

Chapter Eight:

Information sources and data discovery

Maureen Henninger

Journalists use many different sources for obtaining information; these include interviews with experts, proprietary news archive databases such as LexisNexis and Factiva, and the Internet. Furthermore, it is acknowledged that there are different types of information requirements; quick factual information—what is the gross domestic product (GDP) of Singapore; general background information—legislation of the gaming industry in Japan; and detailed data sources for investigative and data-driven journalism. This chapter focusses on the use of the Internet for searching for all three of these types of information. Chapter Eight will examine data-driven journalism requirements in detail. First two points of clarification. The Web (originally called the World Wide Web) is not the Internet—it is an Internet application. While most people use both terms synonymously, the Web consists of pages that are written in a special mark-up language, unseen by the user, but are able to be displayed by browsers. Such Web pages I will refer to as the ‘visible’ Web. This point is very important as we shall see, since the Internet consists of a great deal more which is not Web pages, but rather an invisible, or hidden Web.

Material ‘on the Web’ can be of many types and formats including;

- textual material (reports, journal and newspaper articles, and Web pages themselves)
- graphic material (maps, photographs and charts)
- moving images (videos)
- sound material (for example podcasts).

On top of all this is the content held within social media spaces like Twitter, Facebook, and YouTube—user-created content which can be any of the above formats. In this chapter I will consider all these materials to be information sources and will refer to them as ‘documents’.

Since the early 1990s when the Internet became commercialized, driven by the development of the Web and its concomitant development of hypertext browsers, the amount of material has become vast. In fact one Web blogger wondered if we had reached a world of infinite information (Bloch, 2011). There is no debate about this vastness—one study estimated that at the end of January 2013 there were at least 12.86 billion *indexed* pages¹ and conceivably there are millions more Web documents that have not been indexed by the search engines.

¹ See www.worldwidewebsize.com/ viewed 29January 2013

This chapter covers the general principles of finding information and then explores the tools and techniques for finding firstly, information that has been indexed by search engines—the visible Web, and, secondly, tools for material which has not—the invisible Web.

Information discovery and search

Katz and Lazarsfeld in the 1950s hypothesised that communication was a two-step flow; that ideas often flow from radio and print to opinion leaders who in turn pass them on to every day associates for whom they are influential (Katz and Lazarsfeld, 1955). Donald Case suggested that this information and ideas, traditionally communicated in this two-step flow, is now found easily through the Internet (Case et al., 2004). The idea of ‘ease’ is very much the principle that dictates all human information-seeking behaviour; we generally take the easiest way—what George Zipf referred to as the principle of least effort (Zipf, 1949)—this along with the current attitude of instant gratification, makes the easiest way to ask someone. This, of course, is one of the main ways that journalists work—they find an expert to interview. If you do not know an expert the next ‘easiest’ thing, in fact the almost automatic strategy is to go to the Internet and use a search engine. So automatic is this behaviour that the phrase “to Google it” has become shorthand for searching online; and I often hear searchers say “Google will tell me”. The next section will examine the search process and the impact of search engines on journalistic practice.

How searchers search

According to ComScore search and navigation is the fifth highest Internet usage category world-wide, after portals, entertainment, community, and news/information (ComScore, 2012), and in their monthly statistics for the United States Google consistently dominates the search engine market followed distantly by Bing and Yahoo.² The global trend is similar with statistics for November 2012 having Google leading (90.75 per cent), Bing (3.32 per cent), Yahoo (2.84 per cent) and Baidu (0.58 per cent).³ However there is some research to show that in Asia and the Pacific, searchers tend to rely on local search engines in order to find material that has non-English content. For example Baidu (China) and Naver (South Korea) each has over 60 per cent of the search engine market.

² See ComScore at www.comscore.com

³ See StatCounter Global Stats at <http://gs.statcounter.com/>

If the use of Internet search engines is so wide spread among the general public,

To find a local search engine, use the international search engine directory, Search Engine Colossus at www.searchenginecolossus.com

then it can be assumed that it would also be the same among journalists.

Diekerhof and Bakker (2012) point out that the majority of journalists use the Internet to verify and check facts when it is easy to do so, particularly as most

journalists appear to trust online sources (see Messner and Distaso, 2008, Carlson, 2009, Abdulla et al., 2005). Other studies of journalists' use of search engines bear this out, and that Google is the dominant choice; for example Machill and Beiler (2009) report German journalists use Google 90.4 per cent of the time when carrying out Internet research. As most of the search engines work in similar way (although their algorithms, which are proprietary, may be different and are constantly being adjusted and modified), I will use the big global ones, Google and Bing in most of the examples of searching techniques.

With the colossal number of documents available on the Web and because of the propensity of searchers to use simple queries—most studies show that searchers use only one or two keywords in the search bar—there is the likelihood of retrieving millions of documents per search. This leads to enormous information overload, not to mention the possibility that many of the results are not relevant to the search.

All search engines provide a simple search bar into which the searcher types some keywords, although most search engines do have advanced search screens, it is assumed the average user does not want to use advanced search features or create complex queries (Griffiths and Brophy, 2005, Park et al., 2005, Jansen and Spink, 2006). When journalists use the Internet for research, it was reported they tended to carry out sophisticated searches rarely and not always correctly (Machill and Beiler, 2009). Furthermore this study showed that journalists only rarely flick past the first page of *Google* results.

How search engines work

Search engines, at their most simple level, use 'exact match' retrieval—the words in the search query are present in the index created from all the documents that are found as the search engines crawl the Web. For example a simple search for two words, *football* and *health* would deliver results that have both words in the

document, and the more words you use, the fewer the number of results are returned (see Table 1).

Table 1 Search results from Google (January 2013)

Search query	Results
football health	1,310,000,000
football health benefits	213,000,000
football "health benefits"	3,410,000
intitle:football intitle:"health benefits"	1,360

As can be seen in Table 1 you can make your query more precise by indicating a semantic relationship between words. This is typically done by indicating that words are phrases by using double quotes, for example "health benefits". Most search engines have algorithms that attempt to do this automatically; however the use of double quotes signals you require documents with the phrase to be more highly ranked. Finally, since it can be assumed that words in the title of a document indicate relevancy, that is the words are pertinent to the subject of the document, by putting *intitle:* before the word or phrase can deliver such documents more highly ranked in the results (do not have a space—*intitle:"health benefits"*). However language is very complex and words can have many meanings in different contexts. Here are two examples:

the word 'crop' can mean a plant to be harvested, to remove parts of an image or a whip used in horseback riding;

depending on which country you live football can be soccer, rugby, gridiron, or Australian rules.

Search engines have developed enhanced techniques to retrieve a set of documents 'on topic', for example if your search for the word crop contains also the word market (one of the reasons for including more than one keyword in your search query), the algorithm will deliver documents about plants to be harvested. The results, while about or pertinent to the topic are then ranked in order of relevance, that is, about or pertinent to the topic, but possibly not relevant to the searcher. In the case of the football question, if the searcher is from Brazil, he or she probably wants documents about soccer.

In order to overcome these problems (and to keep the searcher satisfied with the service, the ‘stickiness’ factor that is vital to profitability), search engines have become very sophisticated, attempting to deliver highly relevant and pertinent information to the information seeker, while at the same time delivering profits, based on advertising to their corporate owners. One of the ways to achieve this is by introducing the concept of personalised search.

Personalized search

Other names for personalised search are intent-based search and algorithmic search. Intent-based search is the notion that the search engine can determine the specific intention of the search, that is, the precise piece of information the searcher is looking for; algorithmic search, based on search algorithms that gather data from the searcher’s previous search history, including what has been clicked on (the click stream), and how long the searcher stayed on a retrieved document, in order to deliver *personally* relevant results. Google was the first to introduce personalised search and in 2009 made all search personalised; since then all the other major search engines have followed suite (Bing refers to this as ‘adaptive search’, the concept of adapting to the individual searcher’s intent). The following two examples show how personalised search works.

Suppose you live in India and search on the single word *cricket*; the search engine determines you are probably looking for information about cricket, the sport, since almost everyone in India loves this sport. In fact at the time of writing this chapter a search on www.google.co.in (the Indian version of Google) delivered 368,000,000 documents, and at least the first 50 pages of results were about the sport, not the insect.

This example concerns the search query ‘autocomplete’, where, based on your context, e.g. country, previous search history and popular searches are displayed or suggested. Figure 1 shows two sets of auto complete suggestions for the word *journalism*; on the left is done on Google Australia (www.google.com.au) and on the right is Google Philippines (www.google.com.ph).

Figure 1 Autocomplete suggestions for Google Australia and Google Philippines

journalism	journalism
journalism jobs	journalism
journalism jobs sydney	journalism quotes
journalism courses	journalism terms
journalism internships sydney	journalism online training
journalism courses sydney	
journalism internships	
journalism cadetships	
journalism unsw	
journalism ethics	
journalism at the speed of bytes	

More recently search engines have begun to include social media content (Facebook, Twitter, Flickr, etcetera) in order to further contextualise search results based on the interests of your friends, or, more accurately those persons you are connected to through social media sites.

Finally journalists should be aware of the debates about personalised search, search engine bias, lack of objectivity, and privacy issues that arise from the trend in personalised search.⁴ This debate is not within the objectives of this chapter; rather it concentrates on explaining the functionalities of search engines and provides techniques for producing less personalised and possibly more relevant search results that may be more effective for journalists in certain circumstances. If you are concerned about these issues, here are some ways of circumventing personalised search.

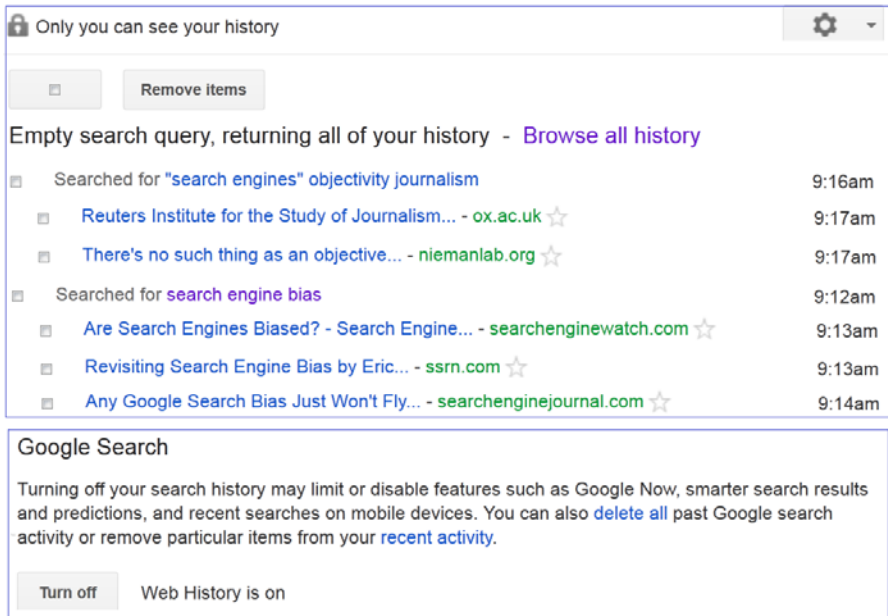
Turn off personalized search; all search engines allow you to do this, Figure 2 below shows how it is done in Google.

⁴ For further discussion concerning this debate see Tavani, H. 2012, 'Search Engines and Ethics', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2012 edn, <<http://plato.stanford.edu/archives/fall2012/entries/ethics-search/>>

Use open source proxy software such as Tor⁵ which prevents traffic analysis, which is the gathering of personal information when using Internet applications, including web browsers. Journalists, for example use it to protect their sources and to allow them to communicate more safely with whistle-blowers and dissidents.

Use DuckDuckGo a search engine which does not personalize search results; “it does not collect or share personal information [search leakage] . . . and prevents search leakage by default.”⁶

Figure 2 Removing items from Google history and permanently turning off history



⁵ Tor can be downloaded from www.torproject.org/

⁶ See <http://duckduckgo.com/privacy>

Even if you use any or all of these techniques to attempt to search for objective, relevant information, it is still important to use some more sophisticated techniques to be a more effective and efficient searcher.

Effective techniques for searching

In order to be an effective searcher, take a little bit of time to think about what you are looking for. Because of the enormous quantity of information available, and the lack of quality of much of it, you should avoid simple keyword searches and think in terms of concepts. For example if you were writing a story on government policy to increase exercise among school children, you might consider the concepts to be, although each of us might translate these concepts into different search terms:

exercise
school children
government policy

While there is no ‘right’ way of conceptualising or doing a search, as we shall see in below there are some very useful ways of entering your concepts and keywords. The major search engines have advanced search forms but are a bit hard to find, since research shows most people do not use them and search algorithms are eliminating the necessity for them. Nevertheless advanced search functionality enables more effective searching, particularly if you are seeking specific and in-depth information for background to an investigative piece of writing. Search engines enable you to use their advanced search functionality in the single search bar as shown below. The following examples show queries which will deliver results that are less likely to be filtered by personalisation and which suggest a more sophisticated approach based on specific concepts.

Tips

- use phrase searching e.g. “shipping lanes”
- use more than one keyword or phrase
- use intitle: to ensure relevance
- for reports use filetype:pdf
- for government or non-government information use inurl:gov or inurl:org

Example one, using Google provides three possible ways of doing the search for government policy to increase exercise among school children, each of which show various keywords/phrases and syntax reflecting differing ways of looking at the information requirements and specificity. Examples two, three and four demonstrate other ways of thinking about search and employ techniques and functionality not only of Google, but also Bing, DuckDuckGo and Exalead.

Example one

This search query uses the concepts above but finds several phrases which reflect the concept of exercise (note in Google and DuckDuckGo you must put OR in upper case):

"school children" "physical exercise" OR "physical fitness" OR "physical activity" OR "physical education" OR sport "government policy"

Narrowing down the results to certain countries and making sure that the documents are about school children—for example by adding Singapore, etcetera requires the name of the country to be somewhere in the document.

intitle:"school children" "physical exercise" OR "physical fitness" OR "physical activity" OR "physical education" OR sport "government policy" Singapore OR "Hong Kong" OR Japan

We can further refine the query using some other techniques:

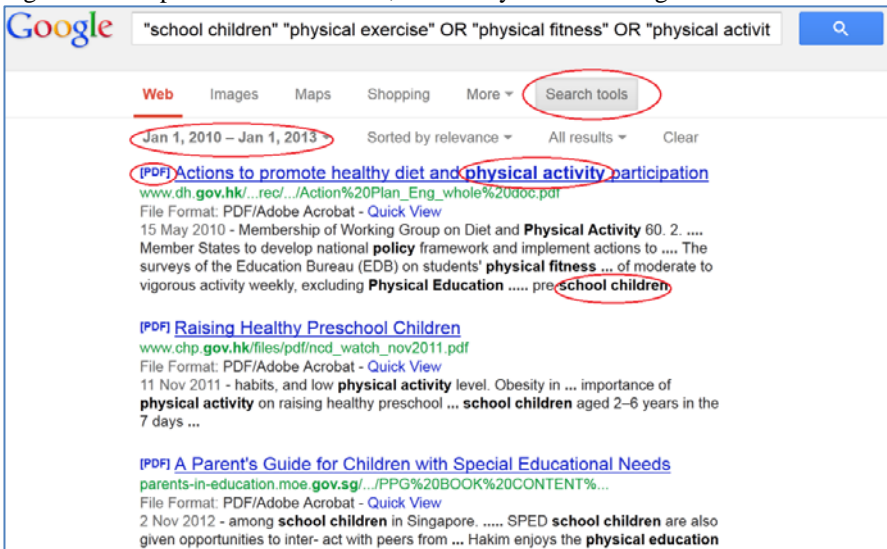
school children" "physical exercise" OR "physical fitness" OR "physical activity" OR "physical education" OR sport ~policy inurl:gov.sg OR inurl:gov.hk OR inurl:gov.jp filetype:pdf

Remove the requirement that the phrase “government policy” is in the document, instead it uses only the word policy and requires the document is from a government website, e.g. inurl:gov.sg (a Singapore government website). Place a tilde (~) directly in front of the word policy, Google searches for both singular and plurals (policy, policies) and for synonyms such as administration (note, only Google has this function).

If one assumes that governments publish documents in pdf format, add the requirement that all the results should be pdf files (filetype:pdf no space before or after the colon).

Filter the results to the past two years using the search tools custom range of dates, as shown below in Figure 3.

Figure 3 Example one search results, filtered by a custom range of dates



Example two

You are doing an investigation on oil pipeline construction in Asia and need some background information and you do not want documents from commercial sites. The following searches use Google and Bing and in each case I have used the currently available search functionality of each search engine.

Google—intitle:oil intitle:pipeline OR intitle:"pipe line" ~construction ~Asia -inurl:com

Bing—(intitle:oil OR intitle:gas) AND (intitle:pipeline OR intitle:"pipe line") AND (construction OR construct OR build) AND (loc:af OR loc:tm OR loc:ru) NOT inurl:com

In Bing you are able to group the ORs—this is called Boolean logic, that is the use of the Boolean operators **AND**, **OR** and **NOT**. George Boole was a mathematician and thus the syntax is the same logic as in mathematics, that is $2 + 2 \times 2 = 6$, whereas $(2 + 2) \times 2 = 8$. I have also used country codes for Afghanistan (af), Turkmenistan (tm) and the Russian Federation (ru)⁷.

Note: Google does not “do Boolean”. Use the minus sign for excluding items.

Example three

Exalead is a search engine that is smaller than Google and Bing, however it has some very powerful functionality and I have included it here to show how you could further create the semantic context using search syntax (and, as we shall see further in the section on the invisible Web, databases often use this functionality). For example if you used the phrase “pipeline construction” conceivably you would not retrieve documents that contain the phrases “the construction of a pipeline” or “to construct pipelines”. To solve this dilemma, Exalead has the facility to use **NEAR**, a proximity operator, which retrieves documents that have two words within ten words of each other. It also allows you to search by date within the query. Thus in a search for documents about arms or weapons shipments to or from Africa or Asia since the 1st of February, 2011, a possible search could be

(arms OR weapon) NEAR (shipment OR shipping OR transport OR transit) NEAR (africa OR asia OR asean) NOT inurl:com
after:2011/02/01

Example four

If you are concerned about personalisation of your search results, try using DuckDuckGo, which does not collect personal information about the searcher;

⁷ The standard Internet country codes can be found at http://www.iso.org/iso/home/standards/country_codes/country_names_and_code_elements.htm

your search history is anonymous. Suppose you were interested in statistical information about HIV/AIDS (acquired immunodeficiency syndrome) among indigenous peoples, particularly the Maoris of New Zealand, and you do not want your search tracked.

(maori OR indigenous) AND (HIV OR AIDS OR "acquired immunodeficiency syndrome") AND (statistics OR statistical) -site:.com -site:.co

In this particular search I have excluded any commercial sites and it should be noted that New Zealand and the United Kingdom use 'co' for commercial domains. Like Google use the minus sign to exclude an item and unlike Google and Bing you need to put the dot in front of the domain code (site:.co).

To sum up - The invisible Web

So far this chapter has concentrated on the visible Web, that part of the Internet which consists of pages written in hypertext mark-up language and which can be displayed by Web browsers. But there is another part of the Internet, in fact which is even bigger than the visible Web and consists of billions of documents that are stored in millions of databases, digital libraries and electronic repositories. In 2001 Bergman estimated that it was "400 to 550 times larger than the commonly defined World Wide Web" and that it contained "nearly 550 billion individual documents" (2001 para 5). This is the hidden, invisible or deep Web because search engines currently are not able to index these pages which are very important sources of information for journalists.

What is in the invisible Web

Much of the invisible Web information is 'grey literature' (documents that have been published but not through any commercial publisher), for example government and non-government reports, briefing papers, technical reports and research papers. While much of this material is found by search engine crawlers (indexing systems) and thus can be listed in search results, to effectively search for with a search engine, you really need to know of its existence and its author or title; and of course, if it does exist it may not necessarily have been found by a search engine. For example a Google search for information on the effects of the global financial crisis on NATO capabilities ("global financial crisis" NATO

capability filetype:pdf) finds about 300,000 documents, many of which no doubt would give you excellent information. But in order to get the one produced by the RAND Corporation you would need to know of its existence and include RAND in your search; at the time of writing the above search did not find the one I required which was written in 2012—*NATO and the Challenges of Austerity*.

The major reason this material often cannot be found and thus not indexed by search engines is that it is stored in databases, hidden behind the database's search interface that search engines, which rely on following hypertext links in Web documents, currently⁸ cannot breach.

When searching the more you know the easier it is to find, IF it exists and has been found by a search engine crawler.

Therefore other tools are required to find the material stored in databases, a two-step process, first to find the databases and then to search within them.

Types of databases

Before examining how to find hidden Web databases we need to look at the types of databases that are available. These fall into several different categories; for our purposes there are three major types:

Bibliographic: this type of database contains the records of published literature such as books, journal articles, reports, conference proceedings, and newspaper articles. The record contains information about the document, such as the information that is shown in the list of references and further readings at the end of this chapter. Sometimes a bibliographic database may also contain the entire text of the documents, in which case it may be called a full-text database. If the 'document' is non-text, such as a photograph or a video, the record is still referred to as a bibliographic record.

Statistical: a statistical database contains numerical data for example population and demographic data, agricultural production data or education statistics.

Factual: this type of database provides direct access to information such as definitions (dictionaries, glossaries), chemical properties, business directories, and postal code directories⁹.

⁸ Search engines are working on this problem, but for the foreseeable future this is not possible

⁹ There is a computational search engine, Wolfram Alpha (www.wolframalpha.com) that often can produce factual, that is statistical information, for example "what is the population of New Guinea"?

Let's go back to the search for information about arms and weapons shipments (Example three above). Once you have begun your investigation you may wish to narrow it to specific data concerning anti-tank missiles shipped to Algeria. This information is not easy to find using Google, but the data is available within a statistical database which you would need to know about and in which you would have to search. So first, you need to be able to find a statistical database which may have this information. Some of the great tools for finding such a database are directories.

Directories

Directories are often referred to as subject directories and are like catalogues in a library where information resources are arranged by subject. Such an arrangement makes them very useful for *browsing* as you might browse the travel sections in a bookstore or a library. Some of the first attempts at organising Web documents used this approach; Yahoo started its life as a subject directory in 1994 and The Virtual Library, started by Sir Tim Berners-Lee in 1991 is still a very valuable information discovery tool.

Directories can be divided into two categories, general ones which include many different subjects and disciplines, and those which specialise in a particular subject or type of resource.

Table 2 Examples of general subject directories

Directory	URL
AcademicInfo Subject Guides	www.academicinfo.net/subject-guides
DADI (European directory of databases)	http://dadi.univ-lyon1.fr
Infomine	http://infomine.ucr.edu
Intute (unfortunately no longer being added to, but still good)	www.intute.ac.uk
IPL2 (Internet Public Library)	www.ipl.org
WWW Virtual Library	http://vlib.org/

Table 3 Examples of some specialised subject directories

Directory	Content	URL
Aerade	Aerospace and defence	http://aerade.cranfield.ac.uk

Complete Planet	Databases	http://aip.completeplanet.com
DocuTicker	Grey literature	www.docuticker.com
Eldis	Development policy and research	www.eldis.org
OFFSTATS	Statistics	www.offstats.auckland.ac.nz
The Guardian's Data Store	Government datasets	www.guardian.co.uk/world-government-data

The next section examines the use of directories to find databases and their content

Use directories for finding databases

There are several strategies for finding information that may be in hidden Web databases. Often this is a two-step process. First finding one or more potential databases and then searching in the database; sometimes as part of the discovery process you need to try more than one directory and more than one database. You also need to examine the Help screens if you have never used the database, since these often have very sophisticated functionality that make your search more efficient and effective. So for the question of arms and weapons supply to Algeria, let us suppose in particular you need data on the supply of 9M133 Kornet/AT-14 anti-tank missiles. Here is a possible strategy:

A search in the subject directory Infomine for **(arm* or weapons) and database*** (Figure 4) retrieves 116 freely available databases containing grey literature and datasets concerning

In many of subject directories and databases you can use a wildcard such as the asterisk (*), for example **arm*** will retrieve arm, arms and armaments; you need to examine the Help details for the available search syntax.

weapons and armaments, including the specialised aerospace and military directory and digital library Aerade and the Rand Corporation International Affaires Digital Library, and the SPIRI Arms Transfers Database which contains the required data (Figures 5-6).

Figure 4 Searching in Infomine for databases

Query: (arm* or weapons) and database*
Expert-selected Resources Found: 116
 Include Computer-Selected Websites
 Include UC Subscription eJournals and eBooks

Modify Search New Search

Titles Display

Result Pages: 1, 2, 3

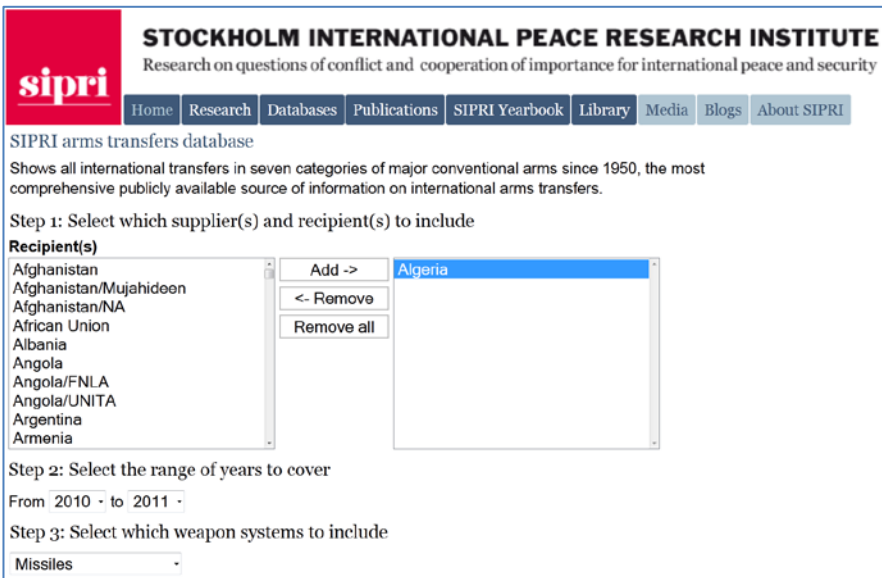
8. [Stockholm International Peace Research Institute](#) Score: 107

 The Stockholm International Peace Research Institute, financed by the Swedish Parliament, conducts "research on questions of conflict and cooperation of importance for international peace and security." From SIPRI's page, you may:

- find out about SIPRI, its staff, and its activities
- use a searchable articles database which covers subjects of peacekeeping, weapons, international relations, etc.
- view country-by-country statistics on arms sales, arms production, and arms transfers
- view statistics on military expenditures worldwide
- view a SIPRI publications catalog

[\[More Info... \]](#) [\[Comment on this resource \]](#)

Figure 5 Searching in the SIPRI database for a specific dataset



STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE
Research on questions of conflict and cooperation of importance for international peace and security

Home Research **Databases** Publications SIPRI Yearbook Library Media Blogs About SIPRI

SIPRI arms transfers database

Shows all international transfers in seven categories of major conventional arms since 1950, the most comprehensive publicly available source of information on international arms transfers.

Step 1: Select which supplier(s) and recipient(s) to include

Recipient(s)

Afghanistan	Add ->	Algeria
Afghanistan/Mujahideen		
Afghanistan/NA	<- Remove	
African Union	Remove all	
Albania		
Angola		
Angola/FNLA		
Angola/UNITA		
Argentina		
Armenia		

Step 2: Select the range of years to cover
From 2010 to 2011

Step 3: Select which weapon systems to include
Missiles

Figure 6 Downloaded dataset from SIPRI

Source: SIPRI Arms Transfers Database
 Information generated: 03 January 2013

Supplier/ recipient (R) or licensor (L)	No. ordered	Weapon designation	Weapon description	Year of order/ licence	Year(s) of deliveries	No. delivered/ produced	Comments
Italy R: Algeria		ASTER-15 SAAM	SAM	(2011)			For BDSL AALS from Italy
Russia R: Algeria	(3000)	9M133 Kornet/AT-14	Anti-tank missile	2005	2006-2010	(3000)	For BMP-2M IFV
	(75)	48N6E2/SA-10E	48N6E2/SA-10E	SAM	2006	2008-2011	(75)
	(40)	53-65	53-65	AS torpedo		2006	(40) For Type-636 (Kilo) submarines
	..	9M131/AT-13 Saxhorn	Anti-tank missile			(2006)	2009-2011 (300)
	(750)	9M311/SA-19 Grison	SA-19 Grison	SAM	(2006)		For Pantsyr-S1 AD systems
	(40)	TEST-71	TEST-71	AS/ASW torpedo		(2006)	2010 40 For Type-636 (Kilo) submarines

From the above example you can see that finding grey literature and datasets can be a two-step process:

Finding an appropriate database in which to search; you may have to try several directories and keep the search broad as you are not searching for the specific report or dataset, but for a database that may contain the information required. Searching in the database using advanced searching techniques; this requires examining the interface for the various options and reading the Help to see the appropriate search syntax.

Here are some further examples to give you some ideas and strategies. You should explore these examples for finding and searching in databases for highly specific information; you will find them invaluable for investigative and data journalism.

Example One


You are writing a story about water resources and one of the angles you want to address is conflicts about water. As background you would like to find a list of violent protests, skirmishes or wars have taken place in the past ten years in Asia. Select, for example, the general directory AcademicInfo Subject Guides and browse through listings for water.

There are four possibilities—in this directory it is often helpful to select a ‘digital library’, in this case, Water Resources Digital Library. Browse through the listings and you will find the Water Conflict Chronology compiled by Peter Gleick, Pacific Institute for Studies in Development, Environment, and Security.

This database gives several ways of accessing the information and enables you to filter by region, date and type of conflict.

Figure 7 shows the partial results and helpfully provides a link to the source by which you can get more information.

Figure 7 Partial results for conflicts over water in Asia

Water Conflict Chronology List 

Region: Conflict Type: Date Range:

Search Showing 24 entries from 2000 to 2010.

Date	Parties Involved	Basis of Conflict	Violent Conflict or In the Context of Violence?	Description	Sources
2000	Central Asia: Kyrgyzstan, Kazakhstan, Uzbekistan	Development dispute	No	Kyrgyzstan cuts off water to Kazakhstan until coal is delivered; Uzbekistan cuts off water to Kazakhstan for non-payment of debt.	Pannier 2000
2000	India: Gujarat	Development dispute	Yes	Water riots reported in some areas of Gujarat to protest against authority's failure to arrange adequate supply of tanker water. Police are reported to have shot into a crowd at Falla village near Jamnagar, resulting in the death of three and injuries to 20 following protests against the diversion of water from the Kanakavati dam to Jamnagar town.	FTGWR 2000
2000	China	Development	Yes	Civil unrest erupted over use and allocation of water from Baiyangdian Lake - the largest natural lake in northern China. Several people died in riots by villagers in July 2000 in Shandong after officials cut off water supplies. In August 2000,	Pottinger 2000

Example Two

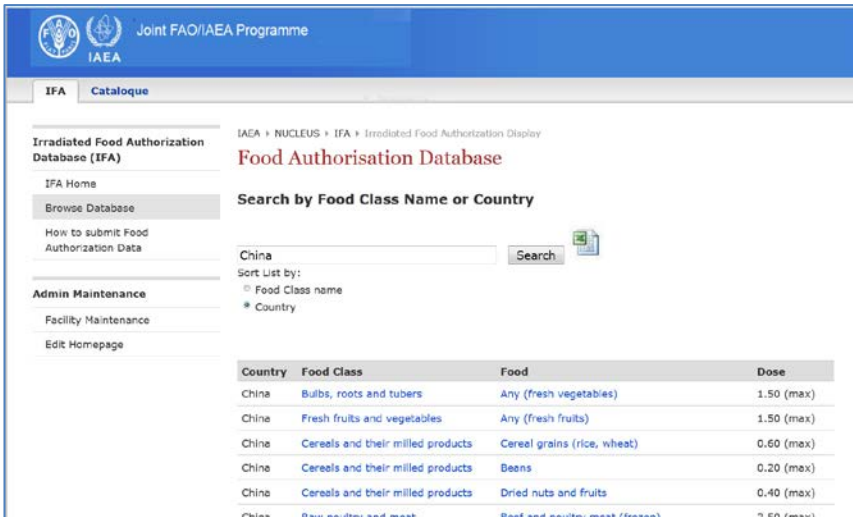
You have an assignment for a story on the safety issues of food irradiation and you would like to examples. For instance what is approved level of irradiation dosages of garlic for Thailand, China and Viet Nam?

Select Intute which has an advanced search that allows you to filter your query by general subject area, which in this case is food, and by the type of resource, data. Enter irradiation as a keyword.

Select the filter for the subject area Agriculture, Food and Forestry and for three of the resource types, Datasets, Non-bibliographic databases, and Statistics.

One of the results is the Food Irradiation Clearances Database that is maintained jointly by the International Atomic Energy Agency (IAEA) and Food and Agriculture Organisation (FAO). Unfortunately the link is no longer correct. However you now know the name of the organisation that produces the database so use Google to find it—“IAEA databases”. The NUCLEUS – IAEA portal shows that the database has a new name, the Irradiated Food Authorization (IFA) Database.

Figure 8 shows the results of the maximum irradiation doses allowed in China. Figure 8 Chinese approved food irradiation dosages (IAEA)



The screenshot shows the 'Food Authorisation Database' interface. It includes a search bar with 'China' entered, a 'Search' button, and a table of results. The table has columns for Country, Food Class, Food, and Dose. The results are filtered for China and show various food classes and their corresponding maximum irradiation doses.

Country	Food Class	Food	Dose
China	Bulbs, roots and tubers	Any (fresh vegetables)	1.50 (max)
China	Fresh fruits and vegetables	Any (fresh fruits)	1.50 (max)
China	Cereals and their milled products	Cereal grains (rice, wheat)	0.60 (max)
China	Cereals and their milled products	Beans	0.20 (max)
China	Cereals and their milled products	Dried nuts and fruits	0.40 (max)
China	Raw poultry and meat	Roof and poultry meat (frozen)	2.50 (max)

Example three

Journalists are very familiar with services and tools which alert them to news and sources such as Twitter, RSS feeds. Another you should be familiar with is DocuTicker a directory of newly published government and non-government grey literature that provides a daily alerting service via Twitter or RSS. The directory can also be browsed and searched.

For example, if you were to be doing a story on human trafficking and wanted an up-to-date report, a quick search on Google, (“human trafficking” report filetype:pdf) found many documents, but not one of the latest (at time of writing January 2012) *Trafficking in Persons: International Dimensions and Foreign Policy Issues for Congress*. However DocuTicker contained it.

To sum up;

Many of the documents and resources available online are hidden away in invisible Web databases, the contents of which cannot be found and indexed by search engines.

Finding this material is often a two-step process; to find an appropriate database and then to search in it.

It is useful to use subject directories, similar to library catalogues, to discover databases.

It is important to read the Help screen of an unfamiliar database since databases often have very sophisticated functionality to enable effective searching.

Statistics and datasets

Statistical information is generally available as analytical and/or formatted reports, which are secondary sources (grey literature), or as raw data often called datasets and are the primary sources from which the analytical reports are written. In most cases the analytical reports do not contain all the detailed data that is available and in many cases journalists wishing to do their own analysis need to be able to access the raw data. Almost all organisations, including governments and non-government organisations such as the World Bank and the United Nations create statistics. Figure 9 is an example of an organisation’s availability of primary and secondary source material; in this case the United States Statistical Abstract. Figure 9 The US Statistical Abstract—raw data and analytical reports

The screenshot shows the 'The 2012 Statistical Abstract' website. At the top, there are navigation tabs: 'Abstract Main', 'Overview', 'PDF Version', 'Earlier Editions', and 'Order'. Below these is a 'BROWSE SECTIONS:' menu with various categories like 'Accommodation, Food, & Other Services', 'Agriculture', 'Arts, Recreation, & Travel', etc. The main content area is titled 'International Statistics: Natural Resources and Energy' and lists several data items with links to Excel and PDF versions. For example, '1379 - World Production of Major Mineral Commodities [Excel 51k] | [PDF 58k]' and '1380 - World Primary Energy Production By Region And Type [Excel 48k] | [PDF 63k]'.

Datasets are presented in a tabular format such a CSV (comma/character separated variables). If you think of a spread-sheet, the values in each cell (or database field) are separated by any character, but most often a comma or a tab. This is an open, albeit not well-defined format, that can be imported into many proprietary software packages including Excel[®]; in many cases, such as those shown in Figure 9 above, the datasets have been converted to Excel spread-sheets. While we will cover the use of datasets in detail in Chapter 8, however it is important that we examine the various methods for finding datasets/

As in the discovery of most information, if you know who has produced the data the easier it is to find; Table 4 gives some examples of major government and non-government official statistical resources

Table 4 Examples of government and non-government statistical resources

Organisation	URL
National governments	
Australian Bureau of Statistics (Australian national statistics)	www.abs.gov.au
Japan e-Stat Portal (Japanese national statistics)	www.nstac.go.jp/en/
US Statistical Abstract (US national statistics, includes some international comparative statistics)	www.census.gov/compendia/statab/

Non-government organisations	
Asian Development Bank (key economic and financial indicators for Asia and the Pacific region)	https://sdfs.adb.org/
OECD (The Organisation for Economic Co-operation and Development) key indicators and other subject statistics for member countries)	www.oecd.org/statistics/
United Nations (economic and social statistics of member countries)	http://unstats.un.org/
World Bank (over 8,000 development indicators)	http://data.worldbank.org

Since the early 2000s there has been great momentum to make statistics, in the form of datasets freely available for public reuse; it this availability that has driven the trend of data-driven journalism. Open Government Data (OGD) initiatives are being introduced in many democratic countries, and led by Sir Tim-Berners-Lee as part of a World Wide Web Foundation grant to promote open data in developing countries.

While many, but not all of the open government data sources use a standard format for the URL—data.gov.state.country, for example

data.nsw.gov.au (New South Wales, Australia), and
data.gov.sg (Singapore),

some catalogues or directories of open government data are currently being developed. As well non-government organisations such as the World Bank also provide dataset catalogues (Table 5).

Table 5 Examples of dataset catalogues/directories

Data services	
datacatalogs.org	http://datacatalogs.org/
Guardian Data store (national, state and local government	www.guardian.co.uk/world-government-data

datasets from around the world)	
OFFSTATS (official statistics portal, by subject, country and region)	www.offstats.auckland.ac.nz
World Bank Data Catalog	http://data.worldbank.org/data-catalog

Reuse of data

It is important to remember that all datasets are the intellectual property of the organisation which creates and/or owns them. As with the reuse of any resource you need to check the conditions under which you may do so; at the very least you must always give attribution of the source (as I have done so with any quote I have used in this chapter). In general any online source, including datasets that may be reused fall into one of several public copyright licencing systems, the most widely used being the Creative Commons (CC) licences. These licences allow content to be “be copied, distributed, edited, remixed, and built upon, all within the boundaries of copyright law” (Creative Commons, n.d. para 13) The most common ones are described in Table6.¹⁰

Table 6 The most widely used Creative Commons (CC) licences

Licence	Conditions
Attribution	The broadest licence—lets others distribute, remix, tweak, and build upon original work, even commercially, as long as they credit the creator for the original creation
Attribution-NoDerivs	Allows redistribution even commercially as long as it is not changed in any way
Attribution-NonCommercial-NoDerivs	The most restrictive licence—can download works and share them with others as long as they credit the creator, but cannot change them or use them commercially

Documents that have Creative Commons licences state so, and displays one of the



¹⁰ Full descriptions of the Creative Commons licences are available at <http://creativecommons.org>

various licence logos, for example, Attribution CC-BY

There are similar licences for government data, for example the United Kingdom has an Open Government Data licence for public sector information which grants a worldwide, royalty-free, perpetual, non-exclusive licence to copy, publish, distribute and transmit, adapt and exploit the information commercially, provided that the source is acknowledged by an attribution statement. Figure 13 below shows downloadable UK datasets that clearly display the licence statement. The following examples show you how to find statistical sources, create your required dataset (example one), and a strategy to find more up-to-date data (example two). Use these as exercises to reproduce the illustrated results.

Example one

Suppose you wanted some statistics on deaths caused by natural disasters in Australia since 2000. Use OFFSTATS to find a statistical source, define the data that you require and download an Excel table of the data (Figures 10 - 11 below).

Figure 10 Browsing OFFSTATS by subject



Figure 11 Deaths by natural disasters in Australia since 2000 (source EM-DAT)

	A	B	C	D	E
1		Australia	China P Rep	Total	
2	Drought	0	134	134	
3	Earthquake (seismic activity)	0	90956	90956	
4	Epidemic	0	423	423	
5	Extreme temperature	347	193	540	
6	Flood	69	8320	8389	
7	Insect infestation	0	0	0	
8	Mass movement dry	0	55	55	
9	Storm	14	3363	3377	
10	Wildfire	206	22	228	
11	Total	636	103466	104102	
12					
13	Created on: Jan-24-2013. - Data version: v12.07				
14	Source: "EM-DAT: The OFDA/CRED International Disaster Database				
15	www.emdat.be - Université Catholique de Louvain - Brussels - Belgium"				
16					

Example two

Often finding current data is more difficult. In this example suppose you are looking for data on tobacco use in various countries, but in particular you are interested in data from the United Kingdom. There are several possible sources including the World Health Organisation (WHO), the OECD Health Statistics, the US Statistical Abstract: International Statistics and the US Center for Disease Control’s Global Tobacco Surveillance System Data (GTSSData) found through OFFSTATS.

Let’s look at some of the choices and a possible strategy.

As we need international comparative statistics I have used the United States Statistical Abstract 2012 since this includes such data over several years—time series—however the latest available figures are 2008 (see Figure 12).

The OECD database covers only the OECD member countries, although this time series does includethe 2009 data for the UK.

WHO data repository’s latest data is that of the OECD.

The UK open government data system—www.gov.uk—provides data for 2010-2011 (see Figure 13).

Figure 12 Downed loaded Excel spread-sheet from US Statistical Abstract (international statistics)

Country	Daily tobacco consumption (percent) Total					Daily tobacco consumption (percent) Females				
	2002	2003	2004	2005	2006	2007	2008	2009	1960	1970
Sweden	17.8	17.5	16.2	15.9	14.5	14.0	14.0	(NA)	(NA)	(1)
Switzerland	26.8	(NA)	(NA)	(NA)	(NA)	20.4	(NA)	(NA)	(NA)	(1)
Turkey	(NA)	34.5	(NA)	(NA)	33.4	(NA)	27.4	(NA)	(NA)	(1)
United Kingdom	26.0	26.0	25.0	24.0	22.0	21.0	22.0	(NA)	42.0	4

Figure 13 OGD (open government data) from data.gov.uk

HM Government

Statistics on Smoking, England

Licence

- UK Open Government Licence (OGL)

[OPEN DATA](#)

Description

Presents a broad picture of health issues related to smoking, smoking habits, behaviour and attitudes, smoking related ill-health and mortality and smoking related costs.

Source agency: NHS Information Centre for Health and Social Care

Designation: National Statistics

Language: English

Alternative title: Statistics on Smoking, England

Data Resources (4)

2008 Report	Details	Download
2009	Details	Download
2010	Details	Download
2011	Details	Download

As we shall see in Chapter 8 we can combine all three datasets to produce the most up-to-date information on world-wide tobacco use as well as a graph for the United Kingdom.

To sum up;

Investigative and data-driven journalism relies on authoritative and current statistics and datasets.

Organisations produce raw data (primary sources) from which they often create statistical and analytical reports (secondary sources).

The Open Government Data and non-government organisation initiatives are making datasets available to the public.

Datasets are often listed in directories or in government and non-government data repositories.

Complete data time-series may have to be compiled from several sources.

Conclusion

This chapter has provided an overview of online search and discovery that acknowledges that while journalists often depend on interviews for their information, research shows that there is a heavy reliance on search engines, particularly Google for finding information online. Research also shows that rarely do online searchers use more than two keywords in their search queries. Furthermore with the trends toward search personalisation there is the potential that possible relevant documents will be filtered out. I have introduced a number of sophisticated and effective techniques for using search engines in order to find documents that are relevant to the searchers' specific needs, particularly for investigative journalism.

The concept of the invisible Web was introduced and I have argued that this part of the Internet is an extremely valuable source of documents for serious journalism. As much of the invisible Web is not currently accessible to search engines, I have presented the case for journalists to become familiar with subject directories that include databases of grey literature—government and on-government reports, for example—and statistical information, and I have given examples for their effective use.

The final section is an introductory overview of methods for finding data and datasets which are the basis of data-driven journalism. While data journalism is

covered in detail in Chapter 8, journalists are reminded of the necessity of examining policies for the use and reuse of public datasets.

In conclusion this chapter provides a basis for journalism students as well as more seasoned journalists for becoming sophisticated searchers of online information and, by including examples that can be explored to enhance their searching experience, as well as ensuring relevant results that satisfy their online information requirements.

Further readings

HENNINGER, M. 2008. *The hidden web: finding quality information on the net*, 2nd edn, Sydney NSW, UNSW Press.

References

ABDULLA, R. A., GARRISON, B., SALWEN, M. B., DRISCOLL, P. D. & CASEY, D. 2005. Online news credibility. In: SALWEN, M. B., GARRISON, B. & DRISCOLL, P. D. (eds.) *Online news and the public*. Mahwah, N.J: Lawrence Erlbaum.

BERGMAN, M. K. 2001. White paper: the deep web: surfacing hidden value. *Journal of Electronic Publishing*, 7.

BLOCH, E. 2011. Have we reached a world of infinite information? *Flowtown*.
CARLSON, M. 2009. Dueling, Dancing, or Dominating? Journalists and their sources. *Sociology Compass*, 3, 526-542.

CASE, D. O., JOHNSON, J. D., ANDREWS, J. E., ALLARD, S. L. & KELLY, K. M. 2004. From two-step flow to the Internet: The changing array of sources for genetics information seeking. *Journal of the American Society for Information Science and Technology*, 55, 660-669.

COMSCORE 2012. State of the Internet: 3rd Quarter, 2012.

CREATIVE COMMONS. n.d. *About Creative Commons* [Online]. Available: <http://creativecommons.org/about> [Accessed 30 January 2013].

DIEKERHOF, E. & BAKKER, P. 2012. To check or not to check: An exploratory study on source checking by Dutch journalists. *Journal of Applied Journalism & Media Studies*, 1, 241-253.

GRIFFITHS, J. R. & BROPHY, P. 2005. Student searching behavior and the Web: Use of academic resources and Google. *Library Trends*, 53, 539-534.

JANSEN, B. J. & SPINK, A. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42, 248-263.

KATZ, E. & LAZARFELD, P. F. 1955. *Personal influence: The part played by people in the flow of mass communications*, Glencoe, Ill, Free Press.

MACHILL, M. & BEILER, M. 2009. The importance of the internet for journalistic research. *Journalism Studies*, 10, 178-203.

MESSNER, M. & DISTASO, M. W. 2008. The source cycle: How traditional media and weblogs use each other as sources. *Journalism Studies*, 9, 447-463.

PARK, S., HO LEE, J. & JIN BAE, H. 2005. End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research*, 27, 203-221.

ZIPF, G. K. 1949. *Human behavior and the principle of least effort*, Cambridge, MA, Addison-Wesley Press.