Doing Digital Methods

Richard Rogers

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Contents

SIX

SEARCH AS RESEARCH

Repurposing Google

Transforming the consumer information appliance into a research machine

Search engine results for repurposing: Google studies and societal search

The chapter is dedicated to the question of search as research, and in particular how Google, the dominant web search engine, may be repurposed as a research machine both for medium as well as social research. After considering the extent to which one is only studying Google when perusing search engine results, ultimately, the goal is to consider how to perform social research with Google, or what, in short, may be termed 'societal search'. Here one would employ Google, in separate exercises elaborated below, for the study of local or national concern as well as the study of partisanship.

Since the mid-2000s Google has offered so-called local domain Googles, where one performs searches at google.nl (in the Netherlands), google.fr (in France), google.de (in Germany), and so forth (see Figure 6.1); they are also called 'regions' and are accessible via the advanced settings if one is not located in the country corresponding to the local domain Google (Kao, 2017). In other words, the engine user is directed by default to the local domain Google on the basis of user's location, read from the IP address or set location preference. After an engine query is made, the results are returned in the local language, as are the advertisements. Google's 'local' also serves its business model, and the results are delivered within national legal jurisdictions, such as European countries where users have the right to oblivion, or certain results removed (Floridi et al., 2015).

Figure 6.1 Local domain Googles, where one is sent by default when located within that country. The graphic shows how Google globalized, or glocalized, in the sense of making its global product local.

Source: Google language tools, https://web.archive.org/web/20111118022541/http://www.google.com/language_tools. See also 'See results for a different country', https://support.google.com/websearch/answer/873?hl=en.

Apart from returning advertisements and legal context, local domain Googles also return 'local' results, and the question concerns how the local is epistemologically constituted by Google. Which types of sources are returned (by default as well as by special advanced settings)? In what sense are they local sources, or how to describe Google's sense of local? Thus, prior to being able to conduct social research with Google ('societal search'), one must interrogate the engine's definition of the local, or conduct medium research ('Google studies').

We start thus by examining the utility of Google's sense of the local. Does it enable the study of local (or national) concern? May one ultimately read societal tendencies or trends

through search engine returns? These questions are posed so as to develop a new form of search engine critique as well as usage, where search becomes research, or the engine becomes more than an information, advertising and legal machine. Put differently, may one perform social research with Google, or is one always only studying Google?

In introducing this line of search engine (epistemological) critique and repurposing, it is instructive to discuss, briefly, how the search engine and its results are often held up to critical examination, before returning, in the main section, to the issues at hand when considering the engine as research machine as opposed to another type. Search engine critique, at least as presented here, could be of at least five varieties: the engine's results biases, cognitive effects, surveillance regime, hegemony as well as its invisible materiality. Each is briefly taken in turn, before the discussion moves on to how to perform search as research in the form of medium research ('Google studies') and social research ('societal search').

Search engine critique

Early investigations into engine result pages have been dominated by search engine bias studies, where the aim of the research is to uncover and explain the absence and presence of certain results, especially at the top of the returns, considered to be the most valuable space or 'real estate'. Early search engine critique focused on the ways in which search engines and other gateways to the web privilege certain information sources over others. 'Preferred placement' critique – which originated when AltaVista was the dominant search engine – concerns placing ads where 'editorial' content is meant to be, and obfuscating the crucial distinction (Rogers, 2000). Search engines, in another classic, idealistic critique, clash with the architecture of the internet as well as its design values, because they undermine 'the substantive vision of the web as an inclusive democratic space' (Introna and Nissenbaum, 2000: 181). From this perspective, differentiation between sources on the basis of ranking algorithms is interpreted as threatening the egalitarian, or democratizing, potential of the web.

Exclusion and deep burying

Search engines supposedly give rise to inequality of information in at least three ways: the exclusion of certain sources, the artificial boosting of others, and the naturalization of results that favour the rich and powerful. First, it is improbable for search engines to index the entire web. In their pioneering and often-cited articles in the late 1990s, computer scientists Lawrence and Giles investigated six search engines' web coverage. In finding it to be incomplete, they concluded that in fact there is a hidden or 'deep web' never visited by crawlers (Lawrence and Giles, 1998, 1999). The deep web thus became an object of study and fascination, as a subset out of reach and difficult to plumb. Second, search engine results may be biased because search engines are prone to manipulation. Since the early

days, search engine optimizers and content producers have been attempting to rank well in search engines by trying to influence engine results, leading Google to issue webmaster guidelines but also to demarcate the web into neighbourhoods, some good and others spammy (Pringle et al., 1998). More recently there have been calls to audit algorithms to uncover not so much epistemological inequalities or advertising in disguise, but rather machine bias and discrimination, such as when African-American name searches trigger Google ads for background checks or when women receive ads for lower-paying jobs on Google's ad network (Sandvig et al., 2014; Datta et al., 2015). Third, search engines have a kind of favouritism built in where 'the rich and the powerful', as well as the web savvy, will dominate the top (Introna and Nissenbaum, 2000: 177; Hindman, 2008). Those sites receiving the most links (from sites which themselves receive many links) are boosted, as if organically, as the results produced by the algorithm are often described. When mention is made of how sources are buried by engines (through a lack of indexing, or by their measurable dearth of authority or influence), one may speak of engines' social effects.

Attention deficit

There is additionally critique about engines' cognitive effects, with the rise of what could be called Google-assisted intelligence. To the cultural critic, Nicholas Carr, we are witnessing the decline of 'contemplative man', and the coming of 'flickering man' (Carr, 2007). Through 'googling' during conversations but also with the intrusion of smartphones, social media and messaging apps, there has been a concomitant rise of what Linda Stone has called 'continuous partial attention' (Stone, 2008). Users are not only distracted but also displaced, owing to the maintenance of remote intimacy as well as ambient friend-following.

User studies also have found what could be termed engine attention effects, whereby fewer and fewer results pages are browsed. As FairSearch (2010: 1), the trade industry alliance, put it more starkly: 'Links below the fold receive less than 1% of users' attention', which also introduces language used by journalists as well as marketeers about the value of not only the top results but also the screen space prior to scrolling. The 'fold' once referred to broadsheet newspapers, folded, whereby those headlines and stories above the fold were most significantly placed. Nowadays it is the size of the real estate affected by the shrinking screen space of the smartphone, and the overpopulation of the top of engine returns with Google properties, that has been a cause for consternation (and industry complaints).

Surveillance vessel

There are further critiques of engines to be aware of, as they also bear upon the potential use of Google as research machine. One concerns the search engine as surveillance vessel, which collects, stores and acts upon user information, such as language and location. One's 'flecks', pieced together by the search engine, form a data body (or second self) that has

agency, for information is readied and delivered to you on the basis of your profile (Critical Art Ensemble, 1998; Fuller, 2005). Search engines may be critiqued for filling out the data body through the collection and retention of one's search history, which, as the AOL release of users' rich search history data revealed, may be thought to contain a personal 'database of intentions' (Battelle, 2003). Perusing one's data dump, a download of everything Google has stored about you, has been likened to 'reading the diary I hadn't intended to keep' (Hanley, 2018).

Projects such as scroogle.org (now discontinued), TrackMeNot and Ruin My Search History are reactions to Google as data collection (and surveillance) vessel, where the former, once an interface atop Google, 'crumbled' cookies, blocked ads and Google properties, and saved no user history. The latter two, a Firefox extension and a webpage, camouflage user queries and mask one's history by sending random words, or noise, in an act of obfuscation (Brunton and Nissenbaum, 2011; Serhat, 2016). On a regulatory level, 'do not track' has been implemented as a feature in browsers (albeit disabled by default), requesting that the website loaded in the browser does not itself track the user's behaviour or allow a third party to do so, such as Google's Doubleclick trackers. From a research standpoint, checking the 'do not track' box in the browser's privacy settings could give the user a sense of which websites are complying, and which are ignoring the user's wishes – by noticing whether ads are still following you around the web, so to speak. DuckDuckGo, one search engine that does not track its users, once ran a billboard ad for a month in Google territory (San Francisco), marketing itself as a privacy enhancing technology (or PET), to use the standard term (see Figure 6.2).

Figure 6.2 DuckDuckGo's billboard advertisement in San Francisco, January 2011.
Picture credit: Gabriel Weinberg.

Googlization

A fourth, broader set of critiques centres on actual and conceptual monopolization, or the 'creep' of Google as well as Google's 'free business model' into more and more markets and more service areas (Rogers, 2009b). A glance at Google's list of products (which appears as a

menu item in its flagship web search service) is telling (maps, places, news, shopping, travel, etc.), as is its market share across search markets, and the related decline of national search engines, with distinctively different algorithms than Google's, with relatively few exceptions such as Baidu in China, Naver in South Korea, Seznam in the Czech Republic, Yandex in Russia and Yahoo! in Japan. More conceptually, one could argue that Google's PageRank and subsequent algorithmic updates that have driven the engine towards personalization has led to algorithmic concentration, which is a variation of market, or media, concentration, describing the search engine company as both monolithic as well as hegemonic.

The term to surface summarizing this political-economy style critique is 'Googlization', put forward by library scientists concerned about the Google Books project, and its march (or creep) into the hallowed halls and shelves of the library (Vaidhyanathan, 2011). Here Google becomes mass media, with a model that strives to serve the greatest audience (with a quality level to match it), together with high barriers of entry for any competitors, given the size (and expense) of the infrastructure needed to compete against it. In an appeal for donations, Jimmy Wales, founder of Wikipedia, once opened by saying that 'Google might have close to a million servers' (Wikimedia, 2011). Wales was comparing the relative muscle of Google to Wikipedia's.

Rematerializing the cloud

The materiality of Google, once understudied, has become the subject of a variety of exposés, ranging from investigative reporting on negotiations between the company and city and state governments, artistic and ethnographic trips to data centres and works of art that show buildings scrubbed from Google maps (Burrington, 2014). To be sure, there is a vast technological infrastructure in the service of delivering fast engine results and seeding 'the cloud'. The infrastructures in turn compete for natural resources with the local population, farmers and others in what are dubbed 'water wars', for they require cooling (Gallucci, 2017). The materiality of the cloud is captured in Timo Arnal's (2014) artwork, *Internet Machine*, as well as in Trevor Paglan's (2016) *Deep Web Dive* where the artist swims to an undersea cable. Paglan's work is about surveillance, though it does point to the physicality of the cloud (and the lengths to which one must go to uncover it). In Arnal's work, the nondescript, often secret, data centre is actually entered (see Figure 6.3). After a series of security layers, the camera takes us down long corridors, laced with cables, and through doors to the server rooms that whirr with the sounds of fans. Apart from surprisingly high noise levels there are also temperature extremes; there are 'hot aisles' and cold ones (Levy, 2012). To keep the systems up there are massive diesel generators for back-up power, and steel containers of cooling water in case of calamity. It is remarkably emptied of people, with few signs of maintenance workers.

The cloud, the airy metaphor that deftly stands in for physical systems of cables, data centres, servers and electricity, is often illustrated with impressive numbers – the billions of

searches served in milliseconds around the world and the number of bytes (zettabytes even) held 'up there', in what technology historians would call exemplary of the 'arithmetic sublime', whereby the reader stands in awe of its incomprehensible vastness, well beyond any human mathematical capacity. The term 'technological sublime' was coined to capture the statistical and other jaw-dropping descriptions of great technological displays such as the illuminations of city streets in the nineteenth century, when electricity and public lighting were introduced (Nye, 1994). Such thinking is often followed by what these numbers (and the awe) obscure: a sprawling political economy of resource extraction, low-wage work and data centre user capture, fuelling growth, such as when Apple OS nudges its users to save files onto iCloud rather than their own hard drive (Merritt, 2013).

Figure 6.3 Facebook data centre signage. 'Other companies don't put their names on their data centers.'

Source and picture credit: Lardinois, 2016.

The clouds of the likes of Amazon, Google, Apple and Facebook have now been brought down to earth through the materialist and environmental critique, one that has found a starting point for research in the lists of data centre locations as well as their resource consumption provided by the companies themselves in displays of corporate social responsibility. That is, for some years now the companies have issued reports not only on 'transparency' (related to requests from governments around the world to block content or identify users) but also on 'environmental responsibility' where in one of Apple's documents, for example, it is stated that in 2016 the company used 630 million gallons of water (up 10 percent from the previous year's consumption owing to the data centres) (Apple, 2017). On back-ups, it also burned 261,580 gallons of diesel. The listing of such figures is couched less in the prose of technological wonderment than in the incremental progress towards a more sustainable pace.

Google studies or societal search?

The very idea that one may use Google as social research machine is not unusual, when one considers that the science built into its algorithms (and entire apparatus) is in the first instance a variation on citation analysis, adapted to the web (Brin and Page, 1998; Rieder, 2012; Marres, 2017). The difference between the engine as it was and as it grows, however, lies largely in a change in the engine's definition of 'relevance' (Van Couvering, 2007). Where once results were deemed to be the best match between document (page) and subject matter (query) (and ranked by influential inlink counts), increasingly that match has been made less on the basis of content than on other variables, too, such as user clicks, page freshness and domain age. Where it once had little or nothing to do with it, now relevance is in some sense user-driven, or perhaps consumer-driven, if one prefers to emphasize the commercialization of Google's results. More conceptually, search results are a product of our 'living within [Google's] lab', meaning that we as users are all a part of Google's experiments and beta-testing that previously would have been performed in-house, with user groups or with students spot checking results against a list of what would constitute desirable outcomes (Davies, 2015: 377). Results now are adjusted according to how they are actually used rather than arriving pre-conceived from the beautiful mind alone or tested in-house.

The delivery of relevant pages based on user feedback could be thought of as one means of determining hierarchies of sources and societal concerns. Or such is the question. What sorts of source hierarchies are revealed when studying engine results? Is one able to study societal concern, or is one always only studying Google? In the following the answer lies somewhere in the middle, given that (on the one hand) engine effects are not to be eliminated, but (on the other) may be identified as well as mitigated.

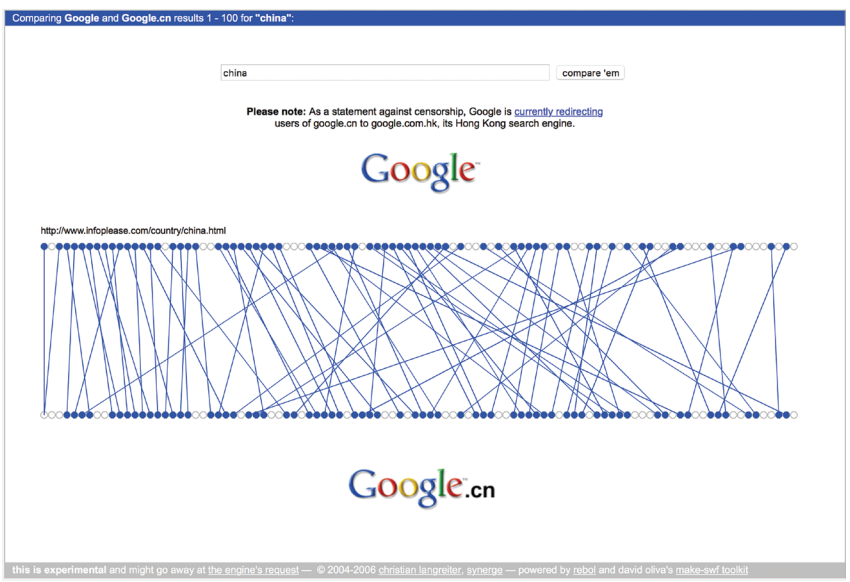

Figure 6.4 Visualization of google.com and google.cn results as technique for comparison.
Source: Langreiter, 2017.

Medium research as Google studies

Below I begin by using the engine for medium research ('Google studies'), before determining how (and whether) it may be used for social research ('societal search'). The Google studies projects ultimately seek to pave the way for societal search.

The first ones concern the types of sites returned in local domain Googles (google.nl, google.be, google.de., google.fr and so forth, now accessible as 'regions' in the advanced settings), and invite questions concerning Google's definition or sense of the 'local.' What is local to Google? Here one is able to critique Google's capacity as research machine for cross-country analysis by showing the extent to which Google returns transnational, regional or some (other) combination of results in its local domain engines. One compares the results of the same query across multiple local domain Googles (see Figure 6.4).

This project is medium research in the sense that the analysis seeks to tease out a Google notion of the local. In this project one queries one ambiguous or underspecified term of relevance in multiple locations or local domains (Rogers, 2013b). The analysis in question queries [diversidad] (or diversity) first in three pertinent local domain Googles (Colombia,

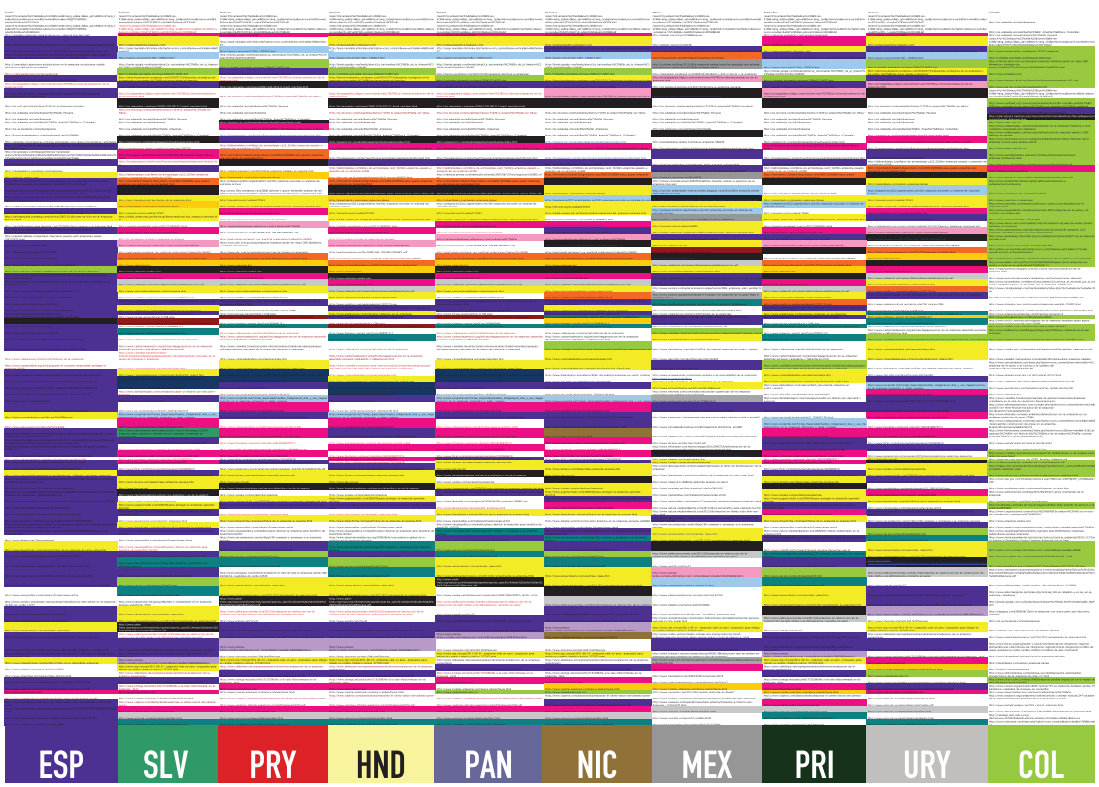

Figure 6.5 Locations of sources compared in local domain Googles in Spanish-speaking countries, where the majority are from Spain. Analysis by Natalia Sánchez Querubín and the Digital Methods Initiative, 2011.
Source: Rogers, 2013b.

Peru and Venezuela, all in the Amazon river basin) and subsequently across Spanish-speaking domains, finding that the vast majority of the results are sources in Spain, rather than from Latin America (see Figure 6.5). Spanish sources are identified not only by the country domain (.es), but also from the 'about us' information as well as the specificity of the Spanish language used on the pages.

In another exercise of this sort [Amazonia] (Amazon) is queried in Spanish-language local domain Googles, and the URLs returned per domain are compared. For Spain (google.es) the results originate largely in Spain. For all the other countries Google provides in each sources from Spain, and the remainder are Latin American results, nearly uniform for each country. It is as if there is a result set for Spain and another one for all of Latin America (see Figure 6.6). Google's local is national for Spain but transnational (and rather colonial) for Latin America, where Spanish sources retain authority. Here one may pursue search engine returns as one expression of the coloniality of knowledge (Grosfoguel, 2004).

Google studies with social research implications

In another comparative source origin project, ["human rights"] is queried in various local domain Googles (in the respective local languages), asking whether the results return local or non-local pages. This undertaking is again medium research, or Google studies, but the implications begin to fall into the realm of social research or societal search. That is, taking the query into account, one also may ask which countries have well-developed content providers for human rights issues, and which rely on non-local, perhaps even establishment sources. The case study explores the distinctiveness of local results across local domain Googles, with the additional consideration of the type of query made, human rights, which to some is a universal as opposed to local or regional issue, as in the first sample project.

Where are the returned information sources based? The aim is to retrieve the location of the information sources outputted per local domain Google engine. The location of a website may be thought of in a number of ways, including country code top-level domain (ccTLD), registration (site owner's geographical location) and/or host (geographical location where a website is stored) (Sottimano, 2013). In this project location is gleaned from the address of the website's registrant (through the contact address on the website and/or its 'whois' information, when available). 'Local' sources are defined as those registered in the country of the local domain Google (e.g., for the results provided by google.com.eg, the source, anhri.net, is considered 'local' because it is registered in Cairo, Egypt).

After the term ["human rights"] is queried in the local language, the top 10 information sources are captured and geolocated, and the results visualized on a geographical map (see Figure 6.7). Remarkably, nearly half of the local domain Googles have no local results in the top ten sources returned. When comparing the number of local sources, the uneven distribution across national webs becomes apparent. The countries with the most local sources are European, some North American, South American and Asian countries.

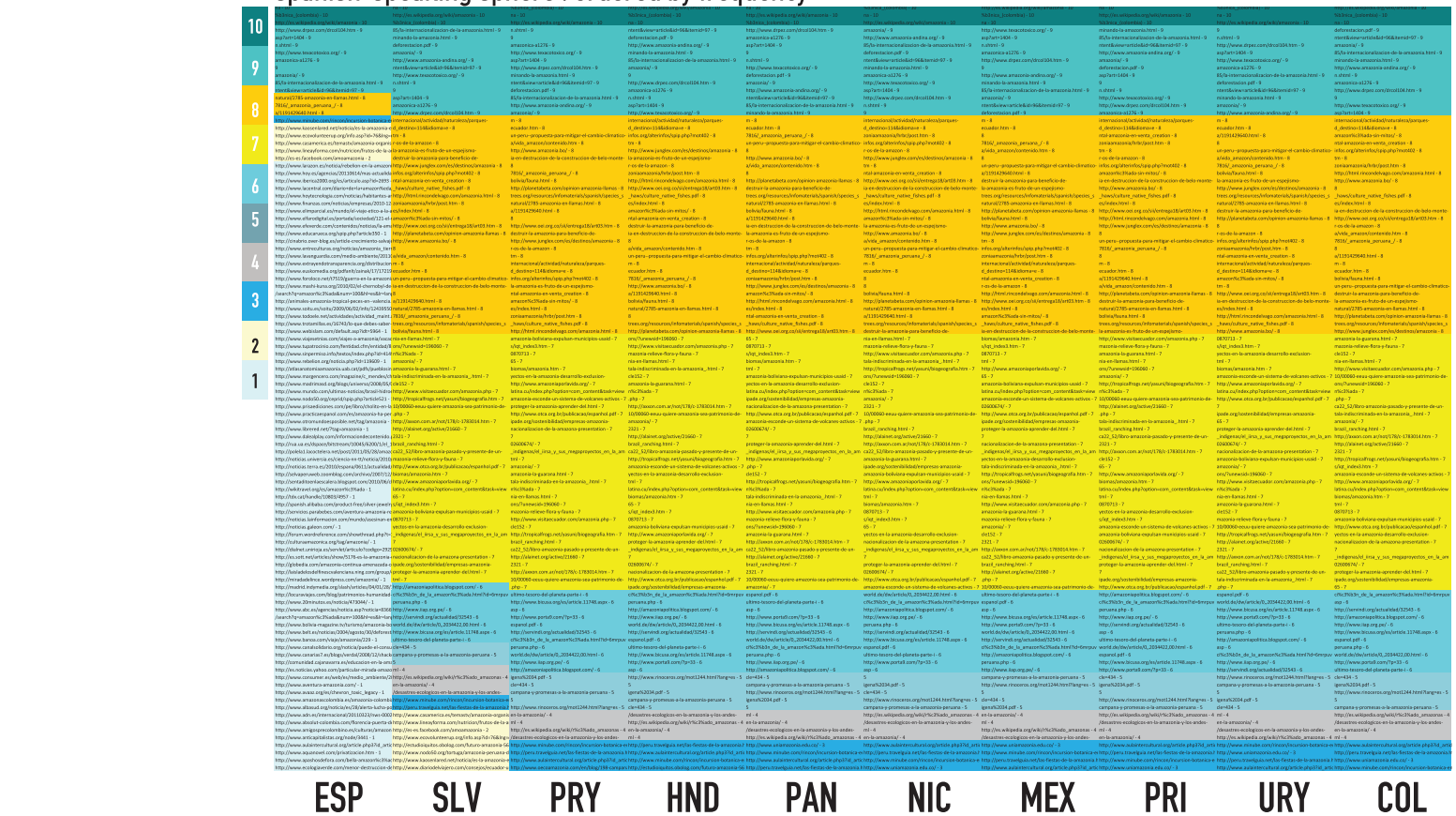

Figure 6.6 URLs compared across local domain Googles in Spanish-speaking countries; colours indicate the number of local domain Googles in which a set of results appear. Analysis by Natalia Sánchez Querubín, Diana Mesa and the Digital Methods Initiative, 2011.

Source: Rogers, 2013b.

Most countries in the top ranks have location-specific languages. African and Middle Eastern countries are found towards the bottom of the list.

The most prominent information source across 121 national Googles, queried in 43 languages, is un.org, with 80 of the local domain Googles returning un.org as one of the top ten results (see Figure 6.7). In the Arab-speaking Middle East and northern Africa (MENA), local

Figure 6.7 Bordered sources: local and international information sources. Graphic by Esther Weltevrede and Erik Borra, Digital Methods Initiative, Amsterdam, 2009.

sources (with the exception of anhri.net) are virtually absent (see Figure 6.8). The shared sources are primarily USA-based, and also include the Brazilian NGO, huquqalinsan.com. Half of the sources returned in the Arab-speaking countries are identical; in the 11 national Googles, un.org, arabhumanrights.org, hrw.org, ar.wikipedia.org and anhri.net appear at the top. On a number of the MENA Googles, we found local results on the second page.

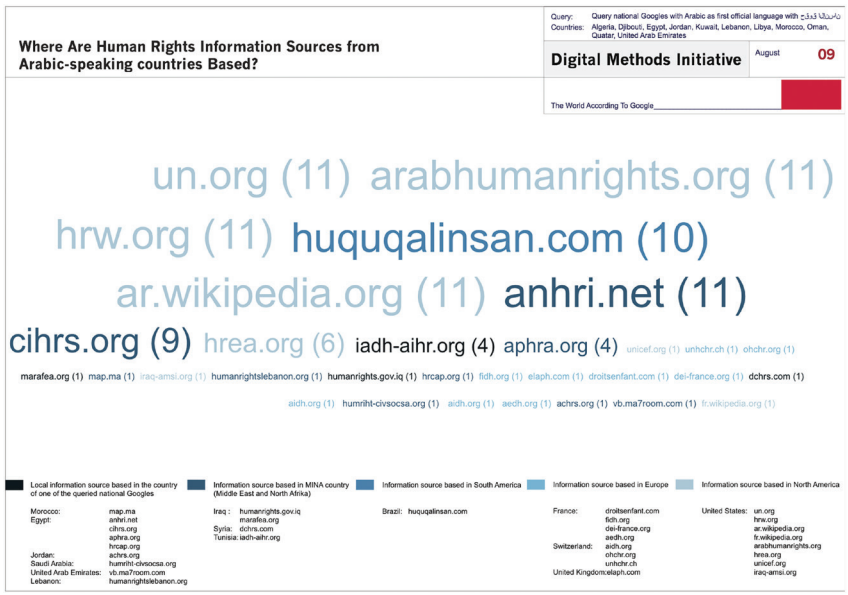

Figure 6.8 Bordered sources: where are the human rights information sources from Arab-speaking countries based? Graphic by Esther Weltevrede, Digital Methods Initiative, Amsterdam, 2009.

Societal search with Google studies artefacts

This project is principally societal research as we are looking to Google to provide a ranked list of societal concerns per local domain Google. Are there distinctive or similar rights that reach the top in Finland, the Netherlands, France, Italy, Switzerland, Germany, Austria, Sweden, Russia, Japan, Canada, the United Kingdom, Australia, Philippines, Ivory Coast and other countries?

The first step is to query the term [rights] in the local languages in the local domain Googles, e.g., [õigused] in google.ee, [direitos] in google.pt, etc. One may use a VPN to be located in the country or local Google domain in question, or use the region setting in advance search. The second step is to read and interpret the results and make lists of the top ten distinctive rights types, leaving them in the order that Google provided.

As noted above, the query design takes advantage of Google as research machine, and particularly its strength in dealing with ambiguous queries such as [rights] rather than its other strength of massive (fresh) site indexing, which is behind a second set of societal search projects below. With respect to its original strength, as Brin and Page (1998: 9) phrase it, 'the benefits of PageRank are the greatest for underspecified queries'. Discrete or less ambiguous keywords would decrease the salubrious algorithmic effects.

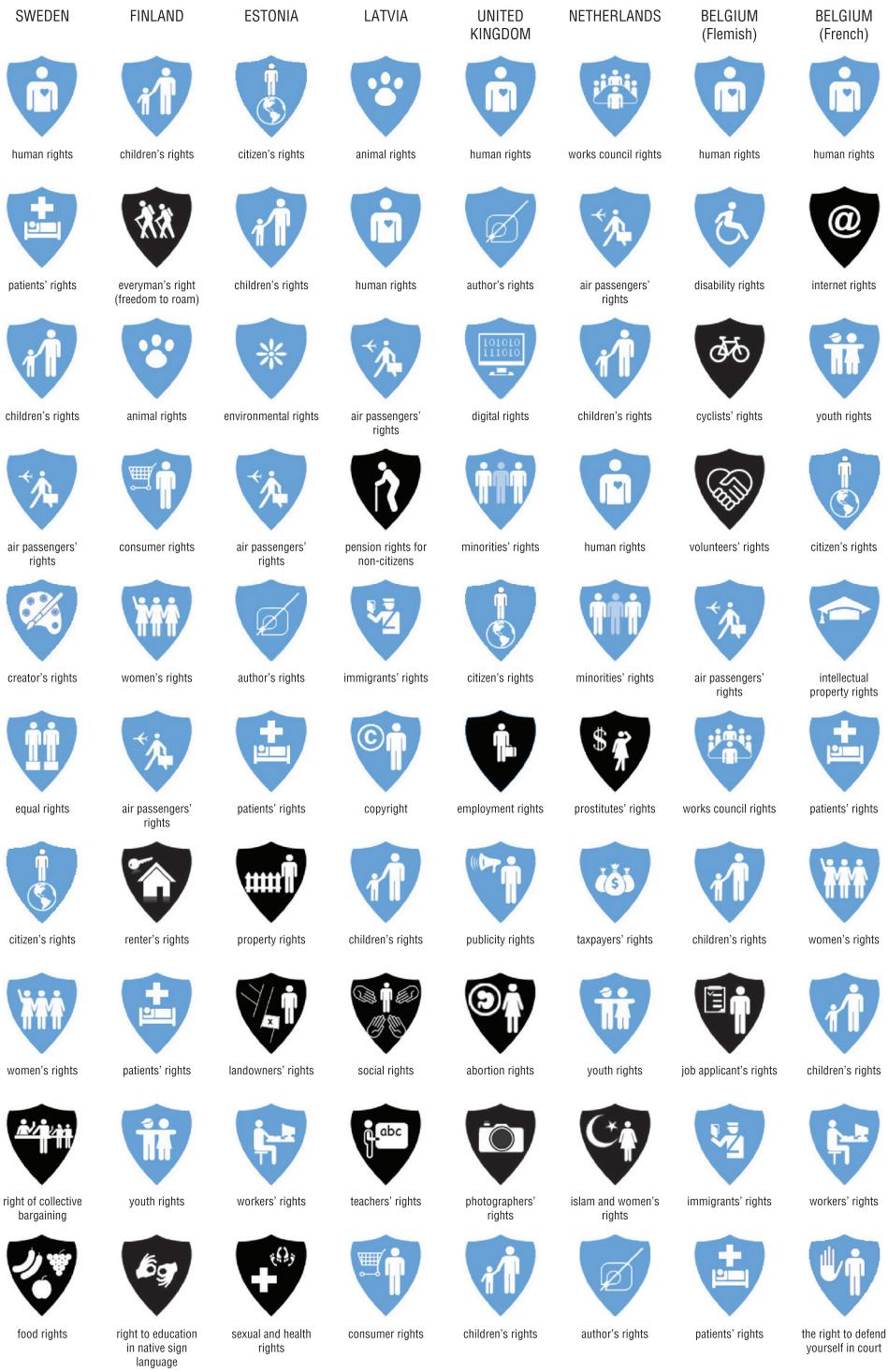

Figure 6.9 The nationality of issues: rights types (excerpt). Digital Methods Initiative, Amsterdam, 2009. *Source*: Rogers et al., 2009a.

When reading and interpreting the results, there are editorial decisions to be taken with respect to Google artefacts. Since the effort is to mitigate them (for societal search) rather than to highlight them (for Google studies), artefacts, however fascinating, may be removed.

For example, Wikipedia is a top result or nearly so for most local domain Google queries. One could make a separate project out of the differences in rights types across Wikipedia language versions, as is the effort in cross-cultural Wikipedia studies. Another Google artefact is the result R.I.G.H.T.S. (rightsforartists.com) in google.com. It is a Google artefact in the sense that it highlights how Google relies on certain 'signals' to boost websites in the rankings (Dean, 2016). Among other indicators, the word 'rights' is part of the URL, and R.I.G.H.T.S. is in the page header.

In the findings, rendered as labelled icons, countries could be said to have diverging hierarchies of concerns per (Google) country (see Figure 6.9). For example, everyman's rights in Finland, prostitutes' rights in the Netherlands, computer programmers' rights in Japan and the right to oblivion in Italy (the right to have personal data deleted) are unique to the respective countries. The order of appearance per country invariably differs.

Given the focus on cultural distinctiveness, it should be noted that the specific issue language per country is retained, rather than grouped as equivalents. Thus, LGBT rights in the United States and homosexual rights in Hong Kong are not considered the same. Indeed, one could make a small sub-study of the terms (and thus inclusiveness) across the local domain Googles for these particular rights as well as others.

In all, the short case study starting with the underspecified query, [rights], has found distinctiveness between rights types and rights hierarchies across the local domain Googles. One could consider techniques to harden these findings, such as returning to the idea of engaging in cross-cultural Wikipedia studies as well as other means of grounding the findings online. It is also a thought piece for discussing rights types cross-culturally.

Google studies and societal search combined: Source distance and partisanship detection

Google, as related above, creates hierarchies of credibility through returning ranked sources for a query. When the query is substantive, such as ["climate change"], sources at the top are given the privilege of providing information on the matter of concern, while others lower down are less likely to be read. Here the question concerns the distance from the top that partisan sites appear, giving voice to a particular side or position. The case in question is the climate change issue. Partisanship concerns giving voice, or a platform on its website, to climate change sceptics. Which sites mention the sceptics, and quote and represent their viewpoints? Are they close to the top of the engine returns for the query ["climate change"]? Source distance is medium research for it asks whether the web via Google (or Google in particular) gives the sceptics top-ranked space.

Indeed, in the first instance, it could be said that we are studying Google. Query Google and consider whether the engine's ranking procedures place sceptic-friendly websites towards the top of the climate change space.

It is a two-step query design (see sample project below). First, query ["climate change"] and save the results. Subsequently, keeping the results in the order they appeared, query each individual result for names of climate change sceptics, through [site:] queries, or the use of the advanced setting 'search one site'. Visualize where the sceptics appear in the top results (see Figure 6.10). (Such work also may be performed with the Lippmannian Device, also discussed below.) The sceptics are represented in a few sites returned (in the first 50 or so) but not at the very top. Put differently, in the journalistic convention both sides of the story are represented, but in the climate change space provided by Google the sceptics' presence is relatively scant, it was found.

When considering the results anew, however, it also could be said that we are undertaking social research as we are considering the presence of sceptics in the climate change source space more generally, and we are identifying specific sources where they are present. Without considering positive or negative mentions, one is studying the 'impact' of the sceptics – whether their overall presence is felt. One is also able to evaluate sources according to sceptic mentions. After closer reading, one notes there are sceptic-friendly 'science' websites as well as sceptic-funders. Another website type where sceptics appear is a watchdog site, with critical mentions of the sceptics. There are also those that do not name sceptics, providing no mentions. Through sceptic presence and absence source evaluation and characterization are performed.

Figure 6.10 Top climate change sources on the web, according to Google Web Search, resized according to the quantity of mentions of a climate change sceptic. Output of the Lippmannian Device and Tag Cloud Generator, Digital Methods Initiative, Amsterdam.

Source: Rogers, 2013b.

Conclusions: From Google critique to repurposing Google results

Above the question was posed concerning the capacity of Google to serve as a research machine, despite becoming a consumer information appliance (as well as a national advertising and legal machine) over the past two decades. Google is still a research machine in how it allows for foraging through online information as its creators envisioned in their seminal paper (Brin and Page, 1998). In the search engine critique that has since arisen, it also has evolved into a hegemon (market-wise), 'googlizing' industries and public resources (such as libraries and art institutions). It has purportedly had cognitive impacts (as illustrated by the coinage of 'flickering man'). Rather than an equalizer, it boosts both through its original algorithmic innovations as well as its subsequent tweaks the rich and now the popular. As an advertising company, Google also has been described as a front-page real estate hog, populating search engine results pages with its own properties, as well as a surveillance machine, inviting privacy-enhancing technologies that mask and obfuscate users as well as competitors as DuckDuckGo that trade on privacy. Google also captures users, nudging one to stay logged in, so disentangling oneself from the device has become burdensome.

But because Google recursively collects a user's data and recommends URLs on the basis of its 'knowledge' of the user, a researcher could consider avoiding obfuscation techniques such as 'track me not' or others (Howe et al., 2011), as they would potentially sully the engine results (garbage out, garbage in). Another approach would be having few traces available in the first place.

By installing a separate instance of a browser (such as Firefox) as a 'research browser', the researcher prepares a clean slate, free of cookies and other engine entanglements such as history and preferences. If one has a Google account, disable customized results, an option in one's web history. ('Do not track' could be enabled.) If one does not have a Google account, the Google cookies should be removed and not allowed to be set. The slate is cleaner (rather than completely refreshed) because Google by default serves localized results zoomed in to a city or similar. In the advanced settings, change the setting to a region (where there is a country drop-down list to choose from). Now results should be rather depersonalized.

Once a browser is so prepared, the work to undertake medium and social research commences. As indicated, medium research concerns engine effects on sources (including their placement in returns), whereas social research is conceived of as source evaluation with the aid of the engine.

To be clear, here we are turning the tables on Google, seeking to use it as a research machine – making social studies via or on top of engines – rather than being used by it as a subject of surveillance and targeted advertising. As you work, be aware that researching with Google requires vigilance, for the engine is continually striving to know you, and customize the results.

PROJECT 8

Determine the impact of climate change sceptics using search engine results

AUTHOR VIDEO

RESEARCH GOAL To show the impact of climate change sceptics through their quantity of hits and their distance from the top of search engine results in a set of sources on climate change (e.g., the top 100 Google returns for the query ["climate change"]).

1 Make a list of climate change sceptics. There is a variety of sources that provide lists of the names of climate change sceptics, as well as the organizations that sponsor them. One may triangulate expert lists or may make a list based on an associative query snowballing, as explained in the Issuecrawler chapter. One may also make a list on the basis of the keynote speakers and/or attendees of climate change sceptics conferences (e.g., hosted by the Heartland Institute).

The following list of climate change sceptics is derived from triangulating mentions of the names of sceptics by Source Watch, *Mother Jones* magazine and the sociologists, Aaron McCright and Riley Dunlap.

List of sceptics (as well as organizations that support sceptics)

Persons

- S. Fred Singer
- Robert Balling
- Sallie Baliunas
- Patrick Michaels
- Richard Lindzen
- Steven Milloy
- Timothy Ball
- Paul Driessen
- Willie Soon
- Sherwood B. Idso
- Frederick Seitz

Sceptical organizations

- American Enterprise Institute
- American Legislative Exchange Council
- Committee for a Constructive Tomorrow
- Competitive Enterprise Institute

- Frontiers of Freedom
- Heartland Institute
- Marshall Institute
- Science and Public Policy Institute

Such a list may be used during the analysis of the results (e.g., when noting the affiliation or partisanship of a source with (many) mentions of the sceptics). Are sceptics predominantly mentioned by the sceptical organizations, or also by other organizations? Or do watchdogs mention the sceptics most significantly? The question also may be whether there are other organizations that mention sceptics apart from sceptical organizations and watchdogs, indicating more widespread uptake. One also could consider undertaking such research across countries (or local domain Googles).

2 To facilitate the work, the Google Scraper may be used. One also may undertake the work manually by querying each term in each site; for example, in Google, query [site:www.epa.gov "Willie Soon"]. For the automated approach using the Google Scraper, first download and install the DMI toolbar, which is a Firefox extension. Install and use a separate instance of Firefox as a 'research browser'. In Firefox, choose preferences > privacy & security, then uncheck 'Block popup windows'. Switch off search history personalization in Google's preferences, log out of Google and enable 'do not track' in the browser settings.

3 Set Google preferences to return 100 results. Query ["climate change"] in google.com. (The list of sceptics provided above is principally American and not necessarily current, so one may wish to curate the list anew if researching another cultural space, though one also could study the globalization of climate change scepticism by compiling a current American list.)

4 Select and copy top 100 Google results. That is, on the Google results page, select all results (avoid the sponsored results), right-click and use 'view selection source' (in Firefox) and then copy the highlighted text. One may also use the Link Ripper (https://tools.digitalmethods.net/beta/linkRipper/), and input the Google results page.

5 .Paste that text into the Harvester, a separate tool at http://tools.digitalmethods.net/beta/harvestUrls/. Choose as output 'exclude URLs from Google and YouTube', 'only return hosts' and 'only return uniques', meaning unique hosts will be returned and later queried (e.g., http://www.epa.gov, rather than http://www.epa.gov/climatechange/kids/index.html).

6 Select and copy the results into the top box of the Google Scraper.

7 In the bottom box of the Google Scraper, enter the keywords (i.e., a list of climate change sceptics), one per line. Place the names in quotation marks.

8 Select the number of desired results (1–100). Use a larger number of results if there is an expectation that most sources mention the keyword, and a lower number if seeking only presence or absence of the mention of the sceptic per source. For the climate change sceptics, use 100 results. Name the output file and press Scrape Google (or Scrape Search Results on the Search Engine Scraper, with Google selected).

9 Keep the Google Scraper window open, and wait until the scrape is completed (i.e., until the output file is available). If a pop-up window appears, type in the captcha and close the pop-up window, and the Scraper will resume.

10 View the results as a source cloud. View multiple sources and single issue for a source cloud of each sceptic. View multiple sources and multiple issues for a source cloud of all sceptics.
11 View visualization for source distance or partisanship. Choose 'order of input in Google Scraper' to view how close to the top of the Google results the sceptic or sceptics resonate (source distance). Choose cloud output 'ordered by size' for a hierarchy of sources mentioning one or more sceptics, with those sources mentioning one or more sceptics the most at the top (impact and partisanship research).

In the case of the climate change sceptics listed above, it was found that there is distance between the sceptics and the top of the search engine returns (see Figure 6.10) Note that few sceptics appear on the websites of the top ten results in Google. When they do appear, their resonance is not particularly resounding.

One may evaluate sources according to the frequency with which each mentions the sceptics. There are sceptic-friendly sites, and less sceptic-friendly sites. From the visualization one is able to see the sceptic-friendly sources, such as realclimate.org and, to a lesser extent, climatescience.gov. Sourcewatch also is prominent, albeit as a progressive watchdog group 'exposing' the sceptics.

Remarkably, news sites, generally speaking, do not mention the climate change sceptics by name. While news watchers and listeners may have the impression that 'uncertainty' in the climate change 'debate' continues in a general sense (as opposed to, say, in more specific, scientific sub-discussions), 'uncertainty' appears to be discussed without resort to the well-known, or identified, sceptics.

With respect to the implications of the findings, one question concerns the extent to which the web stages climate change as a controversy *vis-à-vis* other media spaces, such as the news. Here the web is understood as a search-based medium, and controversy as the relative penetration of the sceptics in the climate change search results space. A comparison between the sceptics' resonance on the web and in the news could be a next step.

Video tutorial

For Project 8 on source distance, watch the video on how to operate the Google Scraper.

- 'The Google Scraper and Lippmannian Device' (12' 18"), https://www.youtube.com/watch?v=-sH5iPmcbQI4

Tool

'Google Scraper', digitalmethods.net. Available at https://wiki.digitalmethods.net/Dmi/ToolGoogleScraper

Google studies with the Google Scraper and societal search with the Lippmannian Device

There is a series of further assignment options, where we are researching how (and for whom) Google works, or societal trends with Google. One option (set out in considerable detail below) compares the results pages of local domain Googles, so as to provide an understanding of Google's sense of the local and discuss the implications of that understanding. In the second option, one employs the engine to output hierarchies of societal (or organizational) concern, following one of the Lippmannian Device research protocols below.

Before choosing one of the options, consider whether you wish to study the medium, some combination of the medium and societal trends, or societal trends. Generally, medium research here is considered to be diagnosing how Google works, for example, by typing [http] into the search bar in order to see how Google ranks URLs generally, or by searching for the same keywords over and again so as to study the effects of personalization on results. Techniques are described for studying Google's sense of the local, comparative source origin, and societal search (with Google artefacts considered). Doing medium studies of Google would be teasing out Google's sense of the local. A combination of medium studies and social research would be to diagnose how Google works and consider the societal implications such as which types sources are privileged, and which ones buried. Finally, studying societal trends refers to relying on Google's rankings either through substantive, underspecified queries such as [rights] or by working with the Lippmannian Device to identify how close to the top are the sceptics in the climate change space, or the global human rights agenda (to name two specific examples).

PROJECT 9

Investigate Google's sense of 'the local'

RESEARCH GOAL To ascertain Google's sense of the local by comparing the results of pages of local domain Googles (otherwise known as Google 'regions').

AUTHOR VIDEO

1. Query design. With respect to choice of term, choose a discrete term, and substantiate your choice (e.g., an unambiguous or underspecified query term). Use an unambiguous term such as [Amazonia] for the question of which sources dominate the results across Latin American countries (i.e., Google regions). Use an underspecified term such as [rights] for the question of which rights are dear per country (i.e., Google region).
2. Language. Apart from dictionaries, there are at least three options to translate a term between languages. Use languages that are available to you or your group, use Google Translate, or use 'languages' in the left-side column of Wikipedia articles.

3 Selection of Google regions to be queried. For the [Amazonia] query, the Google regions may be countries that are in the Amazon river basin, or Latin American countries more broadly. For the [rights] query, the sources may be varied and numerous. Consider building in comparison or contrast into your Google region selection, such as all former Soviet countries (where some are now in the European Union). Unless you use a research browser, and specific (re)search settings, Google will auto-detect your location, and privilege city-level results.
4 Query the term in local language(s) per Google region. Use Google advanced search, setting language and region.
5 Saving results. Set your preferences to the number of returns you wish to save. In your browser, choose File > Save as > html, and name your file using a naming convention such as BE_rechten_50_1DEC2019, where BE is the Belgian Google region, [rechten] the query, 50 the results count and finally the date of the query, with the month indicated in letters in order to avoid confusion between US and western European date formatting conventions.
6 Analysing search engine results pages: source origin and categorization.
   a There are multiple techniques for locating the 'origin' of a source: country domain (ccTLD), 'whois' information of the site registrant, and the contact information located on the websites. The location of the host (which may be local or far-flung) can be derived with the geoIP tool on tools.digitalmethods.net or another such tool. For the [Amazonia] query, the origins of the sources constitute the analytical question concerning Google's sense of the local.
   b Categorization – keyword specificity and A/B schemes. For the [rights] query, the specific rights privileged per country are of interest, such as the 'right to roam' in Finland. Here it is important to retain specificity and resist the urge to group similar rights under umbrella terms. For the [Amazonia] query, one could consider employing an A/B scheme (presence or absence; programme or anti-programme), such as the presence of extractive industries in the top ten results per country.
   c Categorization – source types in the results. For the [Amazonia] query, one may be interested in the presence of non-governmental or more specifically environmental sources in the results per country, or of non-Latin American results. One may glean source types from the top-level and second-level domains. See Wikipedia's articles on 'Top-level domain' and 'Second-level domain' for country-specific ones. For a finer-grained sense of the source type, peruse the 'about' page. One may pose critical questions of the dominance of one source type over another, inquiring into which sites have the privilege of being top sources, and providing information. For example, a ["climate change"] query may be dominated by intergovernmental sites, governmental agencies, NGOs and news outlets, while academic sources may be largely absent.

For finer-grained categorizations, consider using the 'other' category for items that do not fit the scheme, rather than 'neutral' which itself could be efforts made by actors (see the query design chapter).

7 Visualizing the findings. For the [Amazonia] query, a spreadsheet has Latin American countries as columns and source origin countries as rows. The 'visualization' is a colour-coded spreadsheet. For the [rights] query, per country (i.e., Google region), there is a list of rights types that have been artfully rendered as icons. One may consider using the triangulation tool, which takes lists of items as inputs and outputs commonalities and uniques.

The tag cloud generator at tools.digitalmethods.net provides means to visualize hierarchies, as does Wordle. If using Wordle, consider outputting all words horizontally and ordering them by frequency.

One also may consider populating a world map.

8 Drawing conclusions. Note that there are generally three discussions: medium research, some combination of medium and societal research, and societal research only, so to speak. For medium research, one is critiquing Google's sense of the local. If one chooses the combination of medium and societal research, the discussion could concern the extent to which Google is a globalizing or localizing machine, and the related issue of whether it may be used as a research machine, under what conditions and to what ends. If one is undertaking societal research, the capacity of Google to render countries' significant rights types becomes meaningful.

Video tutorials

For Project 9, there is a series of videos on how to transform Google into a research machine, set up a query, localize the outputs of a query and compare multiple results.

- 'The Research Browser' (1' 35"), https://www.youtube.com/watch?v=bj65Xr9GkJM
- 'Google Research Settings' (3' 48"), https://www.youtube.com/watch?v=Zk5Q_3g86qM
- 'Comparing Lists with the Triangulation Tool' (2' 54"), https://www.youtube.com/watch?v=jg9Uz-KcuuOE
- 'Localizing Web Sources' (4' 08"), https://www.youtube.com/watch?v=IyNMDUSBd9s

Consider watching a more general tutorial on analysing engine results in three ways:

- 'Analyzing Engine Results: Organization Types, Hierarchies of Concern, Political Leanings' (3' 50"), https://www.youtube.com/watch?v=MsnSJPXpFno

Tools

Google Scraper, digitalmethods.net. Available at https://wiki.digitalmethods.net/Dmi/ToolGoogleScraper

Harvester, digitalmethods.net. Available at https://wiki.digitalmethods.net/Dmi/ToolHarvester

Triangulation, digitalmethods.net. Available at https://wiki.digitalmethods.net/Dmi/ToolTriangulation

Resources

Google Translate, google.com. Available at http://translate.google.com/

"Top 500 Sites on the Web by Country," alexa.com. Available at http://www.alexa.com/topsites/countries

PROJECT 10

Map and interpret bias with the Lippmannian Device

AUTHOR VIDEO

RESEARCH GOAL Determine source partisanship (side-taking) as well as its distribution of concern

The Google Scraper, when used principally for societal search, is also referred to as the Lippmannian Device. There are two overall use cases for the Lippmannian Device: source partisanship and source distribution of concern. For source partisanship, the question concerns the detection of side-taking by a particular source through its mentioning or failure to mention particular issue language. Above it was noted that particular organizations mentioned the climate change sceptics while others averred. For research on the distribution of concern one is often given a list of issues that a particular organization engages in, advocates for, or otherwise 'does'. The question is whether particular organizations show attention to particular issues (over other issues) through frequency of mentions on their websites. Here one relies on Google's second strength (massive, presentist site indexing) and renders a distribution of attention to a set of issues.

Lippmannian Device?

As a term the Lippmannian Device refers to a piece of equipment for mapping and interpreting bias, or, as indicated, it may be employed to gain a rough sense of a source's partisanship and distribution of concerns. It is named after Walter Lippmann, the American journalism scholar who in his *Public Opinion* book of 1922, and particularly in his sequel to it of 1927, *The Phantom Public*, called for a coarse means of showing actor partisanship:

> The problem is to locate by clear and coarse objective tests the actor in a controversy who is most worthy of public support. (Lippmann, 1927: 120)
>
> The signs are relevant when they reveal by coarse, simple and objective tests which side in a controversy upholds a workable social rule, or which is attacking an unworkable rule, or which proposes a promising new rule. By following such signs the public might know where to align itself. In such an alignment it does not … pass judgment on the intrinsic merits. (Lippmann, 1927, 120)

The device does not answer all of Lippmann's calls, though it seeks to begin with them by addressing a seminal Lippmannian sense (partisanship) as well as an extended one (distribution of partisanship). It also advances the calls by Lippmann, in an attempt to enrich the partisanship notion with the idea of distribution of concern on the part of actors. They may have a list of campaigns or issues they are working on, but which garner more returns? The Lippmannian device queries Google, in a two-step process, and makes the results available in clouds (as well as in a spreadsheet).

Lippmannian Device project: Source clouds for the display of partisanship

The Lippmannian Device may be used to create source clouds that reveal partisanship towards a particular issue. With the tool, one may query a list of sources for one particular issue, or for a set of issues (keywords). Which source mentions 'security fence', which 'apartheid wall' and which neither? Source clouds display sources, each resized according to the number of mentions of a particular issue, according to Google.

Here is an example of employing the Lippmannian Device to study the 'synthetic biology' issue. Craig Venter has been considered a somewhat polarizing figure in the issue space, given that the science in his work often serves commercial interests and the (best-known) work itself is often construed as 'patenting life' (Glasner and Rothman, 2017). Thus we will ask which actors appear sympathetic to Craig Venter in the synthetic biology space.

Automated method

1. Set Google preferences to return 100 results. Query ["synthetic biology"] in Google.
2. Select and copy the top 100 Google results. That is, on the Google results page, select all results (avoid the sponsored results), right-click and use "view selection source" (in Firefox) and then copy the highlighted text.
3. Paste that text into the Harvester, a separate tool. Choose as output 'exclude URLs from Google and YouTube', 'only return hosts' and 'only return uniques' – meaning unique hosts will be returned and later queried (e.g., http://www.synbioproject.org, not http://www.synbioproject.org/topics/synbio101/).
4. Select and copy the results into the top box of the Lippmannian Device.
5. In the bottom box of the Lippmannian Device, enter the keyword [Venter] or, for greater specificity, ["Craig Venter"].
6. Select the number of desired results (1–1000). Use a larger number of results if there is an expectation that most sources mention the keyword. Name the output file and press Scrape Google.
7. Keep the Scraper browser window open and wait until the scrape is completed (i.e., until the output file is available). If a pop-up window appears, type in the captcha and close the pop-up window, and the Scraper will resume.
8. View the source cloud results – multiple sources and single issue.
9. View different orderings. Choose cloud output 'ordered by size' for a hierarchy of sources mentioning Venter, with those sources mentioning Venter the most at the top (see Figure 6.11).

Manual method

1. Query Google for ["synthetic biology"]. Save results. Commit each host in the results to a row in a spreadsheet.
2. Query each individual result in the top 100 for "Craig Venter". Use 'site' queries: [site:http://www.synbioproject.org "Craig Venter"]. For each host queried, place actual and optionally estimated result count in spreadsheet.

3 Show the quantity of mentions of Craig Venter in top sources on synthetic biology with a source cloud. Resize sources (e.g., synbioproject.org) according to the number of mentions.

You may wish to consider normalizing the findings on the basis of the overall sizes of the websites.

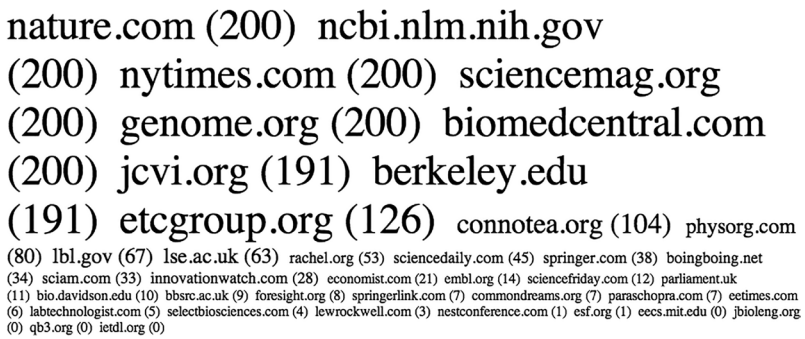

Figure 6.11 Craig Venter's presence in the Synthetic Biology issue space, March 2008. Top sources on "synthetic biology" according to a Google query, with number of mentions of Venter per source. Source cloud ordered by frequency of mentions. Output by the Lippmannian Device, Digital Methods Initiative, Amsterdam.

Lippmannian Device project: Issue clouds for concern distribution

The Lippmannian Device can also be used to create issue clouds that can reveal varying levels of concern by one or more sources. With the tool, one may query one or multiple sources for a set of issues or keywords. For example, Greenpeace International lists several issues for which it campaigns. Are there particular ones that are granted more attention (and perhaps resources)? Issue clouds display the campaign issues, each resized according to the number of mentions on the website (according to Google).

Another case in question are the issues listed by an NGO, Public Knowledge, dedicated to digital rights. Having copied and pasted their issues into the Lippmannian Device, and querying via Google publicknowledge.org for each issue separately, one may gain a sense of a distribution of concern. Here the next step may be to ground the findings with the actor itself and/or compare them to a larger agenda of the (digital rights) field.

1 Extract issues for the NGO by finding and copying its issue list. Public Knowledge's issue list is at http://www.publicknowledge.org/issues.

Copy and paste the issue list to bottom box of the Lippmannian Device, one issue per line, placing quotation marks around multiple-worded issues. An issue such as "Digital Millennium Copyright Act (DMCA)" could be inputted as follows:

"Digital Millennium Copyright Act" OR DMCA.

"700 MHz Spectrum Auction" OR "Spectrum Auction"

"Anti-Counterfeiting Trade Agreement"

"Broadband"
"Broadband Stimulus"
"Broadcast Flag"
"Comcast Complaint"
"Copyright"
"National Broadband Plan"
"Network Neutrality"
"Open Access to Research"
"Opening the White Space"
"Orphan Works"
"Patent Reform"
"Selectable Output Control"
"Text Message Petition"
"Trademark"
"WiFi Municipal Services"
"WIPO Broadcasters Treaty"

2 Place Public Knowledge's URL in the top box of the Lippmannian Device, http://www.publicknowledge.org.
3 Select the number of desired results (1–1000). Use a larger number of results if there is an expectation that the source mentions the issues in great quantity. For the public knowledge case, the setting 1000 results is entered. Name the output file (e.g., publicknowledge_issues_1DEC2019), and press Scrape Google.
4 Keep the Scraper browser window open, and wait until the scrape is completed (i.e., until the output file is available). If a pop-up window appears, type in the captcha and close the pop-up window, and the Scraper will resume.
5 View the issue cloud results. View issues per source. Choose cloud output 'ordered by size' for Public Knowledge's issue hierarchy (see Figure 6.12).

Video tutorials

For Project 10, there are videos on how to extract URLs from a web page and how to operate the Lippmannian Device.

- 'Extracting URLs from a Web Page via the URL Harvester' (1' 25"), https://www.youtube.com/watch?v=kzaq9DXfO_g
- 'The Google Scraper and Lippmannian Device' (12' 18"), https://www.youtube.com/watch?v=-sH5iPmcbQI4

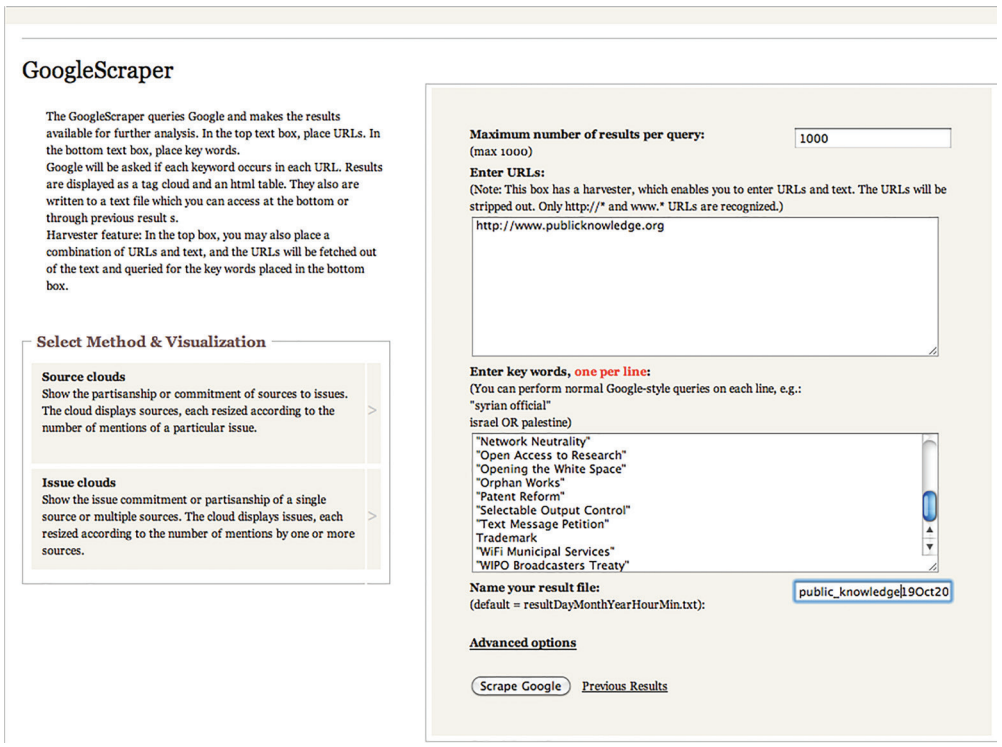

Figure 6.12 The making of Public Knowledge's concern distribution. Input of publicknowledge.org and its issues into Lippmannian Device. Output rendered as word cloud, showing the lower six issues on Public Knowledge's issue list, ranked according to number of mentions of its website according to Google site search, 2 October 2009. Output by the Lippmannian Device, Digital Methods Initiative, Amsterdam.

Tool

Lippmannian Device, *digitalmethods.net*. Available at https://wiki.digitalmethods.net/Dmi/ToolLippmannianDevice