Where are the bodies buried on the web?

Big data for journalists

Find out more about

Websites	Whois, Blekko.com, Bit.ly, Compete.com
Emails	FindByEmail, Email headers
Companies	CrunchBase
Media	Bit.ly, YouTube, Blekko
Topics	PeerIndex, Wikipedia article traffic, Google Insights, Research.ly

Large data sets

Gathering	Extractiv, 80legs, Needlebase, Google Refine
Analyzing	Grep, Mechanical Turk, BigSheets
Visualizing	Tableau Public, OpenHeatMap, GeoCommons, Gephi
Sources	Crawling public websites, CrunchBase, US Census, Google Public Data, Infochimps, Timetric, Factual, Freebase, Wikipedia, World Bank, Kaggle

There's been a revolution in data over the last few years, driven by an astonishing drop in the price of gathering and analyzing massive amounts of information. It only cost me $120 to gather, analyze and visualize 220 million public Facebook profiles. You can use 80legs to download a million web pages for just $2.20.

The technology is not just getting cheaper, it's also getting easier to use. Companies like Extractiv and Needlebase are creating point-and-click tools for gathering data from almost any site on the web, and every other stage of the analysis process is getting radically simpler too.

What does this mean for journalists? You no longer have to be a technical specialist to find exciting, convincing and surprising data for your stories. The rest of this short guide will cover my favorite resources, along with a few examples of how they've been used to create compelling journalism.

The data world is changing so fast that this guide will be out of date before I finish it and I don't have space to go into all the nuances of the tools, so feel free to contact me at pete@petewarden.com with any questions.

Detective work

There's two main uses of the public data that's now available on the web; searching and exploration. In this first section I'll focus on targeted search, tools for when you have a specific person, organization or website in mind. The second section is about casting a wider net over large data sets to uncover patterns or stories.

Websites

When you have a page or domain that you want to know more about, there are a few free services that offer interesting information.

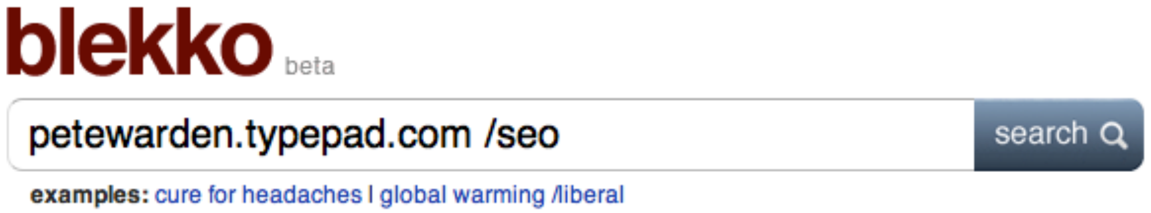
Whois - http://whois.domaintools.com

Many of you will already be familiar with whois, but it's so useful for research it's still worth pointing out. If you go to this site (or just type whois www.example.com in Terminal.app on a Mac) you can get the basic registration information for any website. In recent years, some owners have chosen 'private' registration which hides their details from view, but in many cases you'll see a name, address, email and phone number for the person who registered the site.

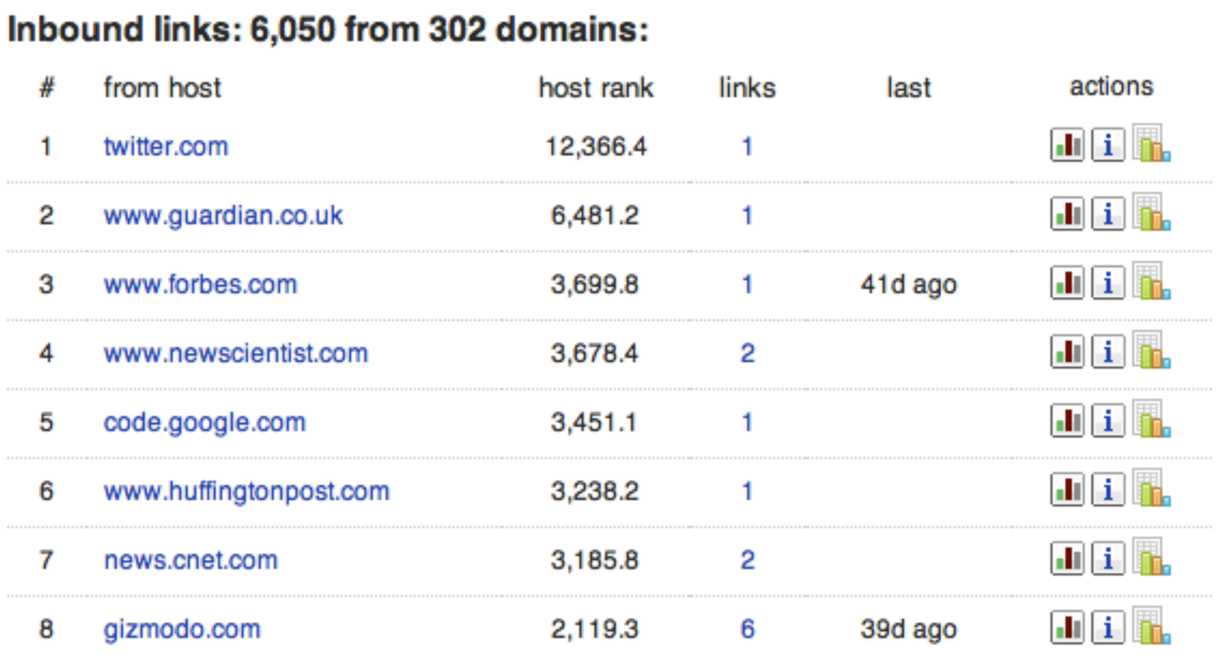
You can also enter numerical IP addresses here and get data on the organization or individual that owns that server. This is especially handy when you're trying to track down more information on an abusive or malicious user of a service, since most websites record an IP address for everyone who accesses them

Blekko - http://blekko.com

The newest search engine in town, one of Blekko's selling points is the richness of the data it offers. If you type in a domain name followed by /seo
you'll receive a page of statistics on that URL


blekko beta
petewarden.typepad.com /seo
search
examples: cure for headaches | global warming /liberal

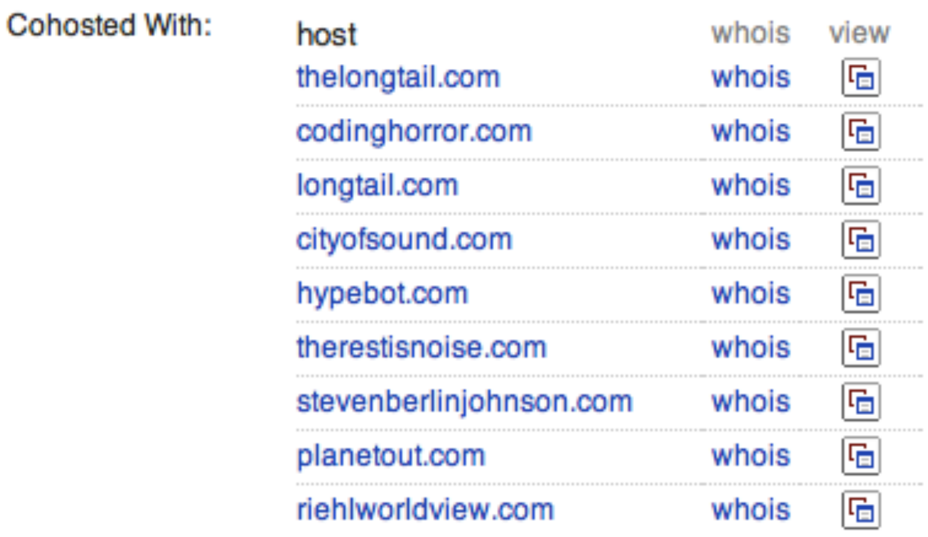
The first tab shows you which other sites are linking to the current domain, in popularity order. This can be extremely useful when you're trying to understand what coverage a site is receiving, and if you want to understand why it's ranking highly in Google's search results, since they're based on those inbound links. It would have been an interesting addition to the recent DecorMyEyes story for example.



Inbound links: 6,050 from 302 domains:

#	from host	host rank	links	last	actions
1	twitter.com	12,366.4	1		📊 i 📊
2	www.guardian.co.uk	6,481.2	1		📊 i 📊
3	www.forbes.com	3,699.8	1	41d ago	📊 i 📊
4	www.newscientist.com	3,678.4	2		📊 i 📊
5	code.google.com	3,451.1	1		📊 i 📊
6	www.huffingtonpost.com	3,238.2	1		📊 i 📊
7	news.cnet.com	3,185.8	2		📊 i 📊
8	gizmodo.com	2,119.3	6	39d ago	📊 i 📊

The other handy tab is 'Crawl stats', especially the 'Cohosted with' section



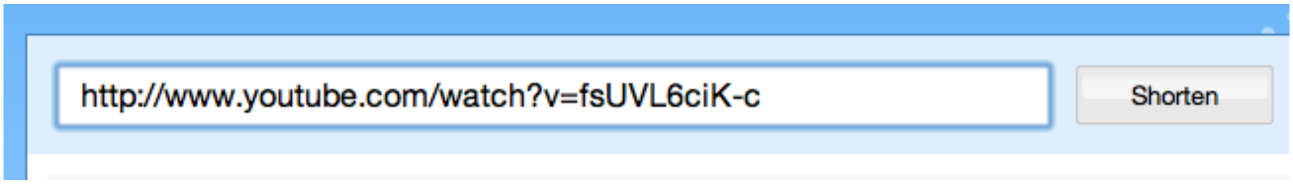
Cohosted With:

host	whois	view
thelongtail.com	whois	⧉
codinghorror.com	whois	⧉
longtail.com	whois	⧉
cityofsound.com	whois	⧉
hypebot.com	whois	⧉
therestisnoise.com	whois	⧉
stevenberlinjohnson.com	whois	⧉
planetout.com	whois	⧉
riehlworldview.com	whois	⧉

This tells you which other websites are running from the same machine. It's common for scammers and spammers to astroturf their way towards legitimacy by building multiple sites that review and link to each other. They look like independent domains, and may even have different registration details but often they'll actually live on the same server because that's a lot cheaper. These statistics give you an insight into the hidden business structure of shady operators.

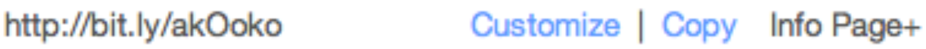
bit.ly - http://bit.ly

I always turn to this site when I want to know how people are sharing a particular link with each other. To use it, enter the URL you're interested in



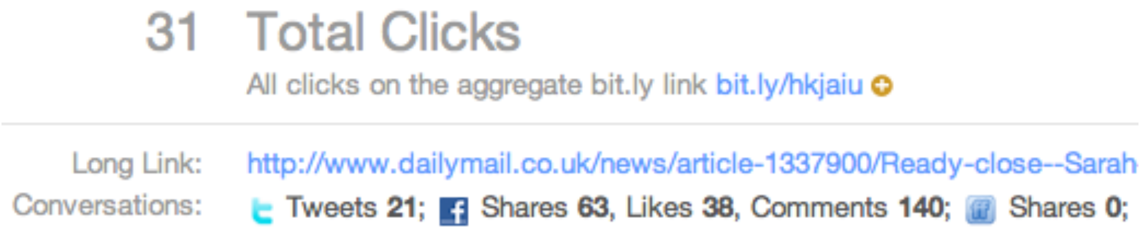
http://www.youtube.com/watch?v=fsUVL6ciK-c

Then click on the 'Info Page+' link



http://bit.ly/akOoko

That takes you to the full statistics page (though you may need to choose 'aggregrate bit.ly link' first if you're signed in to the service).



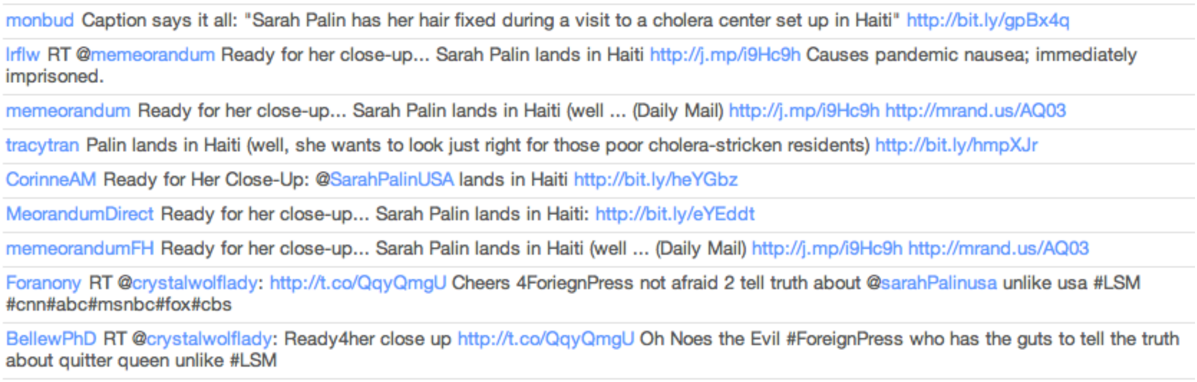
31 Total Clicks

All clicks on the aggregate bit.ly link bit.ly/hkjaiu

Long Link: http://www.dailymail.co.uk/news/article-1337900/Ready-close--Sarah

Conversations: Tweets 21; Shares 63, Likes 38, Comments 140; Shares 0;

This will give you an idea of how popular the page is, including activity on Facebook and Twitter, and below that you'll see public conversations about the link provided by backtype.com.



monbud Caption says it all: "Sarah Palin has her hair fixed during a visit to a cholera center set up in Haiti" http://bit.ly/gpBx4q

lrflw RT @memeorandum Ready for her close-up... Sarah Palin lands in Haiti http://j.mp/i9Hc9h Causes pandemic nausea; immediately imprisoned.

memeorandum Ready for her close-up... Sarah Palin lands in Haiti (well ... (Daily Mail) http://j.mp/i9Hc9h http://mrand.us/AQ03

tracytran Palin lands in Haiti (well, she wants to look just right for those poor cholera-stricken residents) http://bit.ly/hmpXJr

CorinneAM Ready for Her Close-Up: @SarahPalinUSA lands in Haiti http://bit.ly/heYGbz

MeorandumDirect Ready for her close-up... Sarah Palin lands in Haiti: http://bit.ly/eYEddt

memeorandumFH Ready for her close-up... Sarah Palin lands in Haiti (well ... (Daily Mail) http://j.mp/i9Hc9h http://mrand.us/AQ03

Foranony RT @crystalwolflady: http://t.co/QqyQmgU Cheers 4ForiegnPress not afraid 2 tell truth about @sarahPalinusa unlike usa #LSM #cnn#abc#msnbc#fox#cbs

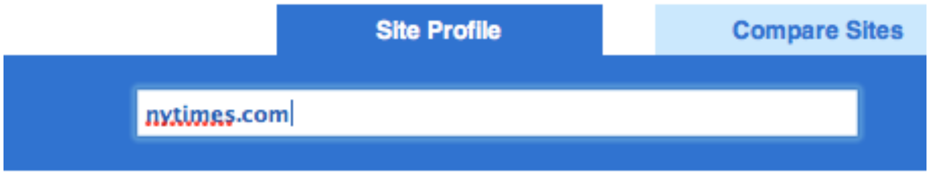
BellewPhD RT @crystalwolflady: Ready4her close up http://t.co/QqyQmgU Oh Noes the Evil #ForeignPress who has the guts to tell the truth about quitter queen unlike #LSM

I find this combination of traffic data and conversations very helpful when I'm trying to understand why a site or page is popular, and who exactly its fans are. For example, it provided me with strong evidence that the prevailing narrative about grassroots sharing and Sara Palin was wrong.

Compete - http://www.compete.com

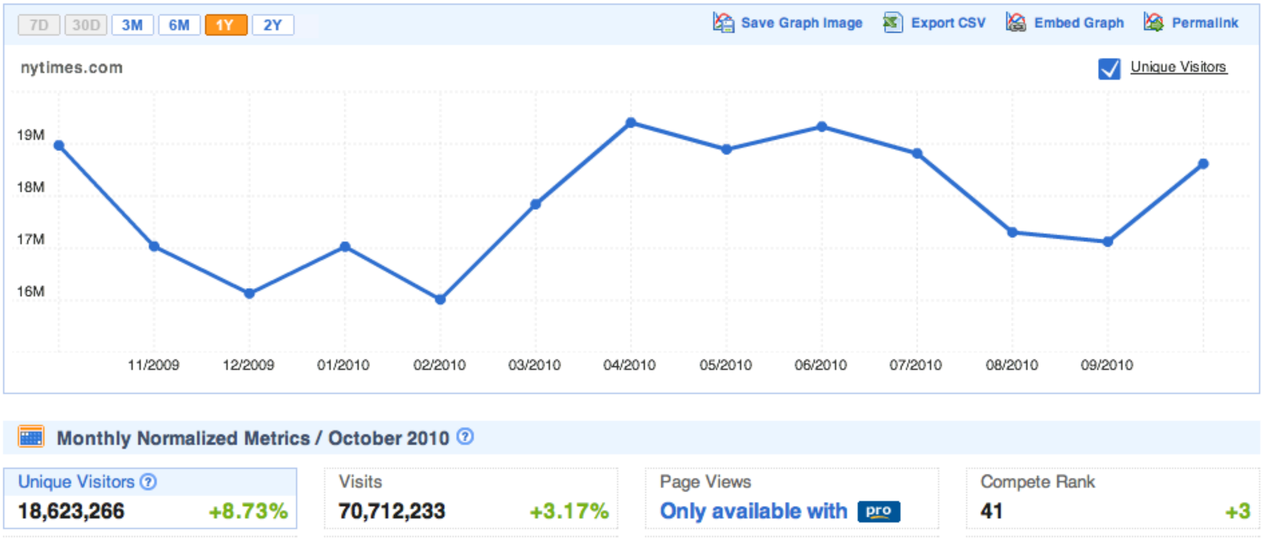
By surveying a cross-section of American consumers, Compete builds up detailed usage statistics for most websites, and they make some basic details freely available.

Choose the 'Site Profile' tab and enter a domain



nytimes.com

You'll then see a graph of the site's traffic over the last year, together with figures for how many people visited, and how often.



Monthly Normalized Metrics / October 2010

Unique Visitors	Visits	Page Views	Compete Rank
18,623,266 +8.73%	70,712,233 +3.17%	Only available with pro	41 +3

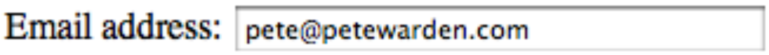
Since they're based on surveys, the numbers are only approximate, but I've found them reasonably accurate when I've been able to compare them against internal analytics. In particular, they seem to be a good source when comparing two sites, since while the absolute numbers may be off for both, it's still a good representation of their relative difference in popularity. They so only survey US consumers though, so the data will be poor for predominantly international sites

Emails

FindByEmail - http://web.mailana.com/labs/findbyemail

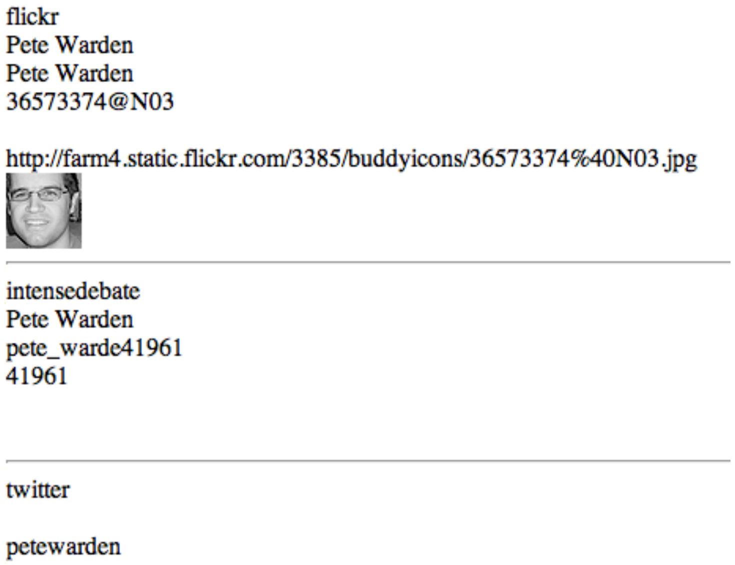
This open-source research project gathers all the services I could find that let you match an email address to information about a person, such as their location, name and social network accounts. It won't always be able to find the data you want, but it can be helpful for locating public Flickr, Twitter, etc profiles that can give interesting leads.

Enter the email address in the search box



Email address: pete@petewarden.com

After a few seconds you'll see a summary of the accounts associated with that address



flickr
Pete Warden
Pete Warden
36573374@N03

http://farm4.static.flickr.com/3385/buddyicons/36573374%40N03.jpg

intensedebate
Pete Warden
pete_warde41961
41961

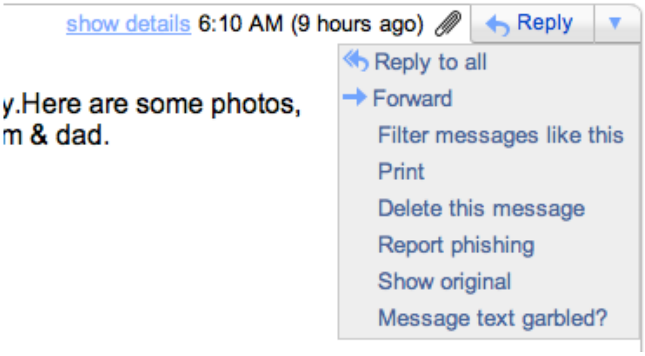
twitter

petewarden

Email headers

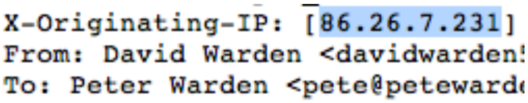
There isn't a good off-the-shelf tool available to help with this, but it can be very helpful to know the basics about the hidden headers included in every email message. These work like postmarks, and can reveal a surprising amount about the sender. In particular, they often include the IP address of the machine that the email was sent from, a lot like caller ID on a phone call. You can then run whois on that IP number to find out which organization owns that machine. If it turns out to be someone like Comcast or AT&T who provide connections to consumers, then you can visit MaxMind to get its approximate location.

To view these headers in Gmail, open the message and open the menu next to reply on the top right and choose Show original



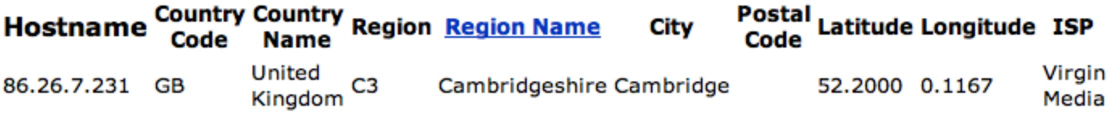
y.Here are some photos,
m & dad.

You'll then see a new page revealing the hidden content. There will be a couple of dozen lines at the start that are words followed by a colon. The IP address you're after may be in one of these, but its name will depend on how the email was sent. If it was from Hotmail, it will be called X-Originating-IP:, but if it's from Outlook or Yahoo it will be in the first line starting with Received:



```
X-Originating-IP: [86.26.7.231]
From: David Warden <davidwarden!
To: Peter Warden <pete@petewardє
```

Running the address through whois tells me it's assigned to Virgin Media, an ISP in the UK, so I put it through MaxMind's geolocation service to discover it's coming from my home town of Cambridge. That means I can be reasonably confident this is actually my parents emailing me, not imposters.



Hostname	Country Code	Country Name	Region	Region Name	City	Postal Code	Latitude	Longitude	ISP
86.26.7.231	GB	United Kingdom	C3	Cambridgeshire	Cambridge		52.2000	0.1167	Virgin Media

Companies

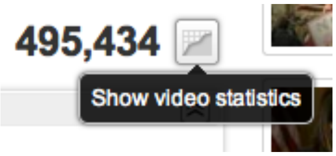
CrunchBase - http://www.crunchbase.com

TechCrunch have created a site that works a lot like Wikipedia, but focused on technology companies. If you want a concise history of any company in this field, including investors and key personnel, this is a fantastic place to look. Even better, they have an API which lets you download their data in a structured form, plus I created an open-source project that lets you pull down the information for all 50,000 firms.

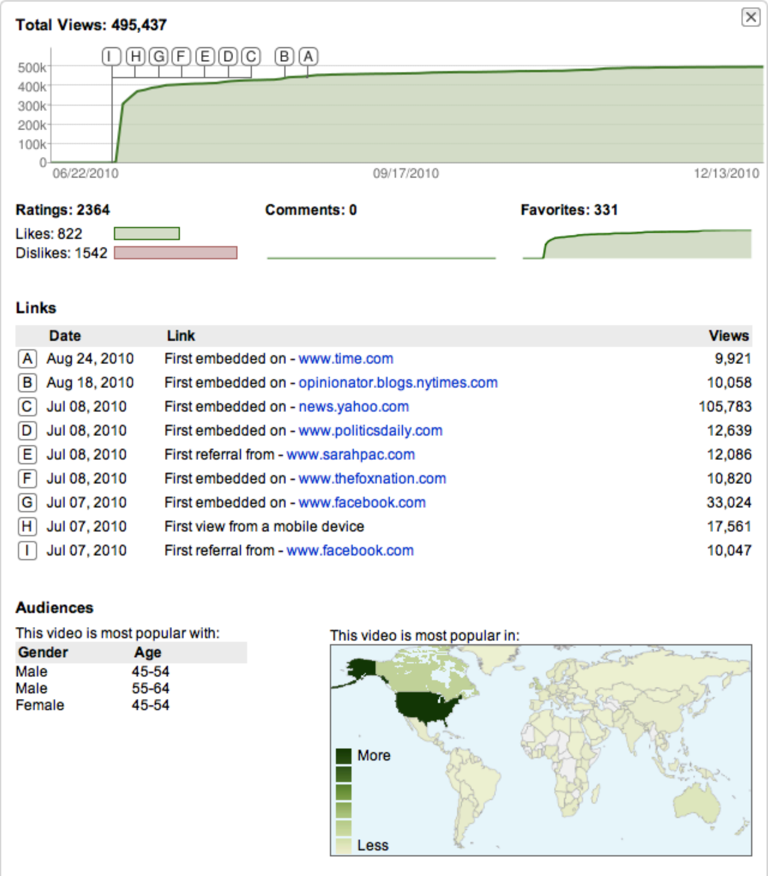
Media

YouTube - http://youtube.com

If you click on the 'statistics' icon to the lower right of any video



you'll see quite a rich set of information about its audience.



Total Views: 495,437

Ratings: 2364

Likes: 822

Dislikes: 1542

Comments: 0

Favorites: 331

Links

	Date	Link	Views
A	Aug 24, 2010	First embedded on - www.time.com	9,921
B	Aug 18, 2010	First embedded on - opinionator.blogs.nytimes.com	10,058
C	Jul 08, 2010	First embedded on - news.yahoo.com	105,783
D	Jul 08, 2010	First embedded on - www.politicsdaily.com	12,639
E	Jul 08, 2010	First referral from - www.sarahpac.com	12,086
F	Jul 08, 2010	First embedded on - www.thefoxnation.com	10,820
G	Jul 07, 2010	First embedded on - www.facebook.com	33,024
H	Jul 07, 2010	First view from a mobile device	17,561
I	Jul 07, 2010	First referral from - www.facebook.com	10,047

Audiences

This video is most popular with:

Gender	Age
Male	45-54
Male	55-64
Female	45-54

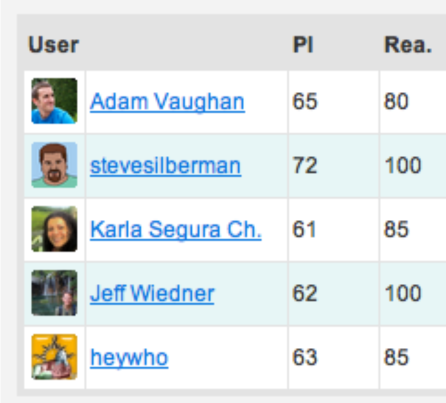
This video is most popular in:

While its not complete (in this example, the statistics total less than half of the views) it is useful for understanding the composition of the audience and the timeline behind its popularity. I was able to use this information to help Ari Melber's argument that Palin's publicity was driven by the traditional media more than social network sharing.

Topics

PeerIndex - http://www.peerindex.net

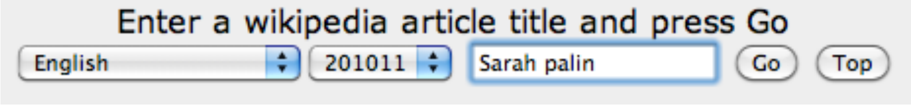
This site is good for identifying the most prominent Twitter voices discussing a topic. For example, if you want to find influential users associated with 'Climate change', you'd enter that into the search box and get back a list of the five most prominent people in that area.



User	PI	Rea.
Adam Vaughan	65	80
stevesilberman	72	100
Karla Segura Ch.	61	85
Jeff Wiedner	62	100
heywho	63	85

Wikipedia article traffic - http://dammit.lt/wikistats

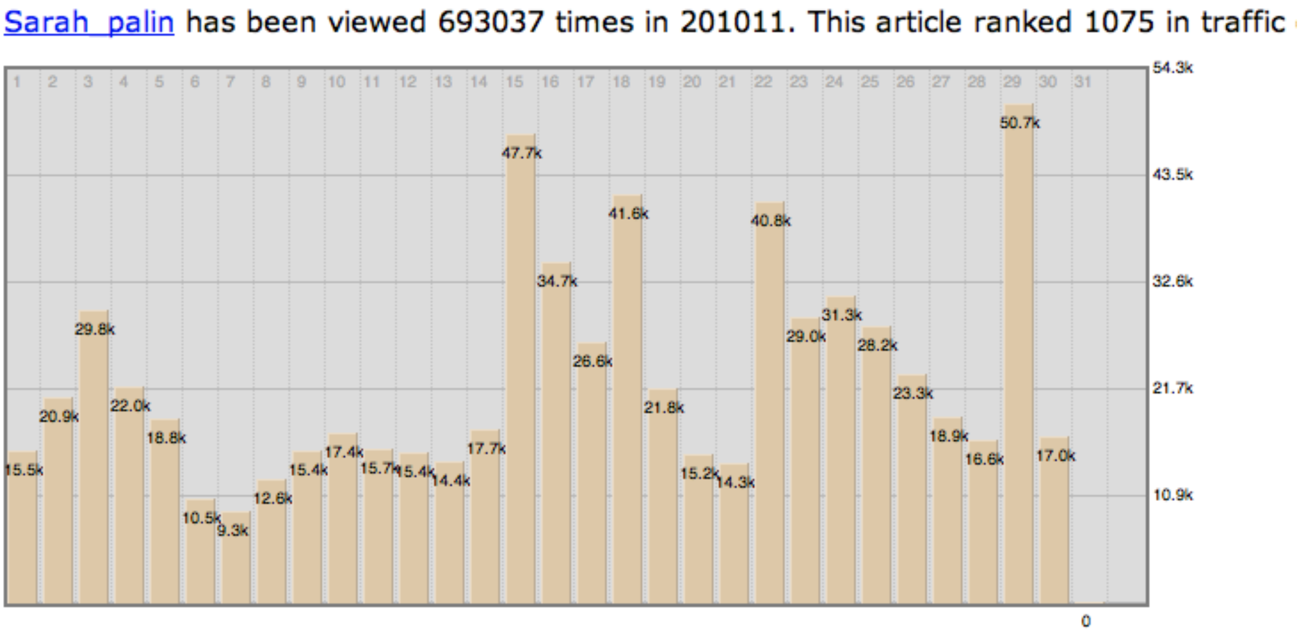
If you're interested in knowing how public interest in a topic or person has varied over time, you can actually get day-by-day viewing figures for any page on Wikipedia. The site I'm linking to is a bit rough and ready, but will let you uncover the information you need with a bit of digging. Enter the name you're interested in to get a monthly view of the traffic on that page



Enter a wikipedia article title and press Go

English | 201011 | Sarah palin

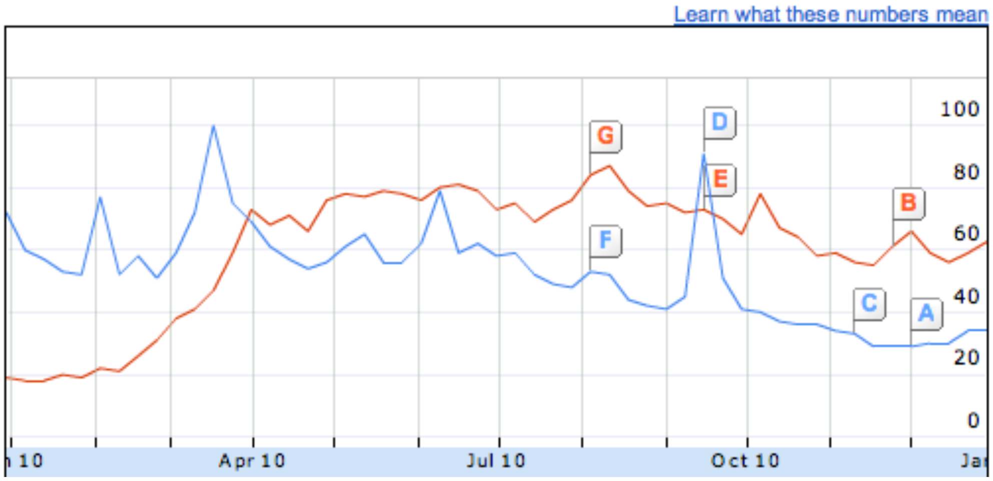
That will bring up a graph showing how many times the page was viewed for each day in the month you specify. Unfortunately you can only see one month at a time, so you'll have to select a new month and search again to see longer-term changes.



Sarah_palin has been viewed 693037 times in 201011. This article ranked 1075 in traffic

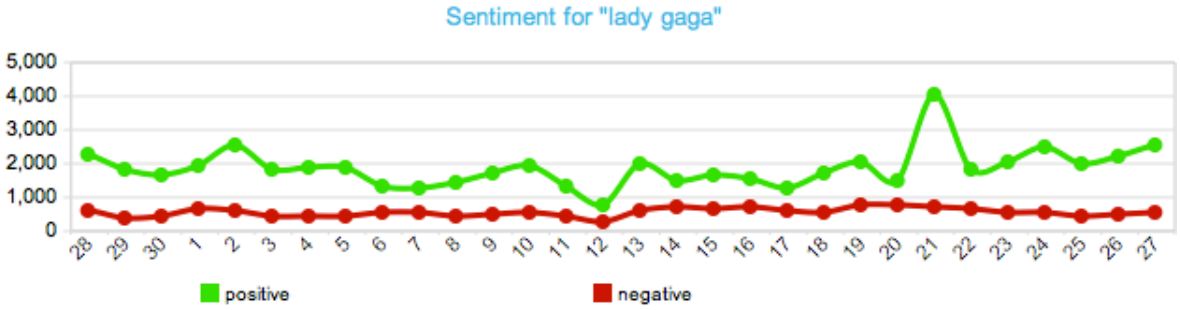
Google Insights - http://www.google.com/insights/search

You can get a clear view into the public's search habits using Insights from Google. Enter a couple of common search phrases, like Justin Bieber vs Lady Gaga, and you'll see a graph of their relative number of searches over time. There's a lot of options for refining your view of the data, from narrower geographic areas, to more detail over time. The only disappointment is the lack of absolute values, you only get relative percentages which can be hard to interpret.



Research.ly - http://research.ly/

The interface can be a bit overwhelming at first, but there's a wealth of information on Twitter conversations available at this site. Enter a topic, and the service will find related conversations over the last few months, track the volume of messages, and even how many were positive or negative comments.



Exploring public data

The tools above are great for digging deep into a very specific subject, but what if you're interested in the bigger picture? There's now a massive number of data sources that are completely open and can be analyzed to uncover stories. I think of this process as interviewing the data. Just like a person, you ask questions and get back answers, which may then lead you to ask more targeted questions, until you have enough to weave into a coherent story. In this section I'll cover the tools I use to ask those questions, and some of the places to look for interesting data to interrogate.

Gathering

Extractiv - http://www.extractiv.com

This service is definitely on the techie end of the spectrum, but is an extremely powerful way of pulling out interesting information from a large number of web pages. You tell it what pages you want it to crawl, and it will give you a summary of each page showing what things the page is talking about. For example, it can spot the names of politicians, times and locations, so you could analyze millions of news reports from a presidential campaign to understand where the candidates were spending their time. Here's a complete list of the types of entities they can understand.

80legs - http://80legs.com

A lot simpler than Extractiv, and also a lot cheaper, 80legs lets you pull out particular strings or regular expression patterns from pages. You use a web interface to tell it which sites you want to crawl, and what parts of the pages you want to save out. I often use this when I want to download information from a large number of pages on a site, and I'll then run the results through a further clean-up phase.

Needlebase - http://needlebase.com

Designed for non-technical users, Needlebase made such an impression that it was recently acquired by Google. Happily that means it was recently released as a free service. There's a point-and-click interface that lets you specify what information you want from each page by simply highlighting it, so you don't need any programming knowledge to get started. They have good tutorial videos to help you get up and running with the service, and Marshall Kirkpatrick has also written a guide for journalists based on his own experiences.

Google Refine - http://code.google.com/p/google-refine/

An update to another project that Google acquired, Refine is a great tool for taking files full of messy, unstructured information and turning it into something that structured tools like Excel and databases can read in. Its real strength is that it takes common clean-up tasks, like merging frequent spelling mistakes in a spreadsheet, and makes it easy to spot and fix the problems.

Analyzing

Grep - http://www.readwriteweb.com/hack/2010/11/how-to-search-your-source-with.php

Once you've downloaded a large data set, how do you start to figure out what it contains? If you've got a Mac, then go to the Utilities folder inside Applications, and start the Terminal program. Then type
grep -iIr 'fraud'
and drag the folder you downloaded into that window. After a few seconds, you'll see which text files contain that word, and where. If you're not used to the command line, it can be a bit frustrating at first, but it's a very powerful and fast way to pull insights from large amounts of data. If you do get comfortable, there are other advanced command-line tools available for processing big data sets.

Mechanical Turk - http://www.mturk.com

Sometimes there's no substitute for human intelligence, and so Amazon's service can be an extremely effective way of extracting meaning from large numbers of documents. All you need to do is create a template that asks a question, upload a list of the data items you want to insert into that template, and set a price you'll pay for each response, usually no more than 10 or 20 cents for something simple. The service will then send the questions out to hundreds of workers, optionally double-checking by sending the same one to multiple people and cross-referencing the answers. Marshall Kirkpatrick has a good example of how he used it to research the backgrounds of hundreds of conference attendees, but you could use the sample approach for almost any simple, repetitive research task.

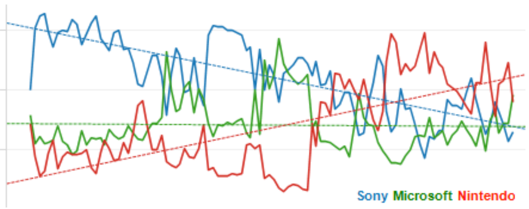
BigSheets - http://www-01.ibm.com/software/ebusiness/jstart/m2/

It's a bit of a pain to set up, but this IBM project brings the power of the MapReduce technology that Google uses to handle processing billions of web pages behind a simple spreadsheet interface. If you're comfortable with Excel but frustrated because it grinds to a halt on massive spreadsheets, this might be the tool for you.

Visualizing

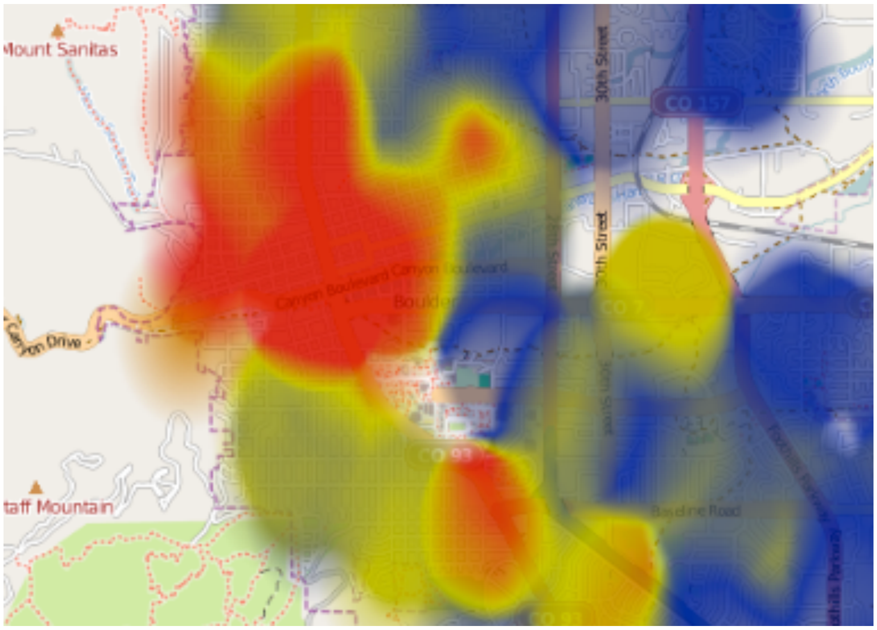
Tableau Public - http://www.tableausoftware.com/public

This free version of the popular graphing package lets you create embeddable charts and graphs, so they're very easy to use as part of an online article. They have put together a good set of training videos to help you get going too.



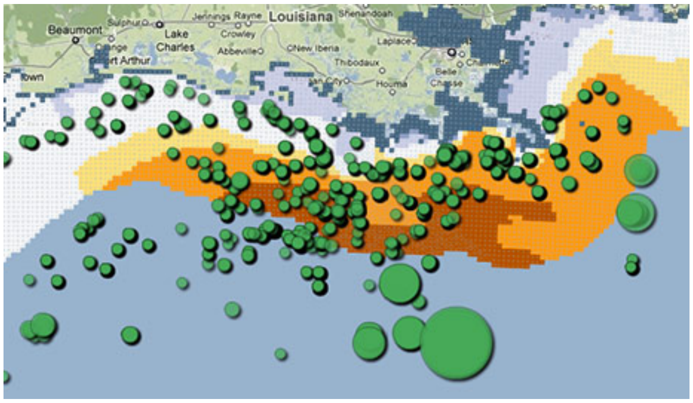
OpenHeatMap - http://www.openheatmap.com

This is my own open-source project, and grew out of my need for better tools to create interactive, animated online maps. It's aimed specifically at journalists, and is designed to transform even very messy, unstructured spreadsheets into beautiful visualizations. It's been used successfully on The Atlantic and Guardian's websites, as well as a lot of regional papers across the globe.



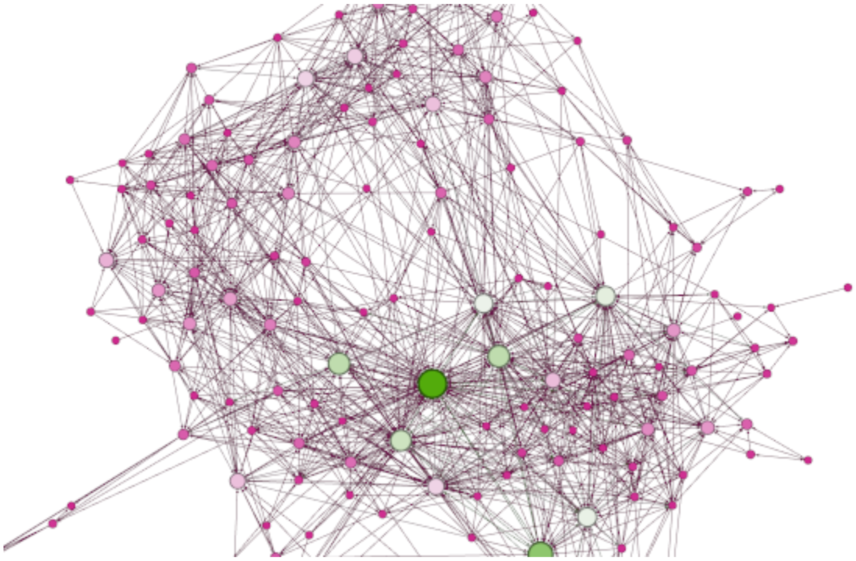
Geocommons - http://geocommons.com

Another widely-used mapping tool, with a slightly different focus than OpenHeatMap. It has a more complex interface, so you have more options and control, but also a steeper learning curve. It does come with some very useful data sets built into the site.



Gephi - http://gephi.org

If you have information about relationships between organizations or companies, it can be very effective to plot that as a network graph, where lines between objects show the relationships and clusters show how they're grouped. Often described as the Photoshop for network graphs, Gephi is an open-source desktop application that will walk you through turning your raw data into a visualization.



Sources

To create an original story, you almost always need a big picture view of a data set, so you can explore what patterns are present by slicing and dicing it in lots of different ways. Normal APIs only allow you to ask the questions the interface designers anticipated, so in this section I'm focused on ways of gathering information in large quantities for later analysis.

Crawling public websites

Most websites welcome being indexed by search engines, and it's now easy to use the Extractiv, 80legs or Needlebase tools I mentioned earlier to run the same process yourself. It's important to be certain that the site owners have rolled out the welcome mat by making sure their robots.txt file gives you permission, but all those services handle that automatically for you.

This represents a massive opportunity, because it means the information on any page you can reach without logging in can become part of your data set. As a practical example, I took the LA Times' series on local school quality and was able to extract the ranking tables they'd generated and turn them into an online map.

Is this legal and ethical? Absolutely, but because these tools have only recently become available outside of a few large companies, I recommend being sensitive to any site owner's concerns about the data you're gathering, reading their robots.txt to see if there's any comments for example. Bear in mind the owners may not realize how much information they're exposing - I was threatened with legal action by Facebook over similar work, but received an apology once things calmed down. Since robots.txt standard doesn't support anything but open or closed, I'm currently proposing an official expansion that allows publishers to set more conditions.

So, what data does this open up? Amazon's entire catalog, including recommendation links and prices. Most US government websites. Almost every blog. If you're researching an industry or organization that has a significant web presence, I bet there's some data available from indexing their pages. For an in-depth course in how to write your own web crawling software, I highly recommend ProPublica's Scraping for Journalism guide.

US Census - http://factfinder.census.gov/

With the 2010 results due to be released in January, there will be a flood of up-to-date information available soon. There truly is a phenomenal amount of data on everything from race, income, family lives and housing to immigration, broken down at a very detailed geographic level, even down to small neighborhoods. The main barrier is the confusing interface you have to navigate to pull down the information you need. I haven't found any good guides to this process, so I recommend resigning yourself to a few days of getting lost initially, and drop me an email if you end up hitting a brick wall in your search.

Google Public Data - http://www.google.com/publicdata/directory

The data sets here are largely just republished versions of information available elsewhere, but they are cleaned up, standardized and nicely presented. This makes them a lot easier to explore and use.

Infochimps - http://www.infochimps.com

A startup building a marketplace for data, they've collected some fascinating sets such as stock market prices for the last 40 years, book sales and crime rates by state. They also offer APIs for a lot of the data they cover, which is handy if you have programming resources.

Timetric - http://timetric.com

Another early-stage company, with a focus on time-series economics data sets and a good understanding of journalist's requirements. Based in the UK, they're especially good on European financial information.

Factual - http://factual.com

On the surface they might seem similar to Infochimps, but Factual are concentrating on large data sets that require lots of users to collaborate to produce them. I think of them as the wikipedia for big data, where people are expected to edit and expand the information, rather than just upload or download it as a set. They also deserve a lot of credit for listing additional external resources for the topics they cover, so even if you don't find the data you want on their site, you have leads to investigate.

Freebase - http://download.freebase.com/datadumps

Now owned by Google, this project collects a wealth of user-contributed facts on a lot of topics, from entertainment to government. The full set of data can be downloaded in bulk form, and while it's not as useful as a primary source, it does capture relationships that aren't available elsewhere.

Wikipedia - http://en.wikipedia.org/wiki/Wikipedia:Database_download

Every article on the site is available in this massive 6GB data set, in structured XML that makes it pretty straightforward to extract the information for further use. Like Freebase, this is a secondary source, but an extremely rich and well-maintained one.

World Bank - http://data.worldbank.org

An extremely useful set of thousands of economic indicators for many countries around the world, stretching back over 40 years. These aren't just financial numbers, they include detailed information on child mortality, education and human rights, as well as how many tractors each country has. The only drawback is that some of the indicators are missing for particular years, or were not collected in certain countries.

Kaggle - http://kaggle.com

I had to include this service because I enjoy it so much, despite the fact that it's not primarily a source of data. They run Netflix-style competitions designed to uncover good algorithms, and as part of this they let you download some interesting data sets. These often come with terms and conditions, and so may not be usable for journalism.

Further Reading

How to be a data journalist
Tools for visualizing data