Richard Rogers

Foundations of Digital Methods: Query Design

2017

https://doi.org/10.25969/mediarep/12536

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Rogers, Richard: Foundations of Digital Methods: Query Design. In: Mirko Tobias Schäfer, Karin van Es (Hg.): *The Datafied Society: Studying Culture through Data.* Amsterdam: Amsterdam University Press 2017, S. 75–94. DOI: https://doi.org/10.25969/mediarep/12536.

5. Foundations of Digital Methods

Query Design

Richard Rogers

digital methods and Online Groundedness

Broadly speaking digital methods may be considered the deployment of online tools and data for the purposes of social and medium research. More specifically, they derive from online methods, or methods of the medium, which are reimagined and repurposed for research. The methods to be repurposed are often built into dominant devices for recommending sources or drawing attention to oneself or one's posts. For an example of how to reimagine the inputs and outputs of one such dominant device, consider the difference between studying search engine results to understand in some manner Google's algorithms, or recent algorithmic updates, or treating them, as in the Google Flu Trends project, as indications of societal concerns. Here, there is a shift from studying the medium to using device data to study the societal. That is, akin to the digital methods outlook generally, Google Flu Trends and other anticipatory instruments use online social signals to measure trends not so much in the online realm but rather 'in the wild'.[1]

Once the findings are made the question becomes how to ground them, that is, with conventional offline methods and techniques, such as the Centers for Disease Control's means of studying flu incidence through hospital and doctor reports, as in the Flu Trends project, or through additional, online methods and sources. In digital methods research, online groundedness, as I have called it, asks whether and when it is appropriate to shift the site of 'ground-truthing', to use a geographer's expression. As a case in point, when verifying knowledge claims, Wikipedians check prior art through Google searches, thereby grounding claims via the search engine in online sources.

Digital methods thereby rethink conditions of proof, first by considering the online as a site of grounding, but also in a second sense. One makes social research findings online, and, rather than leaving the medium to harden them, one subsequently inquires into the extent to which the medium

1 The US Centers for Disease Control and Prevention (CDC) ran a competition in 2013-14 for instruments that use search and social media data to forecast influenza, and the one employing the data from Google Flu Trends won the award.

is affecting the findings. Medium research thus serves a purpose that is distinct from the study of online culture alone. As I will come to shortly, when reading and interpreting social signals online, the question concerns whether the medium, or media dynamics, is overdetermining the outcomes.

Making Use of Online Data: From the Semantic to the Social

As noted, digital methods make use of online methods, by which I refer to an array of techniques from the computational and information sciences – crawling, scraping, indexing, ranking, and so forth – that have been applied to and redeveloped for the Web. They refer to algorithms that determine relevance and authority and thereby recommend information sources as in Google's famed PageRank, but also boost all manner of items, from songs and 'friends' to potential 'followers'.

Many of the algorithms are referred to as 'social', meaning that they make use of user choices and activity (purposive clicks such as liking), and may be contrasted with the 'semantic', meaning that which is categorized and matched (as in Google's Knowledge Graph). Digital methods seek to take particular advantage of socially derived rankings, that is, users making their preferences known for particular sources, often unobtrusively. Secondarily, the semantic (sources that have been pre-matched or taxonomied) are also of value, for example when Wikipedia furnishes a curated seed list of sources ('climate change sceptics' as a case in point), which have been derived manually by information experts or the proverbial crowd guided by the protocols of the online encyclopaedic community.

The distinction between social and semantic is mentioned so as to emphasize Web-epistemological 'crowdfindings' (as implied by the 'social'), as distinct from 'results' from information retrieval.[2] Thus with digital methods, as I relate below, one seeks to query in order to make findings from socialised Web data (so to speak) rather than query in order to find pre-sorted information or sources, however well annotated or enriched with metadata.

Why Query Google (Still) for Research Purposes?

Over the course of the past decade or more Google arguably has transformed itself from an epistemological machine outputting reputational source

2 Crowdfindings is a term coined by Christian Bröer.

hierarchies to a consumer information appliance providing user-tailored results. Here I would like to take up the question of how and to which ends one might still employ Google as an epistemological machine.

There are largely two research purposes for querying Google: medium and social research. With medium research, one studies (often critically) how and for whom Google works. To which degree does the engine serve a handful of dominant websites such as Google properties themselves in a 'preferred placement' critique, or websites receiving the most attention through links and clicks? One would seek to lay bare the persistence of so-called 'googlearchies' that boost certain websites and bury others in the results, as Matthew Hindman's classic critique of Google's outputs would imply. Here the work being done is an engine results critique, where the question revolves around the extent to which the change in 2009 in Google's algorithmic philosophy, captured in the opening chapter of Eli Pariser's *Filter Bubble,* from universal to personalized outputs, dislodges or upholds the pole positions of dominant sites on the Web. Indeed, another critical inroad in engine results critique is the so-called filter bubble itself, where one would examine the effects of personalization, investigating Pariser's claim that Google furnishes increasingly personalized and localised results. In this enquiry, one may reinvigorate Nicholas Negroponte's 'Daily Me' argument and Cass Sunstein's response concerning the undesirable effects of homophily, polarization and the end of the shared public exposure to media which leaves societies without common frames of reference. In this line of reasoning, personalization leads to social atomisation and severe niching, otherwise known as 'markets of one', as described by Joseph Turow in *Niche Envy*. It also would imply the demise of the mass media audience.

In the second research strategy, there is a mode switch in how one views the work of the search engine (and for whom it could work). Google's queries, together with its outputted site rankings, are considered as indicators of social trends. That is, instead of beginning from the democratizing and socializing potential of the Web and subsequently critiquing Google for its reintroduction of hierarchies, one focuses on how examining engine queries and results allows for the study of social sorting. How to study the hierarchies Google offers? Which terms have been queried most significantly (at which time and from which location)? Do places have preferred searches? May we geo-locate temporal pockets of anxiety? The capacity to indicate general and localisable trends makes Google results of interest to the social researcher.[3]

3 Not only Google Trends but also Google Related Search provides means for studying keyword salience as well as the association between keywords, including co-occurrence.

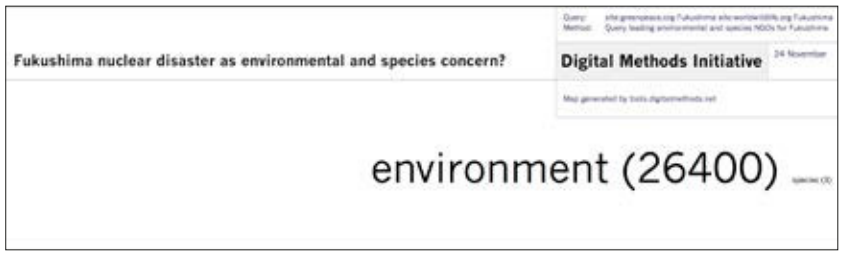
Fig. 5.1: Greenpeace campaigns, 1996-2012, ranked and arrayed as word cloud according to frequency of appearances on Greenpeace.org front page. Source: Data from the Internet Archive, archive.org. Analysis by Anne Laurine Stadermann.

Fig. 5.2: Greenpeace campaigns mentioned on Greenpeace.org as ranked word cloud, 2012. Source: Data from Greenpeace.org gathered by the Lippmannian Device, Digital Methods Initiative. Analysis by Anne Laurine Stadermann.

Apart from trends one may also study dominant voice, commitment and concern. One may ask in the first instance, when and for which keywords do certain actors appear high on the list and others marginal? Which actors are given the opportunity to dominate and drive the meaning of terms and their discussion and debate? Here the engine is considered as serving social epistemologies for any keyword (or social issue) through what is collectively queried and returned.

The engine also can be employed to the study of commitment in terms of the continued use of keywords by individual actors, be they governments, non-governmental organizations, radical group formations or individuals.

Fig. 5.3: Greenpeace with numerous mentions of Fukushima and World Wildlife Fund with few, November 2016. Source: Data and visualization by the Lippmannian Device, Digital Methods Initiative.



Here the researcher takes advantage not of the hierarchies inputted and outputted (socio-epistemological sorting) but of the massive and recent indexing of individual websites. For example, non-governmental organization Greenpeace once had the dual agenda of environmentalism and disarmament (hence the fusion of 'green' and 'peace'). Querying Greenpeace websites lately for issue keywords would show that their commitment to campaigning for peace has significantly waned in comparison to that for environmental causes, for green words resonate far more than disarmament ones. Here one counts incidences of keywords on Web pages for the study of issue commitment (see Figures 5.1 and 5.2).

One also may query sets of actors for keywords in order to have an indication of the levels of concern for an issue. For example, querying a representative environmental group and a species group (respectively) for Fukushima would show that the environmental group is highly active in the issue space whilst the species NGO is largely absent, showing a lack of concern for the matter (see Figure 5.3).

In all, for the social researcher, Google is of interest for its capacity to rank actors (websites) per social issue (keyword), thereby providing source hierarchies, and allowing for the study of dominant voice. It is also pertinent for its ability to count the incidence of issue words per actor or sets of actors, thereby allowing for the study of commitment through continued use of keywords.

Clean Google Results to Remove 'Artefacts'?

One might distinguish between the two research types above by viewing one as primarily doing media studies and the other social research. Yet in practice, the two are entangled with one another. As mentioned in the introduction,

here the entanglement assumes a particular form. Medium research is in service of social research in the sense of concentrating on the extent to which the findings made have been overdetermined by media effects.

It is important to stress from the outset that it not assumed that engine effects can be removed *in toto*, thus enabling a researcher to study 'organic' results, the industry term for editorial content untouched by advertising or preferred placement. Rather there should be awareness of a variety of types of routinely befouling artefacts ('media effects') that nevertheless are returned by the engine. Google properties (e.g. YouTube videos), Google user aids (e.g. 'equivalent results' for queried terms), and SEO'd products (whether through white or black hat techniques) are all considered media effects, and in principle could be removed or footnoted. There are software settings (e.g. remove Google properties from results), query design (use quotation marks for exact matches) and also strategies for detecting at least obviously SEO'd results.

The more problematic issue arises with any desired detection of the effects of personalization. The point here is that users now co-author engine results. The search engine thereby produces artefacts that are of the user's making. The search engine, once critiqued for its social sorting and Matthew effect in the results, leans towards inculpability, since users have set preferences (and had preferences set for them) and some results are affected. There is the question of detecting how many and which results are personalized in one form or another, according to one's location (country as well as locality), language, personal search history as well as adult and violent content filter.

Certain queries would likely have no organic results in the top ten, thus making any content cleaning exercise into an artificial act of removal, given that most users: a) click the top results, b) have the results set to the default of ten, and c) do not venture beyond one page of results. There are also special cases to consider for removal, such as Wikipedia, which is delivered in the top results for nearly all substantive queries, making it appear to be at once an authoritative source (for its persistent presence) and an engine artefact (for its uncannily persistent presence). Wikipedia's supra-presence, so to speak, provides a conundrum for the researcher who may wish to clean content of Google artefacts and media effects, and is perhaps the best case for retaining them at least in the first instance.

One way forward would be to remove the user, so to speak, and strive to have the engine work as unaffected as possible. Removing the user is a means of re-conjuring the pre-2009 distinction between universal results (served to all) and personalized results (served to an individual user). A

research browser would be set up, where one is logged out of Google, and no cookies are set. The ncr (no country redirect) version of Google is used, or one would query from a non-location, or obfuscated one.

Studying Media Effects or the Societal 'in the Wild'?

The question of whether Google merely outputs Google artefacts and medium effects or reveals social trends has been raised in connection with the flagship big data project, Google Flu Trends (Lazer et al. 2014). As mentioned at the outset, the project, run by Google's non-profit Google.org, monitors user queries for flu and flu-related symptoms, geolocates their incidence and outputs the timing and locations of heightened flu activity; it is a tool for tracking where the virus is most prevalent. Yet does the increased incidence of queries for flu and flu-related symptoms indicate a rise in the number of influenza cases 'in the wild', or does it mean that TV and other news of the coming flu season prompt heightened query activity? TV viewers may be using a 'second screen' and fact checking or enhancing their knowledge through search engine queries. Given that Flu Trends was over reporting for a period of time, compared to its baseline at the US Centers for Disease Control (and its equivalents internationally), the project seemed to be overly imbued with media effects.

Thus one may seek research strategies to study medium effects, formulating queries that in a sense put on display or amplify the effects. For which types of queries do more Google properties appear? How can Google be made to output user aids that are telling? How to detect egregiously SEO'd results?

When using Google as a social research machine, the task at hand, however, is to reduce Google effects, albeit without the pretension of completely removing them. This is the main preparatory work, conceptually as well as practically, prior to query design.

When Words are Keywords: A Query Design Strategy

The question of what constitutes a keyword is the starting point for query design, for that is what makes querying and query design practically part of a research strategy. When formulating a query, one often begins with keywords so as to ascertain who is using them, in which contexts and with which spread or distribution over time. In the following a particular keyword query

strategy or design is put forward, whereby one queries competing keywords, asking whether a particular term is winning favour and amongst whom.

The keyword has its origins in the notion of a 'hint' or 'clue'. The *New Oxford American Dictionary* (built into Apple OS's dictionary) calls it 'a word which acts as the key to a cipher or code'. In this rendering keywords do not so much have hidden but rather purposive meaning so as to enable an unlocking or an opening up. Relatedly, Raymond Williams, in his book *Keywords*, discusses them in at least two senses: 'the available and developing meanings of known words' and 'the explicit but as often implicit connections which people are making' (1976: 13). Thus behind keywords are both well-known words (elucidated by Williams's elaborations on the changing meaning of 'culture' over longer periods of time, beyond the high/low distinction) or neologistic phrases such as recent concerns surrounding 'blood minerals' or the more defused 'conflict minerals' mined and built into mobile phones. The one has readily available yet developing meanings and the other are new phraseologies that position. For the query design I am proposing, the purposive meaning of keywords is captured by Williams most readily in his second type (the new language). The first type may apply as well, such as in the case of a new use or mobilization of a phrase, such as 'new economic order' or 'land reform'. The question then becomes what is meant by it *this time*.

Concerning how deploying a keyword implies a side-taking politics, I refer to the work of Madeleine Akrich, Bruno Latour and others, who have discussed the idea that, far from having stable meanings (as Williams also related), keywords can be part of programmes or anti-programmes. Programmes refer to efforts made at putting forward and promoting a particular proposal, campaign or project. Conversely, anti-programmes oppose these efforts or projects through keywords. Following this reading, keywords can be thought of as furthering a programme or an anti-programme. There is, however, also a third type of keyword I would like to add, which refers to efforts made at being neutral. These are specific undertakings made *not* to join a programme or an anti-programme. News outlets such as the BBC, *The New York Times* and *The Guardian* often have dedicated style guides that advise their reporters to employ particular language and avoid other. For example, the BBC instructs reporters to use generic wording for the obstacle separating Israel and the Palestinian Territories:

> The BBC uses the term 'barrier', 'separation barrier' or 'West Bank barrier' as an acceptable generic description to avoid the political connotations of 'security fence' (preferred by the Israeli government) or 'apartheid wall' (preferred by the Palestinians) (BBC Academy, 2013).

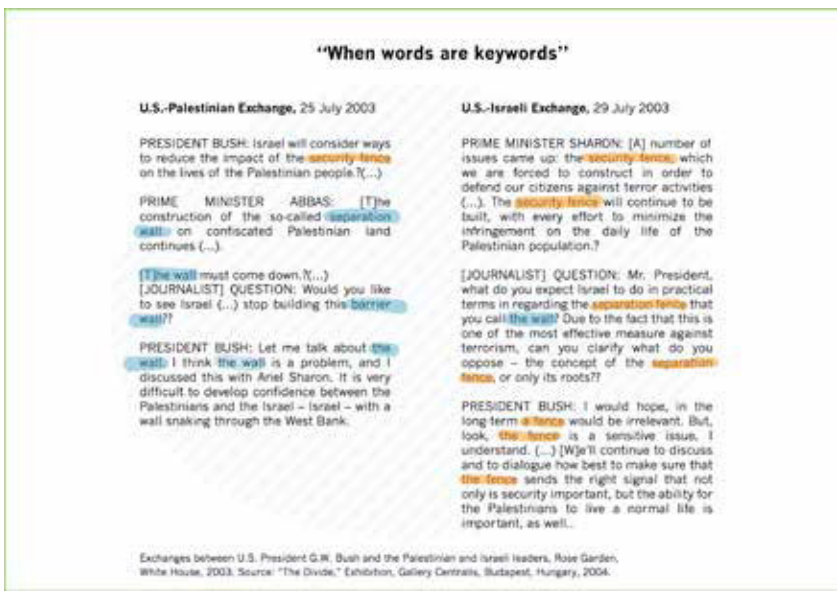
When formulating queries, it is pertinent to consider keywords as being parts of programmes, anti-programmes or efforts at neutrality, as this outlook allows the researcher to study trends, commitments and alignments between actors. To this end (and in contrast to discourse analysis), one does not wish to have equivalents or substitutes for the specific issue language being employed by the programmes, anti-programmes and the neutral programmes. For example, there is a difference between using the term 'blood minerals' or the term 'conflict minerals', or using 'blood diamonds' or 'conflict diamonds', because the terms are employed (and repeated) by particular actors to issuefy, or to make into a social issue forced and often brutal mining practices that fuel war (blood diamonds or minerals) or to have industry recognize a sensitive issue and their corporate social responsibility (conflict diamonds or minerals). Therefore, they should not be treated as equivalent and grouped together. (Here it is useful to return to the point that one should use quotation marks around keywords when querying, because without quotation marks and thus specific key word queries, Google returns equivalents.) Indeed, one should treat 'conflict minerals' and 'blood minerals' as separate, because as parts of specific programmes they show distinctive commitments and they can help to draw alignments. If someone (often a journalist) begins using a third term, such as 'conflict resources', it likely constitutes a conscious effort at being neutral and not joining the programmes using the other terms. Those who then enter the fray and knowledgeably employ *what have become keywords* (in Williams's second sense) can be said to be taking up a position and a side, or avoiding one.

To demonstrate the notion of programmes, anti-programmes and efforts at neutrality further, the Palestinian-Israeli conflict, alluded to above, presents a compelling case for studying positioning as well as (temporary) alignment. There are two famous, recorded exchanges that took place at the White House in the US between then President George W. Bush and the leader of the Palestinian Authority, Mahmoud Abbas; and, secondly, between President Bush and the then Prime Minister of Israel, Ariel Sharon (see Figure 5.4). These exchanges, from the time when the barrier was under construction, show the kinds of positioning efforts that are made through the use of particular terms and thus the kind of specific terminology that one should be aware of when formulating queries. They also reveal temporary alignments that put on display diplomacy, with the US President using the Palestinian and then the Israeli preferred terminology in the company of the respective leaders, but only partly, thereby never fully taking sides.

The first exchange between President Bush and the Palestinian leader, Abbas, begins with a discussion in which Bush refers to the barrier as a 'security fence', which is the official Israeli term. Abbas then makes an attempt to correct this keyword by replying with the term 'separation wall', thereby using a very different adjective – separation instead of security – to allude to the interpretation of the purpose of the barrier as separating peoples and not securing Israel. Abbas also uses a poignant noun, wall. The word 'fence', as in the Israeli 'security fence', connotes a lightweight, neighbourly fence. By calling it a 'wall', however, Abbas connotes the Berlin Wall. The third person in this exchange, the journalist, then steps in with the term 'barrier wall' in an effort not to take sides, though at the moment 'wall' actually gives the Palestinian position some weight. Following this exchange, Bush, being diplomatic, realizes when talking to Abbas that the word 'wall' is being used, so he switches terms and concludes by using the term, albeit without an adjective that would validate Abbas and clash with the official Israeli term.

Four days later, the Israeli Prime Minister, Sharon, visits the White House to talk to President Bush, and he begins by using 'security fence', the official Israeli term. A journalist steps in and seems not to have read any newspaper

Fig. 5.4: The use of keywords by US, Palestinian and Israeli leaders, showing (temporary) terminological alignments and diplomacy. Exchanges between the leaders at the Rose Garden, US White House, 2003.



style guides on the matter, because he first says 'separation fence' and then 'wall'. The journalist, moreover, does not use 'security fence' and, therefore, the question he poses, whilst critical, also seems one-sided for it was preceded by quite some Palestinian language (separation, wall). Bush concludes by being diplomatic once again to both parties involved: he is tactful to Sharon by just using the word 'fence', but he does not use any adjective so as to be wary of Abbas, his recent visitor.

Wall and fence talk in the Middle East, of course, is very specific conflict terminology, but it does highlight a particular programme ('security fence'), an anti-programme ('separation wall') as well as an effort at being neutral ('barrier wall'). It also shows how temporary alignments, often only partial ones, are made with great tact, providing something of a performative definition of diplomacy.

Issue spaces can be analysed with this sort of keyword specificity in mind. A related example in this regard concerns the United Nations (UN) Security Council's debates on the barrier between Israel and the Palestinian Territories, which took place in 2003 and 2005 when it was first being constructed (Rogers & Ben-David 2010). The terms used by each country participating in the debates were lifted directly from the Security Council transcripts. The resultant issue maps, or network graphs, contain nodes that represent countries, clustered by the term(s) that each country uses when referring to the barrier (see Figures 5.5 and 5.6). The network clearly demonstrates the specificity of the terminology put into play by the respective countries at the table as well as the terminological alignments that emerge. When countries utter the same term, groupings or blocs form, to speak in the language of international relations. For example, the largest surrounds 'separation wall', and mention of other terms ('expansionist wall', 'racist wall', 'security wall', 'the barrier', 'the fence', 'the wall', 'the structure', 'separation barrier', and so forth) make for smaller groupings or even isolation.

In 2003 a majority of countries came to terms around 'separation wall' or 'the wall', both Palestinian side-taking terms, and there was a smattering of more extreme terms, e.g. 'racist wall'. On the other side of the divide, the term 'security fence', the official Israeli nomenclature, is only spoken by Israel and Germany, showing terminological alignment between the two countries. Two years later, in 2005, the next UN Security Council debate on the barrier took place, and a similar pattern of terminology use emerged, albeit with two distinct differences. Neutral language had found its way into the debate, with 'the barrier' enjoying support. And this time, Israel was alone in using the term 'security fence', and is thereby isolated.

Fig. 5.5: Cluster graph showing co-occurring country uses of terminology for the structure between Israel and the Palestinian Territories, UN Security Council meeting, 2003. Visualization by ReseauLu.

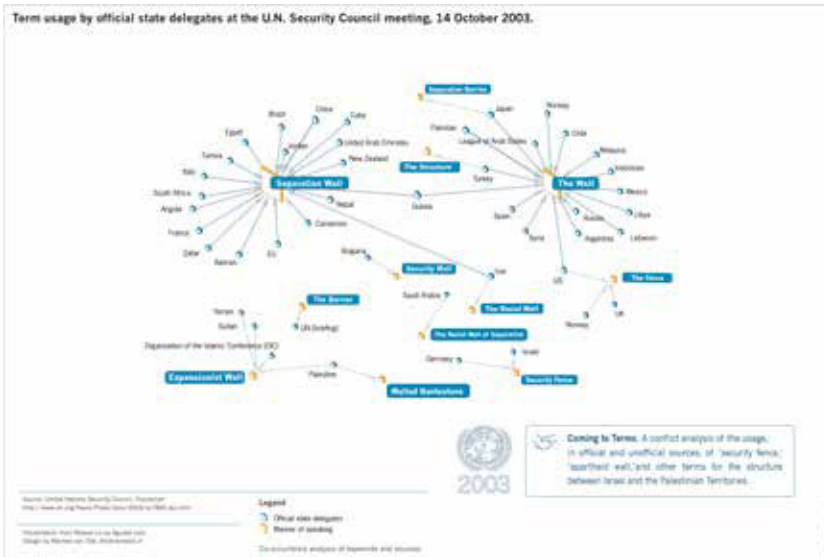
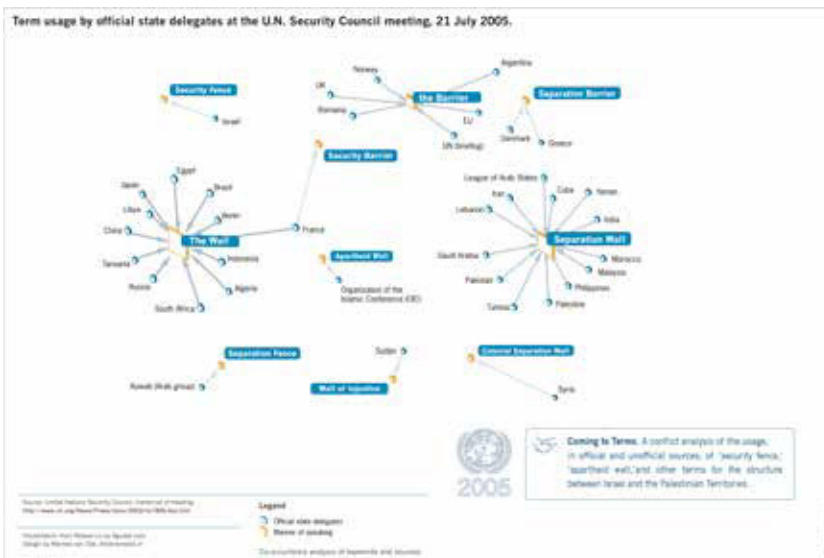


Fig. 5.6: Cluster graph showing co-occurring country uses of terminology for thestructurebetweenIsraelandthePalestinianTerritories,UNSecurityCouncil meeting, 2005. Visualization by ReseauLu.



Countries are 'linked' or isolated by terminology. They settle into a debate by subscribing to programmes, anti-programmes and efforts at neutrality, together with light gestures towards the one side or another (e.g. by using just wall or fence). In some cases, there are evident language blocs. Each bloc shows alignment in that countries (over time) come to terms with other countries by means of using the same language. It is precisely this alignment of actors to programmes, anti-programmes or efforts at neutrality that one seeks to build into query design from the outset.

Unambiguous and Ambiguous Queries

If you peruse the search engine literature, there are mentions of navigational queries, transactional queries and substantive queries, among other types. Yet, on a meta-level, we can broadly speak of two kinds of queries: unambiguous and ambiguous. The original strength of Google and its PageRank algorithms lay in how they dealt with an ambiguous query that matches more than one potential result and thereby is in need of some form of 'disambiguation'. An example that was often used in the early search engine literature is for the query 'Harvard'. This could refer to the university, a city (in Illinois, USA) or perhaps businesses near the university or in the city. By looking at which sites receive the most links from the most influential sites, PageRank would return Harvard University as the top result because it would presumably receive more links from reputable sources than a dry-cleaning business near the university, for example, called Harvard Cleaners. Therefore, without unambiguous matching of keyword to result, the outputs depend on a disambiguating mechanism (Google's PageRank) that places Harvard University at the top. The ability to disambiguate is also thereby socio-epistemological or one that reveals social hierarchies. Harvard University is at the top because it has been placed there through establishment linking practices.

The social researcher may take advantage of how the search engine treats ambiguous queries. In the example, the ambiguous keyword, 'rights', is queried in a variety of local domain Googles (e.g. google.co.jp, google.co.uk etc.), in order to create hierarchies of concerns (rights types) per country, thereby employing Google as a socio-epistemological machine.

Contrariwise, an unambiguous query is one in which it is clear which results one is after. If we return to the cluster maps of countries using particular terms for the barrier between Israel and the Palestinian Territories, precise terms were used. By putting these terms in quotation marks and

querying them, Google would return an ordered list of sources that use those specific terms. If one forgoes the use of quotation marks in the query, Google, as mentioned, 'helpfully' provides the engine user with synonyms or equivalents of sorts. For example, if one does not wish to make a distinction between mobile phones (British English) and cell phones (North American English), you can simply search for [mobile phones] without quotation marks and Google will furnish results for both of them. If one places a term in quotation marks, however, Google will provide results specific to that one term.

It is instructive to point out a particular form of annotation when writing about queries. When noting down the specific query used, the recommendation is to use square brackets as markers. Therefore, a query could be ["apartheid wall"], where the query has square brackets around it and the query is made as unambiguous as possible (for the engine) by using quotation marks. Oftentimes, when a query is mentioned in the literature, it will have only quotation marks without the square brackets. A reader is often left wondering whether the query was in fact made with quotation marks or whether the quotation marks are used in the text merely to distinguish the term as a query. To solve this problem, the square brackets annotation is employed. If one's query does not have quotation marks they are dropped but the square brackets remain.

Doing Search as Research

There are two preparatory steps to take prior to doing search as research. The first one is to install a research browser. This means installing a separate instance of your browser, such as Firefox, or creating a new profile in which you have cleaned the cookies and otherwise disentangled yourself from Google. The second preparatory step is to take a moment to set up one's Google result settings. If saving results for further scrutiny later (including manual interpretation as in the Rights Types project discussed below), set the results from the default 10 to 20, 50 or 100. If one is interested in researching a societal concern, one should set geography in Google to the national level – that is, to the country level setting and not to the default city setting. If one is interested in universal results only, consider obfuscating one's location. In all cases one is not logged into Google.[4]

4 It is also important to note that simply using private browsing tools, such as the incognito tool on Google Chrome, does not suffice as a disentanglement strategy, as this only prevents the saving of one's search history to one's own machine. It is still being saved at headquarters so to speak. When in incognito mode, one is still served personalized results.

I would like to present, first, an example of research conducted using unambiguous queries. The project in question concerns the Google image results of the query for two different terms for the same barrier: ["apartheid wall"], which is the official Palestinian term for the Israeli-Palestinian barrier mentioned previously, versus the Israeli term, ["security fence"] (see Figure 5.7). The results from these two queries present images of objects distinctive from one another. The image results for ["apartheid wall"] contain graffitied, wall-like structures, barbed wire, protests, and people being somehow excluded, whereas with ["security fence"] there is another narrative, one derived through lightweight, high-tech structures. Furthermore, there is a series of images of bomb attacks in Israel, presented as justification for the building of the wall. There are also information graphics, presenting such figures as the number of attempted bombings and the number of bombings that met their targets before and after the building of the wall. In the image results we are thus presented with the argumentation behind the building of the fence. The two narratives resulting from the two separate queries are evidently at odds, and these are the sorts of findings one is able to tease out with a query design in the programme/anti-programme vein. Adding neutral terminology to the query design would enrich the findings by showing, for example, which side's images (so to speak) have become the neutral ones.

When doing search as research as above, the question is often raised whether to remove Google artefacts and Google properties in the results, and under which circumstances. Wikipedia, towards the top of the results for substantive queries, is ranked highly in the results for the query ["apartheid wall"] yet has as the title of its article in the English-language version an effort at neutrality in 'West Bank barrier', however much it includes a discussion of the various names given to it. Whilst a Google artefact, Wikipedia's efforts at neutrality should be highlighted as such rather than removed. A more difficult case relates to a Google artefact in the results for an ambiguous query [rights] in google.com, discussed in more detail below. The R.I.G.H.T.S. organization is returned highly in the results, owing more to its name than to its significance in the rights issue space. Here again the result was retained, and footnoted (or highlighted) as a Google artefact, which in a sense answers questions regarding the extent or breadth of artefacts in the findings. Here the research strategy is chosen to highlight rather than remove an artefact, so as to anticipate critique and make known media effects.

As the last example, I would like to present a project using an ambiguous query that takes advantage of Google's social sorting. In this

Fig. 5.7: Contrasting images for ["Apartheid Wall"] and ["Security Fence"] in Google Images query results, July 2005.

case we undertook a project about rights, conducted by a large group of researchers who spoke some 30 languages amongst them. Using this abundance of diverse language skill, we set about to determine which sorts of rights are held dear to particular cultures relative to others. In the local languages we formulated the query for [rights], and we ran the query in all the various local domain Googles per language spoken, interpreting the results from google.se as Swedish concerns, .fi for Finnish, .ee for Estonian, .lv for Latvian, .co.uk for British, and so forth. With the results pages saved as HTML (for others to check), the researchers were instructed to work with an editorial process where they manually extracted the first 10 unique rights from the search results of each local domain Google.[5] Information designers visualized the results by creating an icon for each right type and a colour scheme whereby unique rights and shared rights across the languages were differentiated. The resultant

5 According to Google's terms of service, one is not allowed to save results, or make derivative works from them. The research thus could be considered to break the terms of service, however much the spirit of those terms is to prevent commercial gain through redistribution rather than to thwart academic research. The results pages are saved as HTML, with a uniform naming convention so that one could return to them, and they, in recognition of the terms of service, were not shared to a data repository.

Fig. 5.8: Rights types in particular countries, ranked from Google results of the query [rights] in the local languages and local domain name Googles (Google.se, Google.fi, Google.ee and Google.it), July 2009.

infographic graphically shows rights hierarchies per country as well as those rights that are unique to a country and those shared amongst two or more countries. One example of a unique right is the case of Finland, in which the 'freedom to roam' is high on the list (see Figure 5.8). Far from being a trivial issue, what this freedom means is that one can walk through someone's backyard, whereas in other countries (e.g. the UK) it is not a right, and organizations are lobbying for the right to ramble and walk the ancient pathways. Another example is in Latvia, where pension rights for non-citizens are of particular importance.

Conclusions

Digital methods have been developed as a distinctive strategy for internet-related research where the Web is considered an object of study for more than online or digital culture only. As a part of the computational turn in social research, digital methods were developed as a counterpart to virtual methods, or the importation of the social scientific instrumentarium into the Web, such as online surveys. Digital methods, as an alternative, strive to employ the methods of the medium, imagining the research affordances

of engines and platforms, and repurposing their methods and outputs for social (and medium) research.

The contribution here is foundational is the sense of outlining certain premises of digital methods but also the nitty-gritty of doing online analysis. In conclusion, I would like to return to the premises of doing digital methods with Google Web Search in particular as well as to the finer points of query design, which underpins 'search as research' as an approach distinctive from other analytical traditions, such as discourse and content analysis.

First, in the digital method, search as research, Google is repurposed from its increasing use as a consumer information appliance, with personalized results that evermore seek to anticipate consumer information needs (such as with autosuggest as well as the Google Instant service). Rather, Google is relied upon as an epistemological machine, yielding source hierarchies and dominant voice studies (through its ranked results for a keyword query) as well as individual actor commitment (through its quantitative counts for a single or multiple site query). Transforming Google back into a research machine (as its founders asserted in the early papers on its algorithms) these days requires disentangling oneself from the engine through the installation of a clean research browser and logging out. Once in use, the research browser is not expected to remove all Google artefacts from the output (e.g. Google properties, SEO'd results, etc.), but in the event they become less obfuscated and an object of further scrutiny (medium research) together with the social research one is undertaking with repurposed online methods.

Query design is the practice behind search as research. One formulates queries whose results will allow for the study of trends, dominant voice, positioning, commitment, concern and alignment. The technique is sensitive to keywords, which are understood as the connections people are currently making of a word or phrase, whether established or neologistic, leaning on Raymond Williams's second definition of a keyword. Indeed, in the query design put forward above, the keywords used could be said to take sides, and are furthermore conceptualized as forming part of a programme or anti-programme, as developed by Madeleine Akrich and Bruno Latour. I have added a third means by which keywords are put into play. Journalists, and others conspicuously not taking sides, develop and employ terms as efforts at neutrality. ["West Bank barrier"] is one term preferred by BBC journalists (and the English-language Wikipedia) over ["security fence"] (Israeli) or ["apartheid wall"]. Querying a set of sources (e.g. country speeches at the UN Security Council debates) for each of the terms and noting use as well as common use (co-occurrence) would show positioning and alignment, respectively.

Secondly, for digital methods practice, I would like to emphasize that for query design in the conceptual framework of programme/anti-programme/efforts at neutrality, one retains the specific language (instead of grouping terms together), because the exact matches are likely to show alignment and non-alignment. Furthermore, language may also change over time. Therefore, if one conducts an overtime analysis, one can determine whether or not certain actors have, for example, left a certain programme and joined an anti-programme by changing the language and terms they use. Some countries may have become neutral, as was noted when contrasting term use in the 2003 versus the 2005 Security Council debates on the barrier. As another example, one could ask, has there been an alignment shift signified through actors leaving the 'blood minerals' programme and joining the 'conflict minerals' programme?

Thirdly, whilst the discussion has focused mainly on unambiguous queries, search as research also may take advantage of ambiguous ones. As has been noted, if we are interested in researching dominant voice, commitment and showing alignment and non-alignment, an unambiguous query is in order. Through an ambiguous query, such as [rights], one can tease out differences and distinct hierarchies of societal concerns across cultures. Here a cross-cultural approach is taken which for search as research with Google implies a comparison of the results of the same query (albeit in each of the native languages) of local domain Google results.

Finally, query design may be viewed as an alternative to forms of discourse and content analysis that construct labelled category bins and toss keywords (and associated items) into them. That is, in query design specificity of the language matters for it differentiates as opposed to groups. More generally, it allows one to cast an eye onto the entire data set, making as a part of the analysis so-called long tail entities that previously would not have made the threshold. One studies it all without categorizing and without sampling, which (following Akrich and Latour) allows not only for the actors to speak for themselves and for the purposes of their programme, anti-programme or efforts at neutrality, but (following Lev Manovich's Cultural Analytics) provides opportunities for new interpretive strategies. That there arises a new hermeneutics (one that combines close and distant reading) could also be seen as the work ahead for the analytical approach.[6]

6 At the lecture delivered at the digital methods Winter School, January 2015, Lev Manovich proposed work on a 'new hermeneutics' after the study and visualization of 'all data', substituting continuous change for periodization and continuous description for categorization.

Acknowledgments

The author would like to thank Anat Ben-David for her work on query design, and Becky Cachia for editorial assistance. Michael Stevenson, Erik Borra and researchers at the Digital Methods Initiative, University of Amsterdam, provided crucial input.

References

Akrich, Madeleine & Bruno Latour. 1992. "The De-Scription of Technical Objects." *Shaping Technology / Building Society: Studies in Sociotechnical Change*, ed. Wiebe Bijker & John Law, 205-224. Cambridge, MA: MIT Press.

BBC Academy. 2013. "Israel and the Palestinians." *Journalism Subject Guide*. London: BBC, www.bbc.co.uk/academy/journalism/article/art20130702112133696.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant. 2009. "Detecting influenza epidemics using search engine query data." *Nature* 457: 1012-1014.

Hindman, Matthew. 2008, *The Myth of Digital Democracy*. Princeton: Princeton University Press.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data." *Science* 343 (6176): 1203-1205.

Pariser, Eli. 2011. *The Filter Bubble*. New York: Penguin Press.

Negroponte, Nicholas. 1995. *Being Digital*. London: Hodder and Stoughton.

Rogers, Richard & Anat Ben-David. 2010. "Coming to Terms: A conflict analysis of the usage, in official and unofficial sources, of 'security fence,' 'apartheid wall,' and other terms for the structure between Israel and the Palestinian Territories." *Media, Conflict & War* 2 (3): 202-229.

Sunstein, Cass. 2001. *Republic.com*. Princeton: Princeton University Press.

Turow, Joseph. 2006. *Niche Envy*. Cambridge, MA: MIT Press.

US Centers for Disease Control. 2014. "CDC Announces Winner of the 'Predict the Influenza Season Challenge'." Press release, 18 June, www.cdc.gov/flu/news/predict-flu-challenge-winner.htm.

Williams, Raymond. 1975. *Keywords: A Vocabulary of Culture and Society*. London: Fontana.