blade paper cutter (a fancier swinging-arm paper cutter, which uses a two-joint system to clamp down and cut uniformly).

      A guillotine blade can cut chunks of 200 pages easily without much slippage, so for books with more pages, I use both: an X-acto to cut along the spine and turn it into several 200-page chunks for the guillotine cutter.

    - *scanning*: at some point, it may make sense to switch to a scanning service like 1DollarScan (1DS has acceptable quality for the black-white scans I have used them for thus far, but watch out for their nickel-and-diming fees for OCR or ""setting the PDF title""; these can be done in no time yourself using `gscan2pdf`/`exiftool`/`ocrmypdf` and will save a *lot* of money as they, amazingly, bill by 100-page units). Books can be sent directly to 1DS, reducing logistical hassles.

- clean up: after scanning, crop/threshold/OCR/add metadata

    - *Adding metadata*: same principles as papers. While more elaborate metadata can be added, like bookmarks, I have not experimented with those yet.

- File format: PDF.

  In the past, I used DjVu for documents I produce myself, as it produces much smaller scans than gscan2pdf's default PDF settings due to a buggy Perl library (at least half the size, sometimes one-tenth the size), making them more easily hosted & a superior browsing experience.

  The downsides of DjVu are that not all PDF viewers can handle DjVu files, and it appears that G/GS ignore all DjVu files (despite the format being 20 years old), rendering them completely unfindable online. In addition, DjVu is an increasingly obscure format and has, for example, been dropped by the IA as of 2016. The former is a relatively small issue, but the latter is fatal—being consigned to oblivion by search engines largely defeats the point of scanning! (""If it's not in Google, it doesn't exist."") Hence, despite being a worse format, I now recommend PDF and have stopped using DjVu for new scans[13] and have converted my old DjVu files to PDF.

- Uploading: to LibGen, usually. For backups, filelockers like Dropbox, Mega, MediaFire, or Google Drive are good. I usually upload 3 copies including LG. I rotate accounts once a year, to avoid putting too many files into a single account.

- Hosting: hosting papers is easy but books come with risk:

  Books can be dangerous; in deciding whether to host a book, my rule of thumb is host only books pre-2000 and which do not have Kindle editions or other signs of active exploitation and is effectively an 'orphan work'.

  As of 23 October 2019, hosting 4090 files over 9 years (very roughly, assuming linear growth, <6.7 million document-days of hosting:), I've received 4 takedown orders: a behavioral genetics textbook (2013), *The Handbook of Psychopathy* (2005), a recent meta-analysis paper (Roberts et al 2016), and a CUP DMCA takedown order for 27 files. I broke my rule of thumb to host the 2 books (my mistake), which leaves only the 1 paper, which I think was a fluke. So, as long as one avoids relatively recent books, the risk should be minimal.

Below are >13 case studies of difficult-to-find resources or citations, and how I went about locating them, demonstrating the various Internet search techniques described above and how to think about searches.

- Missing Appendix: Anders Sandberg asked:

  > Does anybody know where the online appendix to Nordhaus' ""Two Centuries of Productivity Growth in Computing"" is hiding?

  I look up the title in Google Scholar; seeing a friendly `psu.edu` PDF link (CiteSeerx), I click. The paper says ""The data used in this study are provided in a background spreadsheet available at `http://www.econ.yale.edu/~nordhaus/Computers/Appendix.xls`"". Sadly, this is a lie. (Sandberg would of course have tried that.)

I immediately check the URL in the IA—nothing. The IA didn't catch it at all. Maybe the official published paper website has it? Nope, it references the same URL, and doesn't provide a copy as an appendix or supplement. (What do we pay these publishers such enormous sums of money for, exactly?) So I back off to checking `http://www.econ.yale.edu/~nordhaus/`, to check Nordhaus's personal website for a newer link. The Yale personal website is empty and appears to've been replaced by a Google Sites personal page. It links nothing useful, so I check a more thorough index, Google, by searching `site:sites.google.com/site/williamdnordhaus/`. Nothing there either (and it appears almost empty, so Nordhaus has allowed most of his stuff to be deleted and bitrot). I try a broader Google: `nordhaus appendix.xls`. This turns up some spreadsheets, but still nothing.

Easier approaches having been exhausted, I return to the IA and I pull up *all* URLs archived for his original personal website: `https://web.archive.org/web/*/http://www.econ.yale.edu/~nordhaus/*` This pulls up way too many URLs to manually review, so I filter results for `xls`, which reduces to a more manageable 60 hits; reading through the hits, I spot `http://www.econ.yale.edu:80/~nordhaus/homepage/documents/Appendix_Nordhaus_computation_update` from 10 Oct 2014; this sounds right, albeit substantially later in time than expected (either 2010 or 2012, judging from the filename).

Downloading it, opening it up and cross-referencing with the paper, it has the same spreadsheet 'sheets' as mentioned, like "'Manual'" or "'Capital_Deep'", and seems to be either the original file in question or an updated version thereof (which may be even better). The spreadsheet metadata indicates it was created "'04/09/2001, 23:20:43, ITS Academic Media & Technology'", and modified "'12/22/2010, 02:40:20'", so it seems to be the latter—it's the original spreadsheet Nordhaus created when he began work several years prior to the formal 2007 publication (6 years seems reasonable given all the delays in such a process), and then was updated 3 years afterwards. Close enough.

- Misremembered Book: A Redditor asked:

  > I was in a consignment type store once and picked up a book called "'Eat fat, get thin'". Giving it a quick scan through, it was basically the same stuff as Atkins but this book was from the 50s or 60s. I wish I'd have bought it. I think I found a reference to it once online but it's been drowned out since someone else released a book with the same name (and it wasn't Barry Groves either).

  The easiest way to find a book given a corrupted title, a date range, and the information there are many similar titles drowning out a naive search engine query, is to skip to a specialized search engine with clean metadata (ie. a library database).

  Searching in WorldCat for 1950s–1970s, "'Eat fat, get thin'" turns up nothing relevant. This is unsurprising, as he was unlikely to've remembered the title *exactly*, and this title doesn't quite sound right for the era anyway (a little too punchy and ungrammatical, and 'thin' wasn't a desirable word back then compared to words like 'slim' or 'sleek' or 'svelte'). People often oversimplify titles, so I dropped back to just "'Eat fat'".

  This immediately turned up the book: Richard Mackarness's 1958 *Eat Fat and Grow Slim*—note that it is *almost* the same title, with a comma serving as conjunction and 'slim' rather than the more contemporary 'thin', but just different enough to screw up an overly-literal search.

  With the same trick in mind, we could also have found it in a regular Google search query by adding additional terms to hint to Google that we want old books, not recent ones: both `"Eat Fat" 1950s` or `"Eat Fat" 1960s` would have turned it up in the first 5 search results. If we didn't use quotes, the searches get harder because broader hits get pulled in. For example, `Eat fat, get thin 1950s -Hyman` excludes the recent book mentioned, but you still have to go down 15 hits before finding Mackarness, and `Eat fat, get thin -Hyman` requires going down 18 hits.

- Missing Website: Bučar et al 2015, on the phenomenon of disappearing polymorphs quotes striking transcripts from a major example of a disappearing crystal, when ~1998 Abbott suddenly became unable to manufacture the anti-retroviral drug ritonavir (Norvir™) due to a rival (and less effective) crystal form spontaneously infecting all its plants, threatening many AIDS patients, but notes:

  > The transcripts were originally published on the website[42] of the International Association of Physicians in AIDS Care [IAPAC], but no longer appear there.

  A search using the quotes confirms that the originals have long since vanished from the open Internet, turning up only quotes of the quotations. Unfortunately, no URL is given. The Internet Archive has comprehensive mirrors of the IAPAC, but too many to easily search through. Using the filter feature, I keyword-searched for "'ritonavir'", but while this turned up a number of pages from roughly the right time period, they do not mention it and none of the quotes appear. The key turned out to be to use the trademark name instead which pulls up many more pages, and after checking a few, the IAPAC turned out to have organized all the Norvir material into a single subdirectory with a convenient `index.html`; the articles/transcripts, in turn, were indexed under the linked "'Description of the Problem'" index page.

  I then pulled the Norvir subdirectory with a `~/.gem/ruby/2.5.0/bin/wayback_machine_downloader wayback_machine_downloader 'http://www.iapac.org/norvir/'` command and hosted a mirror to make it visible in Google.

- Speech → Book: Nancy Lebovitz asked about a citation in a Roy Baumeister speech about sex differences:

> There's an idea I've seen a number of times that 80% of women have had descendants, but only 40% of men. A little research tracked it back to this, but the speech doesn't have a cite and I haven't found a source.

This could be solved by guessing that the formal citation is given in the book, and doing keyword search to find a similar passage. The second line of the speech says:

> For more information on this topic, read Dr. Baumeister's book *Is There Anything Good About Men?* available in bookstores everywhere, including here.

A search of *Is There Anything Good About Men* in Libgen turns up a copy. Download. What are we looking for? A reminder, the key lines in the speech are:

> …It's not a trick question, and it's not 50%. True, about half the people who ever lived were women, but that's not the question. We're asking about all the people who ever lived who have a descendant living today. Or, put another way, yes, every baby has both a mother and a father, but some of those parents had multiple children. Recent research using DNA analysis answered this question about two years ago. Today's human population is descended from twice as many women as men. I think this difference is the single most under-appreciated fact about gender. To get that kind of difference, you had to have something like, throughout the entire history of the human race, maybe 80% of women but only 40% of men reproduced.

We could search for various words or phrase from this passage which seem to be relatively unique; as it happens, I chose the rhetorical "'50%'" (but "'80%'", "'40%'", "'underappreciated'", etc all would've worked with varying levels of efficiency since the speech is heavily based on the book), and thus jumped straight to chapter 4, "'The Most Underappreciated Fact About Men'". (If these had not worked, we could have started searching for years, based on the quote "'about two years ago'".) A glance tells us that Baumeister is discussing exactly this topic of reproductive differentials, so we read on and a few pages later, on page 63, we hit the jackpot:

> The correct answer has recently begun to emerge from DNA studies, notably those by Jason Wilder and his colleagues. They concluded that among the ancestors of today's human population, women outnumbered men about two to one. Two to one! In percentage terms, then, humanity's ancestors were about 67% female and 33% male.

Who's Wilder? A C-f for "'Wilder'" takes us to pg286, where we immediately read:

> …The DNA studies on how today's human population is descended from twice as many women as men have been the most requested sources from my earlier talks on this. The work is by Jason Wilder and his colleagues. I list here some sources in the mass media, which may be more accessible to laypersons than the highly technical journal articles, but for the specialists I list those also. For a highly readable introduction, you can Google the article "'Ancient Man Spread the Love Around,'" which was published September, 20, 2004 and is still available (last I checked) online. There were plenty of other stories in the media at about this time, when the research findings first came out. In "'Medical News Today,'", on the same date in 2004, a story under "'Genes expose secrets of sex on the side'" covered much the same material.
>
> If you want the original sources, read Wilder, J. A., Mobasher, Z., & Hammer, M. F. (2004). "'Genetic evidence for unequal effective population sizes of human females and males'". *Molecular Biology and Evolution*, 21, 2047–2057. If that went down well, you might try Wilder, J. A., Kingan, S. B., Mobasher, Z., Pilkington, M. M., & Hammer, M. F. (2004). "'Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males'". *Nature Genetics*, 36, 1122–1125. That one was over my head, I admit. A more readable source on these is Shriver, M. D. (2005), "'Female migration rate might not be greater than male rate'". *European Journal of Human Genetics*, 13, 131–132. Shriver raises another intriguing hypothesis that could have contributed to the greater preponderance of females in our ancestors: Because couples mate such that the man is older, the generational intervals are smaller for females (i.e., baby's age is closer to mother's than to father's). As for the 90% to 20% differential in other species, that I believe is standard information in biology, which I first heard in one of the lectures on testosterone by the late James Dabbs, whose book *Heroes, Rogues, and Lovers* remains an authoritative source on the topic.

Wilder et al 2004, incidentally, fits well with Baumeister remarking in 2007 that the research was done 2 or so years ago. And of course you could've done the same thing using Google Books: search "'Baumeister anything good about men'" to get to the book, then search-within-the-book for "'50%'", jump to page 53, read to page 63, do a second search-within-the-book for "'Wilder'" and the second hit of page 287 even luckily gives you the snippet:

> *Sources and References* 287

> …If you want the original sources, read Wilder, J. A., Mobasher, Z., & Hammer, M. F. (2004). ""Genetic evidence for unequal effective population sizes of human females and males"". *Molecular Biology and Evolution*…

- a commenter who shall remain nameless wrote

  > I challenge you to find an example of someone saying ""this den of X"" where X does not have a negative connotation.

  I found a positive connotation within 5s using my Google hotkey for `"this den of "`, and, curious about further ones, found additional uses of the phrase in regard to dealing with rattlesnakes in Google Books.

- Rowling Quote On Death: Did J.K. Rowling say the *Harry Potter* books were about 'death'? There are a lot of Rowling statements, but checking WP and opening up each interview links (under the theory that the key interviews are linked there) and searching for 'death' soon turns up a relevant quote from 2001:

  > Death is an extremely important theme throughout all seven books. I would say possibly the most important theme. If you are writing about Evil, which I am, and if you are writing about someone who is essentially a psychopath, you have a duty to show the real evil of taking human life.

- Crowley Quote: Scott Alexander posted a piece linking to an except titled ""Crowley on Religious Experience"".

  The link was broken, but Alexander brought it up in the context of an earlier discussion where he also quoted Crowley; searching *those* quotes reveals that it must have been excerpts from *Magick: Book 4*

- Finding The Right 'SAGE': Phil Goetz noted that an anti-aging conference named ""SAGE"" had become impossible to find in Google due to a *LGBT* aging conference also named SAGE.

  Regular searches would fail, but a combination of tricks worked: `SAGE anti-aging conference` combined with restricting Google search to 2003–2005 time-range turned up a citation to its website as the fourth hit, `http://www.sagecrossroads.net` (which has ironically since died).

- UK Charity Financials: The Future of Humanity Institute (FHI) doesn't clearly provide charity financial forms akin to the US Form 990s, making it hard to find out information about its budget or results.

  FHI doesn't show up in the CC, NPC, or GuideStar, which are the first places to check for charity finances, so I went a little broader afield and tried a site search on the FHI website: `budget site:fhi.ox.ac.uk`. This immediately turned up FHI's own documentation of its activities and budgets, such as the 2007 annual report; I used part of its title as a new Google search: `future of humanity institute achievements report site:fhi.ox.ac.uk`.

- Nobel Lineage Research: John Maxwell referred to a forgotten study on high correlation between Nobelist professors & Nobelist grad students (almost entirely a selection effect, I would bet). I was able to refind it in 7 minutes.

  I wasted a few searches like `factor predicting Nobel prize` or `Nobel prize graduate student` in Google Scholar, until I search for `Nobel laureate "graduate student"`; the second hit was a citation, which is a little unusual for Google Scholar and meant it was important, and it had the critical word *mutual* in it—simultaneous partners in Nobel work is somewhat rare, but temporally separated teams don't work for prizes, and I suspected that it was exactly what I was looking for. Googling the title, I soon found a PDF like ""Eminent Scientists' Demotivation in School: A symptom of an incurable disease?"", Viau 2004 which confirmed it (and Viau 2004 is interesting in its own right as a contribution to the Conscientious vs IQ question). I then followed it to a useful paragraph:

  > In a study conducted with 92 American winners of the Nobel Prize, Zuckerman (1977) discovered that 48 of them had worked as graduate students or assistants with professors who were themselves Nobel Prize award-winners. As pointed out by Zuckerman (1977), the fact that 11 Nobel prizewinners have had the great physicist Rutherford as a mentor is an example of just how significant a good mentor can be during one's studies and training. It then appears that most eminent scientists did have people to stimulate them during their childhood and mentor(s) during their studies. But, what exactly is the nature of these people's contribution.
  >
  > - Zuckerman, H. (1977). *Scientific Elite: Nobel Laureates in the United States*. New York: Free Press.

  GS lists >900 citations of this book, so there may well be additional or followup studies covering the 40 years since. Or, also relevant is ""Zuckerman, H. (1983). The scientific elite: Nobel laureates' mutual influences. In R. S. Albert (Ed.), *Genius and eminence* (pp. 241–252). New York: Pergamon Press"", and "Zuckerman H. 'Sociology of Nobel Prizes', *Scientific American* 217 (5): 25& 1967."

- Too Narrow: A failure case study: The_Duck looked for but failed to find other uses of a famous Wittgenstein anecdote. His mistake was being *too specific*:

> Yes, clearly my Google-fu is lacking. I think I searched for phrases like ""sun went around the Earth,"" which fails because your quote has ""sun went round the Earth.""

As discussed in the search tips, when you're formulating a search, you want to balance how many hits you get, aiming for a sweet spot of a few hundred high-quality hits to review—the broader your formulation, the more likely the hits will include your target (if it exists) but the more hits you'll return. In The_Duck's case, he used an overly-specific search, which would turn up only 2 hits at most; this should have been a hint to loosen the search, such as by dropping quotes or dropping keywords.

In this case, my reasoning would go something like this, laid out explicitly: '"Wittgenstein" is almost guaranteed to be on the same page as any instance of this quote, since the quote is about Wittgenstein; LW, however, doesn't discuss Wittgenstein much, so there won't be many hits in the first place; to find this quote, I only need to narrow down those hits a *little*, and after "Wittgenstein", the most fundamental core word to this quote is "Earth" or "sun", so I'll toss one of them in and… ah, there's the quote!'

If I were searching the general Internet, my reasoning would go more like "'Wittgenstein' will be on, like, a *million* websites; I need to narrow that down a *lot* to hope to find it; so maybe 'Wittgenstein' *and* 'Earth' *and* 'Sun'… nope nothing on the first page, toss in `'goes around' OR 'go around'`—ah there it is!"

(Actually, for the general Internet, just `Wittgenstein earth sun` turns up a first page mostly about this anecdote, several of which include all the details one could need.)

- Dead URL: A link to a research article in a post by Morendil broke, he had not provided any formal citation data, *and* the original domain blocks all crawlers in its `robots.txt` so IA would not work. What to do?

  The simplest solution was to search a direct quote, turning up a Scribd mirror; Scribd is a parasite website, where people upload copies from elsewhere, which ought to make one wonder where the *original* came from. (It often shows up before the original in any search engine, because it automatically runs OCR on submissions, making them more visible to search engines.) With a copy of the journal issue to work with, you can easily find the official HP archives and download the original PDF.

  If that hadn't worked, searching for the URL without `/pg_2/` in it yields the full citation, and then that can be looked up normally. Finally, somewhat more dangerous would be trying to find the article just by author surname & year.

- Description But No Citation: A 2013 Medical Daily on the effects of reading fiction omitted any link or citation to the research in question. But it is easy to find.

  The article says the authors are one Kaufman & Libby, and implies it was published in the last year. So: go to Google Scholar, punch in `Kaufman Libby`, limit to 'Since 2012'; and the correct paper (""Changing beliefs and behavior through experience-taking"") is the first hit with fulltext available on the right-hand side as the text link ""[PDF] from `tiltfactor.org`"" & many other domains.

- Finding Followups: Is soy milk bad for you as one study suggests? Has anyone replicated it? This is easy to look into a little if you use the power of reverse citation search!

  Plug `Brain aging and midlife tofu consumption` into Google Scholar, one of the little links under the first hit points to ""Cited by 176""; if you click on that, you can hit a checkbox for ""Search within citing articles""; then you can search a query like `experiment OR randomized OR blind` which yields 121 results. The first result shows no negative effect and a trend to a benefit, the second is inaccessible, the second & third are reviews whose abstract suggests it would argue for benefits, and the fourth discusses sleep & mood benefits to soy diets. At least from a quick skim, this claim is not replicating, and I am dubious about it.

- My outstanding research paper/book bounties
- ""The Neural Net Tank Urban Legend""
- ""Leprechaun hunting and historical context""
- Arthur Moulton

- ""Million Short"" (search engine overlay which removes top 100/1k/10k/100k/1m domains from hits, exposing obscurer sites which may be highly novel)
- Practice G search problems: ""A Google A Day""; Google Power Searching course (OK for beginners but you may want to skip the videos in favor of the slides)
- ""Scholarship: How to Do It Efficiently""
- ""How to do hard things""
- outline.com
- ""Tech Support Cheat Sheet""
- ""Do repositories of translated papers exist?""
- /r/DataHoarder//r/Piracy
- Archive Team's Archive Bot
- ""Building personal search infrastructure for your knowledge and code: Overview of search tools for desktop and mobile; using Emacs and Ripgrep as desktop search engine""
- Discussion: HN, Reddit

Searching the Google Reader archives

> A tutorial on how to do manual searches of the 2013 Google Reader archives on the Internet Archive. Google Reader provides fulltext mirrors of many websites which are long gone and not otherwise available even in the IA; however, the Archive Team archives are extremely user-unfriendly and challenging to use even for programmers. I explain how to find & extract specific websites.

A little known way to 'undelete' a blog or website is to use Google Reader (GR). Unusual archive: Google Reader. GR crawled regularly almost all blogs' RSS feeds; RSS feeds often contain the fulltext of articles. If a blog author writes an article, the fulltext is included in the RSS feed, GR downloads it, and then the author changes their mind and edits or deletes it, GR would redownload the new version but it would continue to show the version the old version as well (you would see two versions, chronologically). If the author blogged regularly and so GR had learned to check regularly, it could hypothetically grab different edited versions, even, not just ones with weeks or months in between. Assuming that GR did not, as it sometimes did for inscrutable reasons, stop displaying the historical archives and only showed the last 90 days or so to readers; I was never able to figure out why this happened or if indeed it really did happen and was not some sort of UI problem. Regardless, if all went well, this let you undelete an article, albeit perhaps with messed up formatting or something. Sadly, GR was closed back in 2013 and you cannot simply log in and look for blogs.

Archive Team mirrored Google Reader. However, before it was closed, Archive Team launched a major effort to download as much of GR as possible. So in that dump, there may be archives of all of a random blog's posts. Specifically: if a GR user subscribed to it; if Archive Team knew about it; if they requested it in time before closure; and if GR did keep full archives stretching back to the first posting.

AT mirror is raw binary data. Downside: the Archive Team dump is *not* in an easily browsed format, and merely figuring out what it *might* have is difficult. In fact, it's so difficult that before researching Craig Wright in November–December 2015, I never had an urgent enough reason to figure out how to get anything out of it before, and I'm not sure I've ever seen anyone actually use it before; Archive Team takes the attitude that it's better to preserve the data somehow and let posterity worry about *using* it. (There is a site which claimed to be a frontend to the dump but when I tried to use it, it was broken & still is in December 2018.)

Results

Success: raw HTML. My `dd` extraction was successful, and the resulting HTML/RSS could then be browsed with a command like `cat *.warc | fold --spaces -width=200 | less`. They can probably also be converted to a local form and browsed, although they won't include any of the site assets like images or CSS/JS, since the original RSS feed assumes you can load any references from the original website and didn't do any kind of data-URI or mirroring (not, after all, having been intended for archive purposes in the first place…)

---

1. For example, the `info:` operator is entirely useless. The `link:` operator, in almost a decade of me trying it once in a great while, has never returned remotely as many links to my website as Google Webmaster Tools returns for inbound links, and seems to have been disabled entirely at some point. ↩

2. WP is increasingly out of date & unrepresentative due to increasingly narrow policies about sourcing & preprints, part of its overall deletionist decay, so it's not a good place to look for references. It is a good place to look for key terminology, though. ↩

3. This probably explains part of why no one cites that paper, and those who cite it clearly have not actually read it, even though it invented racial admixture analysis, which, since reinvented by others, has become a major method in medical genetics. ↩

4. University ILL privileges are one of the most underrated fringe benefits of being a student, if you do any kind of research or hobbyist reading—you can request almost anything you can find in WorldCat, whether it's an ultra-obscure book or a master's thesis from 1950! Why *wouldn't* you make regular use of it? Of things I miss from being a student, ILL is near the top. ↩

5. The complaint and indictment are not necessarily the same thing. An indictment frequently will leave out many details and confine itself to listing what the defendant is accused of. Complaints tend to be much richer in detail. However, sometimes there will be only one and not the other, perhaps because the more detailed complaint has been sealed (possibly precisely because it is more detailed). ↩

6. Trial testimony can run to hundreds of pages and blow through your remaining PACER budget, so one must be careful. In particular, testimony operates under an interesting & controversial price discrimination system related to how court stenographers report—who are not necessarily paid employees but may be contractors or freelancers—intended to ensure covering transcription costs: the transcript initially may cost hundreds of dollars, intended to extract full value from those who need the trial transcript immediately, such as lawyers or journalists, but then a while later, PACER drops the price to something more reasonable. That is, the first "original" fee costs a fortune, but then "copy" fees are cheaper. So for the US federal court system, the "original", when ordered within hours of the testimony, will cost <$7.25/page but then the second person ordering the same transcript pays only <$1.20/page & everyone subsequently <$0.90/page, and as further time passes, that drops to <$0.60 (and I believe after a few months, PACER will then charge only the standard $0.10). So, when it comes to trial transcript on PACER, patience pays off. ↩

7. I've heard that LexisNexis terminals are sometimes available for public use in places like federal libraries or courthouses, but I have never tried this myself. ↩

8. I advise prepending, like `https://sci-hub.tw/https://journal.com` instead of appending, like `https://journal.com.sci-hub.tw/` because the former is slightly easier to type but more importantly, Sci-Hub does not have SSL certificates set up properly (I assume they're missing a wildcard) and so appending the Sci-Hub domain will fail to work in many web browsers due to HTTPS errors! However, if prepended, it'll always work correctly.↩

9. To further illustrate this IA feature: if one was looking for Alex St. John's "'Judgment Day Continued…'", a 2013 account of organizing the wild 1996 *Doom* tournament thrown by Microsoft, but one didn't have the URL handy, one could search the entire domain by going to `https://web.archive.org/web/*/http://www.alexstjohn.com/*` and using the filter with "judgment", or if one at least remembered it was in 2013, one could narrow it down further to `https://web.archive.org/web/*/http://www.alexstjohn.com/WP/2013/*` and then filter or search by hand.↩

10. If any Blogspot employee is reading this, *for god's sake stop this insanity*!↩

11. Uploading is not as hard as it may seem. There is a web interface (user/password: "'genesis'"/"'upload'"). Uploading large files can fail, so I usually use the FTP server: `curl -T "$FILE" ftp://anonymous@ftp.libgen.is/upload/`. ↩

12. Although flatbed scanning is sometimes destructive too—I've cracked the spine of books while pressing them flat into a flatbed scanner.↩

13. My workaround is to export from gscan2pdf as DjVu, which avoids the bug, then convert the DjVu files with `ddjvu -format=pdf`; this strips any OCR, so I add OCR with `ocrmypdf` and metadata with `exiftool`.↩