CHAPTER NINE
SEARCH ENGINES

The first stop for many researchers will be a popular search engine. The two big players in the United States are Google and Bing. This chapter will go into great detail about the advanced ways to use both and others. Most of these techniques can apply to any search engine, but many examples will be specific for these two. Much of this chapter is unchanged from the 7th edition.

Google (google.com)

There are entire books dedicated to Google searching and Google hacking. Most of these focus on penetration testing and securing computer networks. These are full of great information, but are often overkill for the investigator looking for quick personal information. A few simple rules can help locate more accurate data. No book in existence will replace practicing these techniques in a live web browser. When searching, you cannot break anything. Play around and get familiar with the advanced options.

Quotation Marks

Placing a target name inside of quotation marks will make a huge difference in a quick first look for information. If I conducted a search for my name without quotes, the result is 147,000 pages that include the words "Michael" and "Bazzell". These pages do not necessarily have these words right next to each other. The word "Michael" could be next to another person's name, while "Bazzell" could be next to yet another person's name. These results can provide inaccurate information. They may include a reference to "Michael Santo" and "Barry Bazzell", but not my name. Since technically the words "Michael" and "Bazzell" appear on the page, you are stuck with the result in your list. In order to prevent this, you should always use quotes around the name of your target. Searching for the term "Michael Bazzell", including the quotes, reduces the search results to 31,800.

Each of these pages will contain the words "Michael" and "Bazzell" right next to each other. While Google and other search engines have technology in place to search related names, this is not always perfect, and does not apply to searches with quotes. For example, the search for "Michael Bazzell", without quotes, located pages that reference Mike Bazzell (instead of Michael). This same search with quotes did not locate these results. Placing quotes around any search terms tells Google to search exactly what you tell it to search. If your target's name is "Michael", you may want to consider an additional search for "Mike". If a quoted search returns nothing, or few results, you should remove the quotes and search again.

When your quoted search, such as "Michael Bazzell", returns too many results, you should add to your search. When I add the term "FBI" after my name, the results reduce from 31,800 to

12,000. These results all contain pages that have the words "Michael" and "Bazzell" next to each other, and include the term "FBI" somewhere on the page. While all of these results may not be about me, the majority will be and can be easily digested. Adding the occupation, residence city, general interest, or college of the target may help eliminate unrelated results. This search technique can be vital when searching email addresses or usernames. When searching the email address of "michael@inteltechniques.com", without quotes, I receive 14,200 results. When I search "michael@inteltechniques.com" with quotes, I receive only 7 results that actually contain that email address (which does not reach my inbox).

Search Operators

Most search engines allow the use of commands within the search field. These commands are not actually part of the search terms and are referred to as operators. There are two parts to most operator searches, and each are separated by a colon. To the left of the colon is the type of operator, such as "site" (website) or "ext" (file extension). To the right is the rule for the operator, such as the target domain or file type. The following will explain each operator and the most appropriate uses.

Site Operator

Google, and other search engines, allow the use of operators within the search string. An operator is text that is added to the search, which performs a function. My favorite operator is the "site:" function. This operator provides two benefits to the search results. First, it will only provide results of pages located on a specific domain. Second, it will provide all of the results containing the search terms on that domain. I will use my name again for a demonstration. I conducted a search of "Michael Bazzell" on Google. One of the results is a link to the website forbes.com. This search result is one of multiple pages on that domain that includes a reference to me. However, this search only displayed one of the many pages on that domain that possessed my name within them. If you want to view every page on a specific domain that includes your target of interest, the site operator is required. Next, I conducted the following exact search.

site:forbes.com "Michael Bazzell"

The result was all eight pages on forbes.com that include my name within the content. This technique can be applied to any domain. This includes social networks, blogs, and any other website that is indexed by search engines.

Another simple way to use this technique is to locate every page that is part of a specific domain. A search query of site:inteltechniques.com displays all 628 pages that are publicly available on my personal website. This can be a great way to review all the content of a target's personal website without attempting to navigate the actual site. It is very easy to miss content by clicking around within a website. With this technique, you should see all of the pages in a format that is easy to digest. Also, some of the pages on a website that the author may consider "private" may actually

be public if he or she ever linked to them from a public page. Once Google has indexed the page, we can view the content using the "site" operator.

Real World Application: While conducting private background checks, I consistently use the site operator. A search such as "site:amazon.com" and the target name can reveal interesting information. A previous background check of an applicant that signed an affidavit declaring no previous drug or alcohol dependencies produced some damaging results. The search provided user submitted reviews that he had left on Amazon in reference to books that he had purchased that assisted him with his continual addiction to controlled substances. Again, this result may have appeared somewhere in the numerous general search results of the target; however, the site operator directed me exactly where I needed to look.

File Type Operator

Another operator that works with both Google and Bing is the file type filter. It allows you to filter any search results by a single file type extension. While Google allows this operator to be shortened to "ext", Bing does not. Therefore, I will use the original "filetype" operator in my search examples. Consider the following search attempting to locate PowerPoint presentation files associated with the company Cisco.

"Cisco" "PowerPoint"

The result is over 10,000,000 websites that include the words Cisco and PowerPoint in the content. However, these are not all actual PowerPoint documents. The following search refines our example for accuracy.

"Cisco" filetype:ppt

The result is 15,200 Microsoft PowerPoint presentations that contain Cisco within the content. This search only located the older PowerPoint format of PPT, but not newer files that may have the PPTX extension. Therefore, the following two searches would be more thorough.

"Cisco" filetype:ppt
"Cisco" filetype:pptx

The second search provided an additional 12,700 files. This brings our total to over 27,000 PowerPoint files, which is overwhelming. I will begin to further filter my results in order to focus on the most relevant content for my research. The following search will display only newer PowerPoint files that contain the exact phrase Cisco Confidential within the content of the slides.

"Cisco Confidential" filetype:pptx

The result is exactly 1,080 PowerPoint files of interest. There are many uses for this technique. A search of filetype:doc "resume" "target name" often provides resumes created by the target which can include cellular telephone numbers, personal addresses, work history, education information, references, and other personal information that would never be intentionally posted to the internet. The "filetype" operator can identify any file by the file type within any website. This can be combined with the "site" operator to find all files of any type on a single domain. By conducting the following searches, I was able to find several documents stored on the website irongeek.com.

site:irongeek.com filetype:pdf
site:irongeek.com filetype:ppt
site:irongeek.com filetype:pptx

Previously, Google and Bing indexed media files by type, such as MP3, MP4, AVI, and others. Due to abuse of pirated content, this no longer works well. I have found the following extensions to be indexed and provide valuable results.

7Z: Compressed File
BMP: Bitmap Image
DOC: Microsoft Word
DOCX: Microsoft Word
DWF: Autodesk
GIF: Animated Image
HTM: Web Page
HTML: Web Page
JPG: Image
JPEG: Image
KML: Google Earth
KMZ: Google Earth
ODP: OpenOffice Presentation
ODS: OpenOffice Spreadsheet
ODT: OpenOffice Text
PDF: Adobe Acrobat
PNG: Image
PPT: Microsoft PowerPoint
PPTX: Microsoft PowerPoint
RAR: Compressed File
RTF: Rich Text Format
TXT: Text File
XLS: Microsoft Excel
XLSX: Microsoft Excel
ZIP: Compressed File

Hyphen (-)

The search operators mentioned previously are filters to include specific data. Instead, you may want to exclude some content from appearing within results. The hyphen (-) tells most search engines and social networks to exclude the text immediately following from any results. It is important to never include a space between the hyphen and filtered text. The following searches were conducted on my own name with the addition of excluded text. Following each search is the number of results returned by Google.

"Michael Bazzell" 31,800
"Michael Bazzell" -police 28,000
"Michael Bazzell" -police -FBI 22,100
"Michael Bazzell" -police -FBI -osint 6,010
"Michael Bazzell" -police -FBI -osint -books 4,320
"Michael Bazzell" -police -FBI -osint -books -open -source 604
"Michael Bazzell" -police -FBI -osint -books -open -source -"mr. robot" 92

The final search eliminated results which included any of the restricted words. The pages that were remaining referenced other people with my name. My goal in search filters is to dwindle the total results to a manageable amount. When you are overwhelmed with search results, slowly add exclusions to make an impact on the amount of data to analyze.

InURL Operator

We can also specify operators that will focus only on the data within the URL or address of the website. Previously, the operators discussed applied to the content within the web page. My favorite search using this technique is to find File Transfer Protocol (FTP) servers that allow anonymous connections. The following search would identify any FTP servers that possess PDF files that contain the term OSINT within the file.

inurl:ftp -inurl(http|https) filetype:pdf "osint"

The following will dissect how and why this search worked.

inurl:ftp - Instructs Google to only display addresses that contain "ftp" in the URL.

-inurl(http|https) - Instructs Google to ignore any addresses that contain either http or https in the URL. The separator is the pipe symbol (|) located above the backslash key. It tells Google "OR". This would make sure that we excluded any standard web pages.

filetype:pdf - Instructs Google to only display PDF documents.

"osint" - Instructs Google to mandate that the exact term osint is within the content of results.

Obviously, this operator could also be used to locate standard web pages, documents, and files. The following search displays only blog posts from inteltechniques.com that exist within a folder titled "blog" (WordPress).

inurl:/blog/ site:inteltechniques.com

InTitle Operator

Similar to InURL, the "InTitle" operator will filter web pages by details other than the actual content of the page. This filter will only present web pages that have specific content within the title of the page. Practically every web page on the internet has an official title for the page. This is often included within the source code of the page and may not appear anywhere within the content. Most webmasters carefully create a title that will be best indexed by search engines. If you conduct a search for "osint video training" on Google, you will receive 2,760 results. However, the following search will filter those to 5. These only include web pages that had the search terms within the limited space of a page title.

intitle:"osint video training"

Note that the use of quotation marks prevents the query from searching "video training" within websites titled "osint". The quotes force the search of pages specifically titled "osint video training". You can add "all" to this search to force all listed words to appear in any order. The following would find any sites that have the words osint, video, and training within the title, regardless of the order.

allintitle:training osint video

An interesting way to use this search technique is while searching for online folders. We often focus on finding websites or files of interest, but we tend to ignore the presence of online folders full of content related to our search. As an example, I conducted the following search on Google.

intitle:index.of OSINT

The results contain online folders that usually do not have typical website files within the folders. The first three results of this search identified the following publicly available online data folders. Each possess dozens of documents and other files related to our search term of OSINT. One provides a folder structure that allows access to an entire web server of content. Notice that none of these results points to a specific page, but all open a folder view of the data present.

http://cyberwar.nl/d/
http://bitsavers.trailing-edge.com/pdf/
http://conference.hitb.org/hitbseccon f2013kul/materials/

OR Operator

You may have search terms that are not definitive. You may have a target that has a unique last name that is often misspelled. The "OR" (uppercase) operator returns pages that have just A, just B, or both A and B. Consider the following examples which include the number of results each.

"Michael Bazzell" OSINT 61,200
"Mike Bazzell" OSINT 1,390
"Michael Bazzell" OR "Mike Bazzell" OSINT 18,600
"Michael Bazell" OR "Mike Bazell" OSINT 1,160
"Michael Bazzel" OR "Mike Bazzel" OSINT 582

Asterisk Operator (*)

The asterisk (*) represents one or more words to Google and is considered a wild card. Google treats the * as a placeholder for a word or words within a search string. For example, "osint * training" tells Google to find pages containing a phrase that starts with "osint" followed by one

or more words, followed by "training". Phrases that fit this search include: "osint video training" and "osint live classroom training".

Range Operator (..)

The "Range Operator" tells Google to search between two identifiers. These could be sequential numbers or years. As an example, OSINT Training 2015..2018 would result in pages that include the terms OSINT and training, and also include any number between 2015 and 2018. I have used this to filter results for online news articles that include a commenting system where readers can express their views. The following search identifies websites that contain information about Bonnie Woodward, a missing person, and between 1 and 999 comments within the page.

"bonnie woodward" "1..999 comments"

Related Operator

This option has been proven to be very useful over the past year. It collects a domain, and attempts to provide online content related to that address. As an example, I conducted a search on Google with the following syntax.

related:inteltechniques.com

The results included no references to that domain, but did associate it with my other websites, my Twitter page, my Black Hat courses, and my book on Amazon. In my investigations, this has translated a person's personal website into several social networks and friends' websites.

Google Search Tools

There is a text bar at the top of every Google search result page. This allows for searching the current search terms within other Google services such as Images, Maps, Shopping, Videos, and others. The last option on this bar is the "Tools" link. Clicking this link will present a new row of options directly below. This provides new filters to help you focus only on the desired results. The filters will vary for each type of Google search. Figure 9.01 displays the standard search tools with the time menu expanded.

The "Any time" drop-down menu will allow you to choose the time range of visible search results. The default is set to "Any time" which will not filter any results. Selecting "Past hour" will only display results that have been indexed within the hour. The other options for day, week, month, and year work the same way. The last option is "Custom range". This will present a pop-up window that will allow you to specify the exact range of dates that you want searched. This can be helpful when you want to analyze online content posted within a known time.

Real World Application: Whenever I was assigned a missing person case, I immediately searched the internet. By the time the case is assigned, many media websites had reported on the incident and social networks were full of sympathetic comments toward the family. In order to avoid this traffic, I set the search tools to only show results up to the date of disappearance. I could then focus on the online content posted about the victim before the disappearance was public. This often led to more relevant suspect leads.

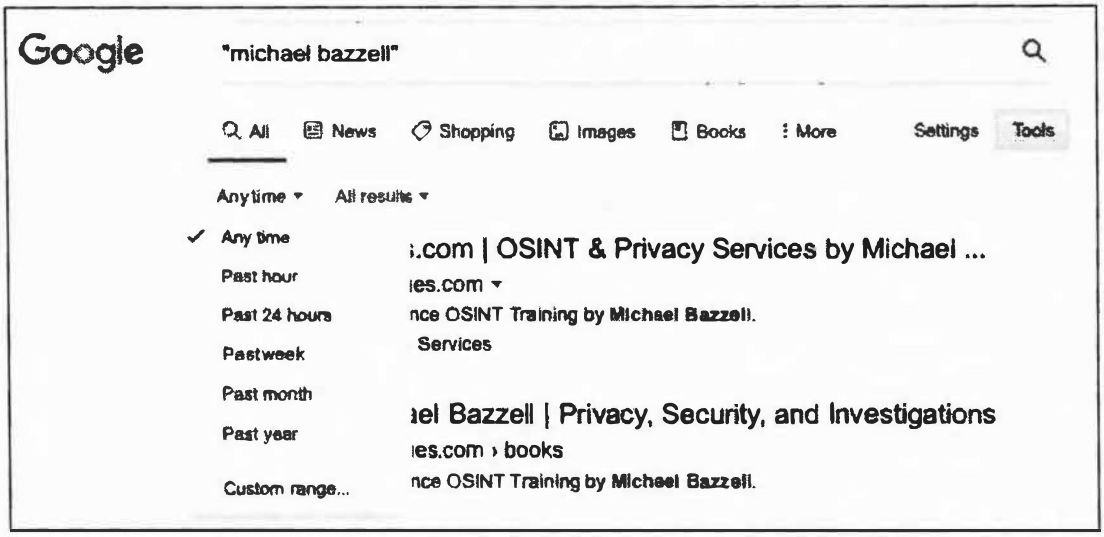

Figure 9.01: A Google Search Tools menu.

Dated Results

Google can be very sporadic when it comes to supplying date information within search results. Sometimes you will see the date that a search result was added to the Google index and sometimes you will not. This can be frustrating when you desire this information in order to identify relevant results. There is a fairly unknown technique that will force Google to always show you the date of each search result.

When you modify the "Any Time" option under the Search Tools menu, you will always see a date next to each result. If you are only searching for recent information, this solves the issue. However, if you are conducting a standard search without a specific date reference, the dates next to each result are missing. To remedy this, you can conduct a specific search that includes any results indexed between January 1, 1 BC and "today". The appropriate way to do this is to add "&tbs=cdr:1,cd_min:1/1/0" at the end of any standard Google search. Figure 9.02 displays the results of a search for the term "Michael Bazzell". The exact URL of the search was as follows.

https://www.google.com/search?q="michael+bazzell"

Notice that the result does not include a date next to the item. Figure 9.03 displays the results of this same search with the specific data added at the end. The exact URL of this search was the following address.

https://www.google.com/search?q="michael+bazzell"&tbs=cdr:1,cd_min:1/1/0

Notice that the result now has the date when the content was first indexed by Google. You can also now sort these results by date in order to locate the most recent information. The search tools menu also offers an "All results" menu that will allow you to choose to see "all results" or "Verbatim". The All Results will conduct a standard Google search. The Verbatim option searches exactly what you typed. One benefit of the Verbatim option is that Google will often present more results than the standard search. It digs a little deeper and gives additional results based on the exact terms you provided.

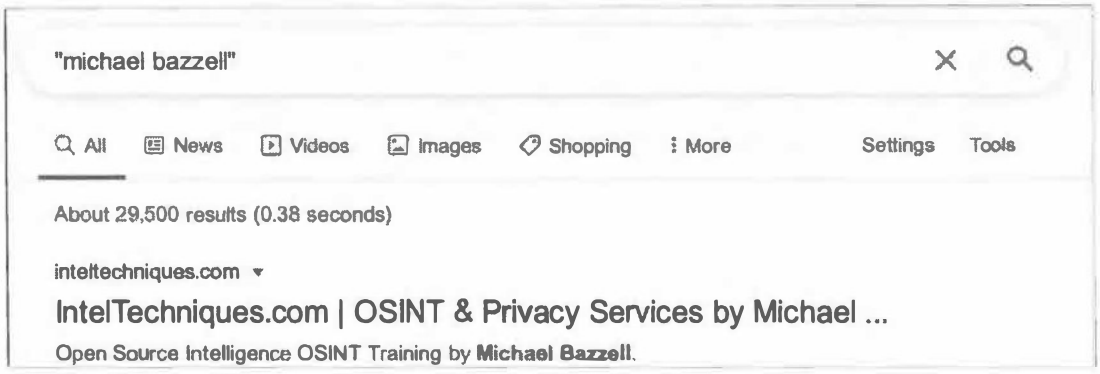

Figure 9.02: Google results without date injection.

Figure 9.03: Results with date injection.

Google Programmable Search Engines (programmablesearchengine.google.com)

Now that you are ready to unleash the power of Google, you may want to consider creating your own custom search engines, which Google has rebranded to Programmable Search Engines. Google allows you to specify the exact type of searches that you want to conduct, and then create an individual search engine just for your needs. Many specialty websites that claim to search only social network content are simply using a custom engine from Google. For our first example, we will create a basic custom search engine that only searches two specific websites.

After you log in to a Google account, navigate to the website listed above. If you have never created an engine, you will be prompted to create your first. Enter the first website that you want to search. In my example, I will search inteltechniques.com. As you enter any website to search, Google will automatically create another field to enter an additional website. The second website that I will search is inteltechniques.net. Provide a name for your custom engine and select "Create". You now have a custom search engine. You can either embed this search engine into a website or view the public URL to access it from within any web browser.

This basic functionality can be quite powerful. It is the method behind my custom Pastebin search engine discussed in a later chapter. In that example, I created a custom search engine that scoured dozens of specific websites in order to retrieve complete information about specific topics. This is only the first layer of a Google custom search engine. Google offers an additional element to its custom engines. This new layer, labeled Refinements, allows you to specify multiple actions within one custom search engine. The best way to explain this is to offer two unique examples.

For the first example, I wanted to create a custom search engine that allowed us to search several social networks. Additionally, we will isolate the results from each network across several tabs at the top of our search results. The first step will be to create a new custom search engine by clicking "New search engine" in the left menu. Instead of specifying the two websites mentioned earlier, we will identify the websites to be searched as the following.

Facebook.com	Instagram.com	YouTube.com
Twitter.com	LinkedIn.com	Tumblr.com

While this is not a complete list of active social networks, it represents the most popular social networks at the time of this writing. At this point, our custom search engine would search only these websites and provide all results integrated into one search result page. We now want to add refinements that will allow us to isolate the results from each social network.

After you have added these websites, provided a name, and created your engine, navigate to the control panel option in order to view the configuration of this custom search engine. On the left menu, you should see an option called "Edit search engine". Expanding this should present a list of your engines. Select your test engine and click "Search features". This will present a new option at the top of the page labeled "Refinements". Click the "add" button to add a new refinement for

each of the websites in this example. You should create these in the same order that you want them to appear within the search results. For this demonstration, I created the following refinements in order, accepting the default options.

Facebook	Instagram	YouTube
Twitter	LinkedIn	Tumblr

When each refinement is created, you will have two options of how the search will be refined. The option of "Give priority to the sites with this label" will place emphasis on matching rules, but will also reach outside of the rule if minimal results are present. The second option of "Search only the sites with this label" will force Google to remain within the search request and not disclose other sites. I recommend using the second option for each refinement.

Now that you have the refinements made, you must assign them each to a website. Back on the "Setup" menu option, select each social network website to open the configuration menu. Select the dropdown menu titled "Label" and select the appropriate refinement. Repeat this process for each website and save your progress. You should now have a custom search engine that will not only search several specific social network websites, but it should also allow you to isolate the results for each network. Navigate back to "Setup" in the left menu and select the Public URL link to see the exact address of your new engine. Go to that address and you should see a very plain search engine. You can now search any term or terms that you want and receive results for only the social networks that you specified. Additionally, you can choose to view all of the results combined or only the results of a specific network. Figure 9.04 displays the results when I searched the term osint. In this example, I have selected the Twitter refinement in order to only display results from twitter.com.

Figure 9.04: A Twitter refinement in a Google Custom Search.

You can now bookmark this new search engine that you created and visit it whenever you have a target to search. You can take your custom search engines to another level by adding refinements that are not website specific. In the next example, we will make a search engine that will search the entire internet and allow us to filter by file type.

Create a new custom search engine and title it "Documents". Add only "google.com" as the website to be searched. We do not actually want to search google.com, but a website is required to get to the edit panel. Save your engine, click "Edit search engine", and then click "Setup". In the "Sites to search" portion, enable the "Search the entire web" toggle. Delete google.com from the sites to be searched. You now basically have a custom search engine that will search

everything. It will essentially do the same thing as Google's home page. You can now add refinements to filter your search results. Navigate to the search features menu and add a new refinement. Title the new refinement "PDF"; change the default setting to "Give priority to the sites with this label"; and enter the following in the "Optional word(s)" field.

ext:pdf

This will create a refinement that will allow you to isolate only PDF documents within any search that you conduct. Save this setting and create a new refinement. Title it DOC; change the default search setting; and place the following in the "Optional word(s)" field.

ext:doc OR ext:docx

This will create a new tab during your search results that will allow you to isolate Microsoft Word documents. By entering both the doc and docx formats, you will be sure to get older and newer documents. The word "OR" tells Google to search either format. Repeat this process for each of the following document types with the following language for each type.

XLS (Excel Spreadsheets) - ext:xls OR ext:xlsx OR ext:csv
PPT (PowerPoint Files) - ext:ppt OR ext:pptx
TXT (Text Docs) - ext:txt OR ext:rtf
WPD (Word Perfect Docs) - ext:wpd
ODT (OpenOffice Docs) - ext:odt OR ext:ods OR ext:odp
ZIP (Compressed Files) - ext:zip OR ext:rar OR ext:7z

Figure 9.05 displays the results of a search for the term osint within this new engine. The All tab is selected which reveals 717,000 results. Clicking the PowerPoint presentations option (PPT) reveals 45 files which contain the term. There are endless possibilities with this technique. You could make an engine that isolated images with extensions such as jpg, jpeg, png, bmp, gif, etc. You could also replicate all of this into a custom engine that only searched a specific website. If you were monitoring threats against your company, you could isolate only these files that appear on one or more of your company's domains.

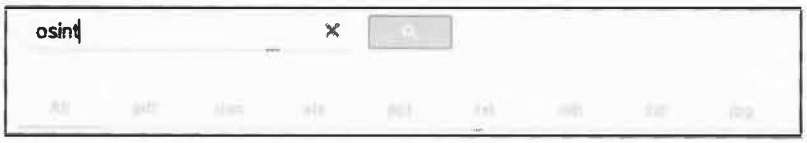

Figure 9.05: A Documents File Type Google Custom Search.

One negative aspect to custom Google search engines is that they only display the most relevant 100 results. These are presented in ten pages of ten results per page. If you are searching very specific terms, this may not be an issue. However, standard searches can be limiting.

Google Alerts (google.com/alerts)

When you have exhausted the search options on search engines looking for a target, you will want to know if new content is posted. Checking Google results every week on the same target to see if anything new is out there will get mundane. Utilizing Google Alerts will put Google to work on locating new information. While logged in to any Google service, such as Gmail, create a new Google Alert and specify the search term, delivery options, and email address to which to send the alert. In one of my alerts, Google will send an email daily as it finds new websites that mention "Open Source Intelligence Techniques" anywhere in the site. Another one of my alerts is for my personal website. I now receive an email when another site mentions or links to my website. Parents can use this to be notified if their child is mentioned in a website or blog. Investigators that are continuously seeking information about a target will find this beneficial.

Real World Application: A police detective was assigned a runaway case where a 15-year-old had decided to leave home and stay with friends at an unknown location. After several extensive internet searches, a Google Alert was set up using the runaway's name and city of residence. Within three days, one of the alerts was for a blog identifying the runaway and where she was currently staying. Within 30 minutes, the unhappy child was back home.

Bing (bing.com)

Google is not the only great search engine. While Google is the overwhelming choice of search engines used today, other sites should not be ignored, especially when having trouble locating any information on a subject. Bing is Microsoft's competition to Google and provides a great search experience. In 2009, Yahoo search (yahoo.com) began using the Bing search engine to produce search results. This makes a Yahoo search redundant if a Bing search has already been conducted. The same tactics described previously, and in the numerous Google books, can be applied to any search engine. The site operator and the use of quotes both work with Bing exactly as they do with Google. Bing also introduced time filtered searching that will allow you to only show results from the last 24 hours, week, or month. There are a couple of additional operators that are important that only apply to Bing. Bing offers an option that will list every website to which a target website links, and is the only search engine that offers this service.

Bing LinkFromDomain

I conducted a search on Bing of "LinkFromDomain:inteltechniques.com". Note that there are no spaces in the entire search string and you should omit the quotation marks. This operator creates a result that includes every website to which I have a link, located on any of the pages within my website. This can be useful to an investigator. When a target's website is discovered, this site can be large and contain hundreds of pages, blog entries, etc. While clicking through all of these is possible, sometimes links are hidden and cannot be seen by visually looking at the pages. This operator allows Bing to quickly pull links out of the actual code of the website.

Bing Contains

Earlier, I discussed searching for files with specific file extensions on Google. The "filetype" and "ext" operators that were explained both work on Bing the same way. However, Bing offers one more option to the mix. The "contains" operator allows you to expand the parameters of the file type search. As an example, a Bing search of "filetype:ppt site:cisco.com" returns 13,200 results. These include PowerPoint files stored on the domain of cisco.com. However, these results do not necessarily include links on the cisco.com website to PowerPoint files stored on other websites. A search on Bing for "contains:ppt site:cisco.com" returns 36,200 results. These include PowerPoint files that are linked from pages on the domain of cisco.com, even if they are stored on other domains. This could include a page on cisco.com that links to a PowerPoint file on hp.com. In most cases, this search eliminates the need to conduct a filetype search, but both should be attempted.

Google Images (images.google.com)

Google Images scours the web for graphical images based on a search term. Google obtains these images based on keywords for the image. These keywords are taken from the filename of the image, the link text pointing to the image, and text adjacent to the image. This is never a complete listing of all images associated with a subject, and will almost always find images completely unrelated to a target. In the case of common names, one should enclose the name in quotes and follow it with the city where the subject resides, place of employment, home town, or personal interests. This will help filter the results to those more likely to be related to the subject. When results are displayed, clicking the "Tools" button will present five new filter menus. This menu will allow you to filter results to only include images of a specific size, color, time range, image type, or license type. The most beneficial feature of Google Images is the reverse image search option. This will be explained in great detail later in the book.

Bing Images (bing.com/images)

Similar to Google, Bing offers an excellent image search. Both sites autoload more images as you get toward the end of the current results. This eliminates the need to continue to load an additional page, and leads to faster browsing. Bing also offers the advanced options available on Google, and adds the ability to filter only files with a specified layout such as square or wide. Bing provides a "filter" option in the far right of results that provides extended functionality. The People tab offers restriction for images of "Just faces" and "Head & shoulders". It also provides suggested filters with every image search. Clicking image search links may provide additional photographs of the specific target based on the listed criteria. This intelligence can lead to additional searches of previously unknown affiliations.

International Search Engines

Search engines based in the U.S. are not the primary search sites for all countries. Visiting search sites outside of the U.S. can provide results that will not appear on Google or Bing. In Russia, Yandex is the chosen search engine. Yandex offers an English version at yandex.com. These results are often similar to Google's; however, they are usually prioritized differently. In the past, I have found unique intelligence from this site when Google let me down. In China, most people use Baidu. It does not offer an English version; however, the site is still usable. Striking the "enter" key on the keyboard after typing a search will conduct the search without the ability to understand the Chinese text. New results not visible on Google or Bing may be rare, but an occasional look on these sites is warranted.

Yandex (yandex.com)

In a previous edition of this book, I only made a brief reference to Yandex and quickly moved on. In the past few years, I have discovered many advanced features of Yandex which justify an expanded section. Visually, the Yandex home page and search results pages do not possess additional search operators. These options are only available by issuing a direct command within your search. While this can be more cumbersome than a Google search, the results can include much new data. Some of these searches can be overkill for daily use, but those who conduct brand reputation monitoring or extensive background checks may take advantage of this.

Exact terms: Similar to Google and Bing, quotation marks will search for exact terms. Searching "Michael Bazzell" inside of quotes would search those terms, and would avoid "Mike" or "Bazel".

Missing word: You can search an exact phrase without knowing every word of the phrase. A search for "Open Source * Techniques" inside of quotation marks will identify any results that include that phrase with any word where the asterisk (*) is located. This identified not only results with the title of this book, but also results for "Open Source Development Techniques" and "Open Source Responsive Techniques". This search can be very useful for identifying a person's middle name. "Michael * Bazzell" produced some interesting results.

Words within the same sentence: The ampersand (&) is used in this query to indicate that you want to search for multiple terms. "Hedgehog & Flamingo", without the quotation marks, would identify any websites that contained both of those words within one sentence. If you want the results to only include sentences that have the two words near each other, you can search "Hedgehog /2 Flamingo". This will identify websites that have a sentence that includes the words Hedgehog and Flamingo within two words of each other.

Words within the same website: Similar to the previous method, this search identifies the searched terms within an entire website. "Hedgehog && Flamingo", without quotation marks, would identify pages that have both those words within the same page, but not necessarily the same sentence. You can also control the search to only include results that have those two words

within a set number of sentences from each other. A search of "Hedgehog && /3 Flamingo", without the quotation marks, would identify websites that have those two words within three sentences of each other.

Include a specific word: In Google and Bing, you would place quotation marks around a word to identify pages that contain that word in them. In Yandex, this is gained with a plus sign (+). Michael +Bazzell would mandate that the page has the word Bazzell, but not necessarily Michael.

Search any word: In Google and Bing, you can use "OR" within a search to obtain results on any of the terms searched. In Yandex, this is achieved with the pipe symbol (|). This is found above the backslash (\) on your keyboard. A search of "+Bazzell Michael|Mike|M", without quotation marks, would return results for Michael Bazzell, Mike Bazzell, and M Bazzell.

Exclude a word: Google and Bing allow you to use a hyphen (-) to exclude a word in a search. Yandex does not technically support this, but it seems to work fine. The official Yandex operator is the tilde (~). A typical search would look like "Michael Bazzell ~ Mike", without the quotation marks. This would identify websites that contained Michael Bazzell, but not Mike Bazzell. I prefer to stick with the hyphen (-) until it no longer works.

Multiple identical words: This is a technique that I have needed several times in the past before I learned of Yandex's options. You may want to search for websites that contain a specific word more than once. An example might be if you are searching for someone that has two identical words in his or her full name. "Carina Abad Abad" would fit in this scenario. You could use quotation marks to identify the majority of the results, but you would filter out anything that was not exact such as Abad,Abad, Abad-Abad, or AbadAbad. This is where the exclamation point (!) comes in. A search of "!Carina !Abad !Abad", without quotation marks, would identify any results that included those three words regardless of spacing or punctuation.

Date specific searches: While Google provides a menu to filter your searches by date, Yandex makes you work harder for it. You must specify the date range within the search. The following queries should explain the options.

date:20111201..20111231 OSINT - Websites mentioning OSINT between December 1-31, 2011
date:2011* OSINT - Websites mentioning OSINT in the year 2011
date:201112* OSINT - Websites mentioning OSINT in December of 2011
date:>20111201 OSINT - Websites mentioning OSINT after December 1, 2011

Standard operators: Most of the operators explained earlier for Google and Bing should also work in Yandex. The commands for Site, Domain, Inurl, and Intitle should work the same way. Yandex maintains a list of operators at https://yandex.com/support/search/how-to-search/search-operators.html. All Yandex operators work together and multiple operators can be used to form very specific searches. Figure 9.06 displays the results for a search of any websites from 2013 with the phrase Michael Bazzell and the word OSINT while excluding the word Mike.

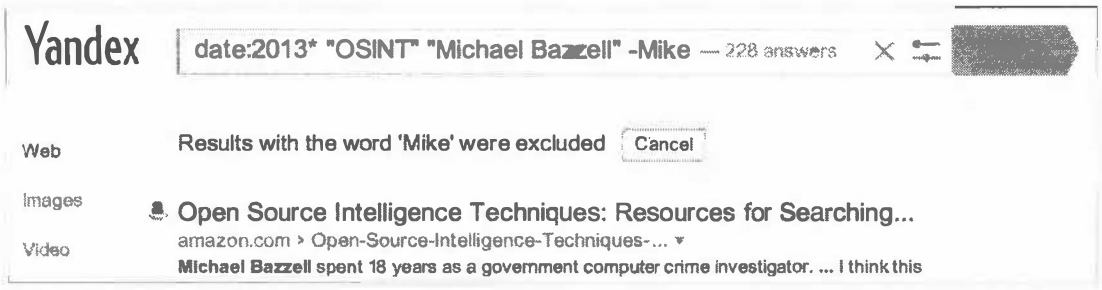

Figure 9.06: A custom Yandex search.

There are hundreds of additional international search engines. Of those, most are extremely specialized and do not offer great general search. The following have been most beneficial to my international investigations, in order of usefulness. I have included a direct search URL, which could be useful for your custom search tools.

Baidu http://www.baidu.com/s?wd=osint
Sogou https://www.sogou.com/web?query=osint
So https://www.so.com/s?q=osint
Mail.ru https://go.mail.ru/search?q=osint
Goo https://search.goo.ne.jp/web.jsp?MT=osint
Daum https://search.daum.net/search?w=tot&q=osint
Parseek http://parseek.com/Search/?q=osint
Parsijoo http://parsijoo.ir/web?q=osint
Naver https://search.naver.com/search.naver?query=osint
Coccoc https://coccoc.com/search?query=osint
Pipilika https://www.pipilika.com/search?q=osint
Seznam https://search.seznam.cz/?q=osint

I Search From (isearchfrom.com)

If you want to search Google within a version specified for another country, this site simplifies the process. Choose the country and language, and the tool will do the rest. While testing this service, I entered Japan as my country, English as my language, an iPad as my device, and OSINT as my search term. I was presented a google.co.jp search page in tablet view. Many results were similar to the U.S. version, but all were in a unique order. I find this useful when searching for international targets when I do not want bias toward a U.S. user. The "News" tab of foreign searches is often catered toward that geographical audience. This can display emphasis on news articles which would otherwise be buried in a typical Google result page.

Web Archives

Occasionally, you will try to access a site and the information you are looking for is no longer there. Maybe something was removed, amended, or maybe the whole page was permanently removed. Web archives, or "caches" can remedy this. I believe that these historical copies of websites are one of the most vital resources when conducting any type of online research. This section will explain the current options in order from most effective to least.

Google Cache (google.com)

When conducting a Google search, notice the result address directly below the link to the website. You will see a green down arrow that will present a menu when clicked. This menu will include a link titled "Cached". Clicking it will load a version of the page of interest from a previous date. Figure 9.07 (first image) displays a search for phonelosers.org which returns a result that includes a cached version of the page. This version was taken four days prior to the current date, and displays information different from the current version. The second option visible within this menu, titled "Similar", identifies web pages that contain content similar to the listed result.

If you have a specific page within a website that you want to view as a cached version, type the exact website into Google to link to the cached page. For example, if I wanted to see a previous view of the podcast for The Phone Show, an audio archive about telephone pranks, I would conduct a Google search for the site "www.phonelosers.org/snowplowshow". This will return the main landing page as well as sub-pages that will each have a cached view. If any of these pages were to go offline completely, Google would hold the last obtained version for viewing. I could have also typed the following directly into any Google search page to be navigated directly to the cached page.

Cache:www.phonelosers.org/snowplowshow

Bing Cache (bing.com)

Similar to Google, Bing offers a cached view of many websites. Searching for a domain name, such as phonelosers.org, will present many results. The first result should link to the actual website. Directly next to the website name is a small green down arrow. Clicking it will present the option of "Cached page". Clicking this link will display a previous version of the target website as collected by Bing. Figure 9.07 (second image) displays their menu option.

Yandex Cache (yandex.com)

The Russian search engine Yandex will be explained in great detail later, but it is important to note now that it also possesses a cache option. Very similar to Google and Bing, Yandex presents a green drop-down menu directly under the title of the search result. Figure 9.07 (third image) displays their cache menu option. Selecting the Cached page option opens a new tab displaying

the most recent Yandex archive of the page. The top banner displays the date and time of capture, the original website address, and a search option to highlight selected keywords within the result. The biggest strength of the Yandex cache is the lack of updates. While this may sound counterintuitive, an older cache can be very helpful in an investigation. Assume that the Phone Losers website was your target. At the time of this demonstration, September 7, 2019, the Google, Bing, and Yandex caches of this page were dated as follows.

Google: September 6, 2019
Bing: September 7, 2019
Yandex: September 1, 2019

Google and Bing tend to have very recent results which often appear identical to the live view. However, the Yandex option from a week prior is more likely to contain modified content. You can often locate a cached version of a page that is older than the Yandex version on Baidu.

Baidu Cache (baidu.com)

This Chinese search engine is the least productive as far as cached copies of websites are concerned, but it should not be ignored. It will be explained further during a later discussion about international engines. The results of a search on Baidu are mostly in Chinese, but can still be valuable to those that cannot read the text. At the bottom of each search result is a green link to the website that hosts the content of the result. While this also includes a drop-down menu, the cache option is not there. Instead, look for a word in Chinese directly to the right of this link. In Figure 9.07 (fourth image) it is displayed as 百度快照. Clicking this link will open a new tab with the cache result, which Baidu refers to as a snapshot. In my experience, the presence of this linked option does not always mean that a cached version exists.

The Wayback Machine (archive.org/web/web.php)

The Wayback Machine will provide a much more extensive list of options for viewing a website historically. Searching for phonelosers.org displayed a total of 1,280 captures of the site dating from 12/21/1997 through 6/10/2019 (Figure 9.08). Clicking the links presents quite a display of how the site has changed. Graphics are archived as well, proving that we should always think twice about which photos we post to the internet. Each view of the archived page will allow the user to click through the links as if it were a live page on the original web server. Clicking through the timeline at the top of each page will load the viewed page as it appeared on the date selected.

Wayback Search

Until 2016, you could not search keywords across Wayback Machine data. You had to know the exact URL of a target website, or at least the domain name. Today, we can search any terms desired and connect directly to archived data. At the time of this writing, a search bar was present at the top of every Wayback Machine page. If that should change, you can also conduct a search

via a direct URL. The following address searched "Michael Bazzell" throughout the entire archive of information.

https://web.archive.org/web/*/Michael Bazzell

The results identify over twenty websites that include these terms. Within those sites are dozens of archived copies of each. This data represents decades of content at your fingertips. Much of it is offline and unavailable on the current public internet. Many domains have completely shut down. Furthermore, websites that I own appear within the results, even though I have specifically blocked archiving them through a configuration file on my server. You would not find these by searching the domains directly through the Wayback Machine. This is a reminder that we should check all available resources before completing our investigations.

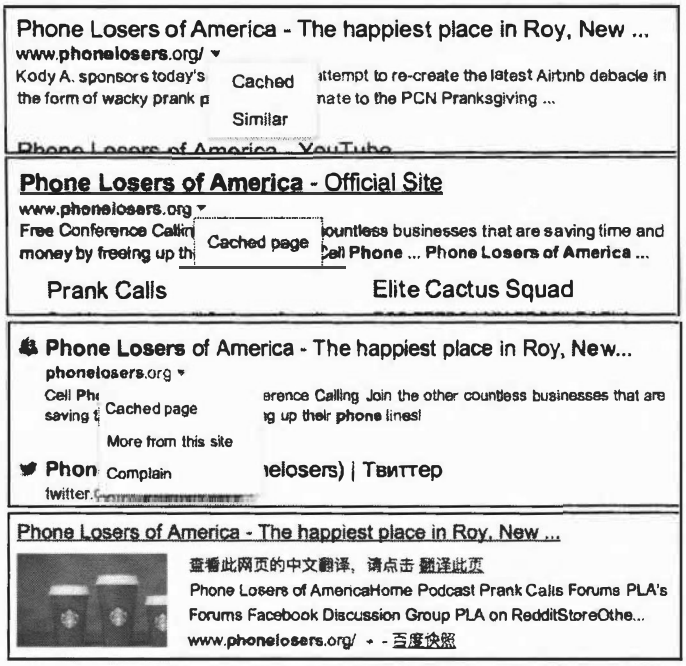

Figure 9.07: Cache menu options on Google, Bing, Yandex, and Baidu.

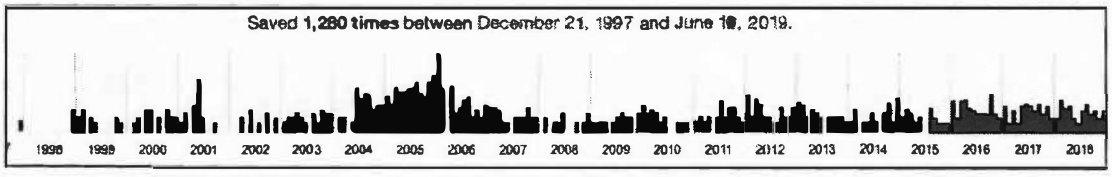

Figure 9.08: Wayback Machine results for an archived website.

Searching All Resources

Occasionally, there are websites that surface claiming to be able to extract and rebuild entire websites from online caches. In my experience, none of these have ever provided a complete historical view versus a manual approach. Engines such as Bing and Yandex generate a unique code when a cache is displayed. This action prevents most automated search tools from collecting archived information. I do not believe any option, other than navigating to each resource, will present you with the content that you need. I bookmark each of these services in an individual folder titled Archives and open each tab when I have a domain as a target. I have also created an online tool that will collect your target domain and forward you to the appropriate archive page. This will be explained later when discussing domain searches.

Finally, it is important to acknowledge that these resources can be beneficial when everything on a website appears to be present and unaltered. While caches work well on websites that have been removed and are completely empty, they also can tell a different story about websites that appear normal. Any time that I find a website, profile, or blog of interest, I immediately look at caches hoping to identify changes in content. These minor alterations can be very important. They highlight information that was meant to be deleted forever. These details can be the vital piece of your investigation puzzle. Most people have no idea that this technique exists.

Non-English Results

Not every piece of information that will be useful to you will be obtained by standard searches within English websites. Your target may either be from another country or have associates and affiliations in another country. While Google and Bing try to pick up on this, the technology is not perfect. Google has a search site and algorithm that change by location. For example, google.fr presents the French search page for Google. While this may produce the same overall results, they are usually in a different order than on google.com. Google no longer maintains a page with links to each international version of its search, but I have a preferred method.

2Lingual (2lingual.com)

This page will allow you to conduct one search across two country sites on Google. The Google search will display a plain search box and choices of two countries. The results will display in single columns next to each other. Additionally, the foreign results will be automatically translated to English. This feature can be disabled, if desired. The first few sponsored results (ads) will be similar, but the official results following should differ. This site can also be helpful when demonstrating to someone the importance of searching targets through multiple countries.

Google Translator (translate.google.com)

Many websites exist in non-English languages. As internet enthusiasts, we tend to focus on sites within our home area. There is a wealth of information out there on sites hosted in other countries which are presented in other languages. Google Translator will take text from any site or document and translate the text to a variety of languages. Usually, the service will automatically identify the language of the copied and pasted text. Selecting the desired output will provide the translation. Alternatively, you can translate an entire website in one click which will give a native view of the layout of the site. Instead of copying individual text to the search box, type or paste in the exact URL (address) of the website you want translated. Clicking the "Translate" button will load a new page of the site, which will be translated to English. This translation is rarely, if ever, perfect. However, it should give you an idea of the content presented on the page. This will also work on social network sites such as Twitter and Instagram.

Bing Translator (bing.com/translator)

A few years after Google introduced free translation services, Bing created their own product. At first glance, it looks like a replica of Google's offering. However, Bing's translations are usually slightly different than Google's results. Similar to Google, you can also type or paste an entire foreign website to conduct a translation of everything on the target page.

DeepL (deepl.com/translator)

While smaller than Google or Bing, this may be the most accurate translator service I have found. The page appears and functions identical to the previous options, but the results may be substantially different.

PROMT Online Translator (online-translator.com)

There are dozens of additional online translation tools available. Almost all of them allow translation of a small amount of text at a time. Some use either the Google or Bing translation service. One last online translation tool worth mentioning is PROMT Online Translator. It is unique from the dozens of other options in that it allows translation of entire websites similar to Google and Bing. This service provides an independent translation and can be considered a third source.

I am often asked during training which of the services I use during investigations. My answer is all of them. This is important for two reasons. The obvious benefit is that you will receive four unique translations that will be very similar. The minor variations may be important, especially when translating Tweets and other shortened messages that may not be grammatically correct in any language. The second reason is to show due diligence during my investigation. I always want to go above and beyond what is required. Translating a foreign web page through four different services emphasizes my desire to conduct an unbiased investigation.

Google Input Tools (google.com/inputtools/try)

There is one last feature regarding foreign language searching that I have found useful. Google's Input Tools allow you to type in any language you choose. Upon navigating to the above website, choose the language of your target search. In Figure 9.09, I have chosen Arabic as the language and typed "Online Investigation" on a standard English keyboard. The result is how that text might appear in traditional Arabic letters. I have had the most success with this technique on Twitter. When supplying any search term on Twitter, the results are filtered by the presence of the keywords entered and only in the language provided. Searching "Online Investigation" on Twitter only provides results that have that exact spelling in English characters. However, searching the Arabic output provides Tweets that include the Arabic spelling of the selected words. This technique is extremely important when you have located a username in a foreign language. As with all computer-generated translation services, the results are never absolutely accurate. I expect this technology to continue to improve.

Figure 9.09: A Google Input Tools translation from English to Arabic.

Google Groups (groups.google.com)

Google Groups provides access to both Usenet groups and Non-Usenet Google groups. Usenet groups are similar to mailing lists. The Usenet archive is complete and dates back to 1981. Since many people posted to these groups using their real name or email address, identifying their opinions on controversial topics is effortless. Additionally, searching a real name will often provide previous email addresses that may not be known to the searcher. This provides new intelligence for future searches. While none of this is usually damaging to the submitter, it helps provide an overall view of the target of interest. Many of the newer groups used are created through Google and conform to practically any interest imaginable. Most users continue to use a real name, screen name, email address, or a combination of all three. Searching these posts is similar to any Google search.

I have had multiple successes with searches in Google Groups. Most have been associated with pedophiles or background checks. In pedophile cases, I have identified new evidence based on historic conversations in various forums. This never identified new victims or generated new cases, but the details strengthened the current charges by showing a pattern of inappropriate interest in children. With background checks, this content has been extremely valuable. While applicants can easily clean their blogs and social profiles, they cannot easily purge their history within these groups. Many people simply have forgotten about the content, which was often posted a decade earlier.

Google News Archive (news.google.com)

This can be an amazing resource of information about a target. In the past, if someone relocated to a new geographical area, he or she could leave the past behind and start over. Today, that is difficult. Google's News Archive is continually adding content from both online archives and digitized content from their News Archive Partner Program. Sources include newspapers from large cities, small towns, and anything in between. The link referenced above will allow for a detailed search of a target's name with filters including dates, language, and specific publication. In order to display this menu, click on the down arrow to the right of the search box. This can quickly identify some history of a target such as previous living locations, family members through obituaries, and associates through events, awards, or organizations.

Google Newspaper Archive (news.google.com/newspapers)

The previous option focused solely on digital content, such as your local newspaper website. Google's Newspaper archive possesses content from printed newspapers. All results on this site consist of high-resolution scanned newspaper pages. In my experience, this collection is not as extensive as the next option discussed. However, it is definitely worth a look, and will likely continue to grow.

Newspaper Archive (newspaperarchive.com)

This paid service provides the world's largest collection of newspaper archives. The high-resolution PDF scans of entire daily newspapers range in date from the 1800's until present. The first four editions of this book explained a method of using the Google Site operator and cached results to obtain practically any page of this newspaper collection without paying or subscribing. These vulnerabilities have all been patched and none of those techniques work today. Fortunately, Newspaper Archive still offers a 14-day free trial with unlimited access to every archive. While multiple trials can be obtained, each require a unique credit card number and email address. Many libraries have asked this service to scan their entire microfilm archives and make them freely available online. You will not find any mention of this free alternative on their home page, but a bit of searching will guide you to the right place. The following search on Google identifies hundreds of public libraries that pay for your access to their archives.

site:newspaperarchive.com "This archive is hosted by" "create free account"

The first part of the search tells Google to only look at the website newspaperarchive.com. The second part mandates that the exact phrase "This archive is hosted by" appears in the result. The final piece isolates only the newspaper collections that are available for free and without a credit card. This identifies the landing pages of the various libraries that have made their collections freely available. While you will still be required to register through the service, payment is not required for these collections. Consider the following usage that will likely present you with free views of Newspaper Archive whenever you need them.

On 12/13/2017, I navigated to newspaperarchive.com/advancedsearch/ and conducted an advanced search for anyone named Michael Williams from Cedar Rapids, Iowa. Newspaper Archive presented several results from the Cedar Rapids Gazette. Clicking on any of these results prompted me to create an account and forced me to enter a valid credit card number to proceed. I could not create an account from any of the pages without providing payment. Instead, I conducted the following Google search.

site:newspaperarchive.com "This archive is hosted by" "cedar rapids gazette"

The first result was a direct connection to crpubliclibrary.newspaperarchive.com. Clicking this link presented a page dedicated to searching over 40 newspapers within the Cedar Rapids and Des Moines areas. In the upper right corner was a link titled "Create Free Account". I clicked this link and provided generic details and a throwaway email address. The membership choices now include a completely free option, which will only allow access to the Iowa newspapers. After creating my free account, I returned to crpubliclibrary.newspaperarchive.com and repeated the search of my target. Every link allowed me full unrestricted access to the high-resolution images.

While still logged in to this account, I navigated to delawarecolib.newspaperarchive.com, the direct page associated with the Delaware County Library (which I found through the original Google search in this section). I was not authorized to view this newspaper collection. However, after clicking "Create Free Account" on this page, I entered the same data as previously provided to the Iowa newspaper. After verifying my email address, I was allowed immediate access to this series of newspapers.

This technique will not obtain access to every collection on Newspaper Archive. However, it will provide a surprising amount of free access to huge collections internationally. During an hour of downtime, I created a free account on every library collection I could locate, using the same credentials on each. I can now log in to my single Newspaper Archive account and navigate the site from any page. When I reach a newspaper of interest after a search, I will be given full access if it is within a free collection. This is all thanks to the local libraries that have paid this site to give free access to the public. If the free trial of Newspaper Archive or the free library collections do not offer enough content, consider the following options.

Old Fulton (fultonhistory.com/Fulton.html): 34,000,000 scanned newspapers from the United States and Canada.

Library of Congress US News Directory (chroniclingamerica.loc.gov):
Scanned newspapers from the United States dated 1836-1922.

Library of Congress US News Directory (chroniclingamerica.loc.gov/search/titles):
Scanned newspapers from the United States dated 1690-Present.

Google Advanced Search (google.com/advanced_search)

If the search operators discussed in this chapter seem too technical, Google offers an advanced search page that simplifies the process. Navigating to the above website will present the same options in a web page that are possible by typing out the operators. This will help you get familiar with the options, but it will be beneficial to understand the operators for later use. The Advanced Search page will allow you to specify a phrase for which you are searching, just like the quotes in a search will allow. The site and filetype operators used earlier can be achieved by entering the desired filters on this page. It should be noted that the file type option on this page is limited to popular file types, where the filetype operator can handle many other file extensions.

Bing Advanced Search (search.yahoo.com/web/advanced)

Bing does not technically provide an advanced search page similar to Google's. However, since Yahoo uses Bing's search, you can use Yahoo's advanced search page as a replacement. This page will allow you to easily create a search that filters by individual terms, exact phrases, omitted terms, specific domains, file formats, and languages.

Additional Google Engines

Google isolates some search results into specialized smaller search engines. Each of these focuses on a unique type of internet search. The following engines will likely give you results that you will not find during a standard Google or Bing search. While some results from these unique searches will appear within standard Google results, the majority will be hidden from the main page.

Google Blogs (google.com)

Google removed its original blog search in 2014. It was quite helpful and focused mostly on personal websites, especially those with a blogging platform. Today, this is nothing more than a subsection of Google News. You can load the "Blogs" option under the " News" menu within the "Tools" option on any Google News results page. Alternatively, you can navigate to the following address, replacing TEST with your search terms.

google.com/search?q=TEST&tbm=nws&tbs=nrt:b

The website above displays a standard Google search option, but the results appear much differently. A standard Google search of my name reveals my website, Twitter, and Amazon pages in the first results. The Google Blogs option reveals several personal and professional (media) blogs that mention my name. These results are likely buried within the standard Google search.

Google Patents (google.com/?tbm=pts)

Google probably has the best patent search option on the internet. It allows you to search the entire patent database within any field of a patent. This can be useful for searching names associated with patents or any details within the patent itself. If you need further help, Google offers an advanced patent search at google.com/advanced_patent_search.

Google Scholar (scholar.google.com)

Google Scholar is a freely accessible web search engine that indexes the full text of scholarly literature across an array of publishing formats. It includes most peer-reviewed online journals of Europe's and America's largest scholarly publishers, plus many books and other non-peer reviewed journals. My favorite feature of this utility is the case law and court records search. I have located many court records through this free website that would have cost money to obtain from private services.

Advangle (advangle.com)

This is a simple and convenient builder of complex web search queries for both Google and Bing. The service allows you to quickly build a query with multiple parameters, such as the domain, language, or date published. Immediately you will see the result of this query in Google or Bing search engines. You can save your queries in an Advangle account if you want to restore a search to identify new results. Any condition in a query can be temporarily disabled without removing it completely. This allows you to quickly try several combinations of different conditions and choose the most suitable for your needs. Figure 9.10 displays the search page with filters for my exact name, on my website, within the past month, and only PDF files. The Google and Bing "Open" options will launch a new tab with the exact terms required for these filters.

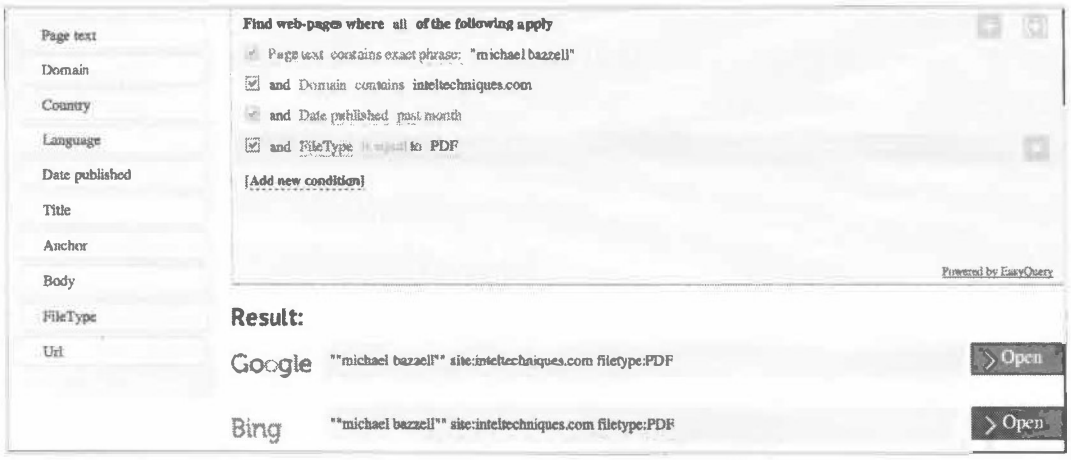

Figure 9.10: An Advangle search menu in use.

Keyword Tool (keywordtool.io)

Keyword Tool displays autocomplete data from Google, Bing, YouTube, and the App Store. You have likely noticed that Google quickly offers suggestions as you type in your search. This is called autocomplete. If I were to type "macb" into Google, it would prompt me to choose from the most popular searches when people typed those letters. This information may lead you to new terms to search in reference to your investigation. The advantage of Keyword Tool over Google is that Google only provides the five most popular entries. Keyword Tool provides the ten most popular entries. Additionally, you can choose different countries to isolate popular terms. You can also see results from similar searches that Google does not display.

Real World Application: I have successfully used this technique during the investigation of many businesses. I was once asked by a medium-sized business to investigate reports of a faulty product that they had recently recalled. They wanted to see customer complaints. After searching the typical review websites, I conducted a search with Keyword Tool. I discovered that the 9th most popular search involving this specific product name included a term that was a misspelling of the product name. It was different enough in spelling that my searches were missing this content. Knowing this information, I was able to locate more relevant data for the client.

This can also be very valuable for marketing and promotion. Assume I want to know what additional terms people search when they start with the word osint. Maybe I want to buy Google ads or tweak my website to be noticed more often. With this tool, I now know that the following are the most popular osint-related searches on Google.

osint meaning
osint websites
osint techniques
osint training

The YouTube tab tells me that people are searching for videos related to the following terms.

osint tools
osint investigations
osint phone number
osint analysis

Finally, I see that Bing users seem to be a bit more focused with the following queries.

osint resources
osint api
osint mind map
osint michael bazzell

Other Alternatives

Google and Bing are great, but they do not do it all. There will always be a need for specialized search engines. These engines usually excel in one particular search method which justifies the lack of search power in other areas. The sites listed in this next section represent the extreme minority when it comes to search traffic. It is often sites like these that implement the technologies that we later take for granted in more popular engines.

Searx (searx.me)

This is considered a meta-crawler, as it presents results from Google, Bing, and others. It often gets dismissed as another comparison search site, but there are many other advantages to using this service. First, conducting a search will provide results from the main search engines, but will remove duplicate entries. This alone is a quick way to conduct your due-diligence by checking Google and Bing. Next, the top row of options will allow you to repeat this redundancy-reducing option by checking results on Images, News, and Videos sections. Next to each result on any search page is a "cached" link. Instead of opening the Google or Bing cache, clicking this will open the cached page of the target website through the Wayback Machine. Finally, a "proxied" option next to each result will connect you to the target website through a proxy service provided by Searx. This is basically a layer of privacy preventing the website owner from collecting data about you, such as your IP address. Technically, Searx.me opened the target site, and their data would be tracked instead of yours. There are ways for adversaries to bypass this "anonymity", but it is decent protection for most sites.

The final benefit of this service over all others is the easy ability to export search results as a file. The "Links" section to the right of all search pages displays options to download a csv, json, or rss file of the results. The csv option is a simple spreadsheet that possesses all of the search results with descriptions and direct links. I find this helpful when I have many searches to conduct in a short amount of time, and I do not have the ability to analyze the results until later.

Exalead (exalead.com/search)

Headquartered in Paris, this search engine has gained a lot of popularity in the United States. The main search engine provides many results on popular searches. I have found that individual targets without a strong internet presence do not get many, if any, results on this site. However, this site excels in two areas. It works well in finding documents that include the target mentioned within the document. The "filetype" operator used in other engines works the same here. Voxalead, an Exalead search engine, searches within audio and video files for specific words. This is thanks to speech to text technologies. Voxalead will search within all of the spoken audio of a file for references to the text searched. The results are presented in a timeline view. Currently, the majority of the results of this new product link to news media and public news video files.

DuckDuckGo (duckduckgo.com)

This search engine with a clean interface offers two unique services. It has gained a lot of popularity because it does not track anything from users. Engines, such as Google, record and maintain all of your search history and sites visited. This can be a concern to privacy advocates and those with sensitive investigations. Additionally, it uses information from crowd-sourced websites such as Wikipedia and Wolfram Alpha to augment traditional results and improve relevance. You will receive fewer results here than at more popular search engines, but the accuracy of the results will improve.

Start Page (startpage.com)

Similar to DuckDuckGo, Start Page is a privacy-focused search engine that does not reveal your connection information to traditional search engines. The difference here is that Start Page only includes Google results versus DuckDuckGo's collaboration of multiple sources. The benefit to this is the ability to use Google's advanced search options while still protecting your identity. This includes filtering by date, images, and videos. Another benefit is the ability to open any result through a "proxy" link. This option, labeled "Proxy" next to each search result, opens the linked page through Start Page's servers and displays the content within their site. This protects your IP address from anyone monitoring connections at the target website. While this technique is not foolproof, it provides a valid layer of protection. My search strategy involves Start Page whenever I have a sensitive search that I do not want to associate with my computer or internet connection. This might include investigations that involve highly sensitive topics such as tech-savvy stalker suspects.

Qwant (qwant.com)

Qwant attempts to combine the results of several types of search engines into one page. It was launched in 2013 after two years of research. It has an easily digestible interface that displays results in columns titled Web, News, Images, Videos, Maps, and Music. There is a Google "feel" to it and the layout can be changed to your own preferences. A default search of my own name provided the expected results similar to Google and Bing. Clicking on the tabs at the top introduced new results not found on the other engines. The results included recent posts from Twitter, Facebook, LinkedIn, and Myspace from and about people with my name.

Million Short (millionshort.com)

This website offers a unique function that is not found on any other search engine. You can choose to remove results that link to the most popular one million websites. This will eliminate popular results and focus on lesser-known websites. You can also select to remove the top 100,000, 10,000, 1,000, or 100 results.

Tor Search Engines

Tor is free software for enabling anonymous communication. The name is an acronym derived from the original software project name The Onion Router. Tor directs internet traffic through a free, worldwide, volunteer network consisting of more than six thousand relays to conceal a user's location and usage from anyone conducting network surveillance or traffic analysis. Using Tor makes it more difficult for internet activity to be traced back to the user. This also applies to a website that is hosted on the Tor network. Usually, these sites include illegal drug shops, child pornography swaps, and weapon sales. Because these sites are not hosted on publicly viewable networks, they are hard to locate and connect. Tor-based search engines and a proxy aid this process.

Ahmia (ahmia.fi)

This is a very powerful Tor search engine. While no engine can index and locate every Tor website, this is the most thorough option that I have seen. It should be the first engine used when searching Tor related sites. The links in the results will not load if searching through a standard browser and connection. Using the Tor Browser discussed previously is the ideal way to use this service.

Dark Search (darksearch.io)

This engine appeared in 2019 and appears quite promising. When I conducted a search of "OSINT" within Ahmia, I received 5 results. The same query on Dark Search revealed 51 results. It appears to index Tor sites well, and I received many results when querying email addresses of targets. This has replaced Ahmia for many of my Tor-based investigations.

Onion Link (onion.link)

Similar to Ahmia, Onion Link attempts to identify websites within the Tor network. It uses a Google Custom Search Engine (CSE) and appends ".link" to each search result. This allows you to open these links through Onion Link's own Tor connection which appears to pull the data from their own cached sources. This makes the viewing of these pages much quicker with faster page loads, and eliminates the need to be on the Tor Browser. While relying on Google to index these pages is a bit amateur, the minimal results here are often different than other Tor engines.

Tor Search Engine (torsearchengine.net)

This service also relies on Google's indexing of Tor sites which possess URL proxy links. However, I find some "hidden" sites with this utility which were not presented by the previous option. Another alternative is **Onionland Search** (onionlandsearchengine.com).

Tor2Web (www.tor2web.org / onion.ly)

Whenever you see a URL like libertygb2nyeyay.onion, it is a Tor Onion website. As mentioned earlier, you cannot connect directly to these links without being connected to the Tor network. However, you can replace ".onion" within the address to ".onion.ly" in order to view the content. In the above example, navigating to the website libertygb2nyeyay.onion.ly will display the content using the Tor2Web proxies. This connects you with Tor2web, which then talks to the onion service via Tor, and relays the response back. This is helpful when locating Tor links on Twitter.

Tor Search Sites

I believe some of the strongest Tor search engines exist only on the Tor network. You cannot access them from a standard internet connection, and the Tor Browser is required for native use. My favorite is "Not Evil", which can be found at the following address if connected to Tor.

hss3uro2hsxfogfq.onion

Since Tor2Web allows us to use their proxy, we can connect to "Not Evil" by navigating directly to the following Tor2Web proxy address, without being on the Tor Browser.

hss3uro2hsxfogfq.onion.ly

This presents the home page of the search site, and allows for a keyword search. However, searching through this portal while being connected through the Tor2Web proxy can present difficulties. Instead, consider conducting a search within a URL submission. In the following web address, I am connecting to Tor2Web's proxy of the search engine and requesting a search results page for the term OSINT.

hss3uro2hsxfogfq.onion.ly/index.php?q=OSINT

This type of submission will be much more reliable than counting on the proxy to conduct your search and return an additional proxy-delivered page. An alternative to Not Evil is **Haystack** (haystakvxad7wbk5.onion). Similar to the previous query, we can search this Tor service without the Tor browser with the following standard URL.

http://haystakvxad7wbk5.onion.ly/?q=osint

All of the options presented within these two pages are available for automatic queries within your custom search tool, as presented in a moment.

Search Engine Collections

I believe I could fill several chapters with tutorials for the hundreds of search engines available today. Instead, I point you toward two of the best collections I have found.

Search Engine Colossus (searchenginecolossus.com)

This website is an index of practically every search engine in every country. The main page offers a list of countries alphabetically. Each of these links connects to a list of active search engines in that country. I stay away from this service when searching American-based subjects. However, if my target has strong ties to a specific country, I always research the engines that are used in that area through this website.

Fagan Finder (faganfinder.com)

This website offers an interactive search page which populates your query into hundreds of options. Figure 9.11 displays a very partial view of the options. Enter your search term in the field and click any of the hundreds of buttons to begin a search. Many of the search services are targeted toward niche uses, but you may find something valuable there which you do not see on the custom offline search tool provided at the end of this chapter.

Figure 9.11: The Fagan Finder search page.

FTP Search

I believe that the searching of File Transfer Protocol (FTP) servers is one of the biggest areas of the internet that is missed by most online researchers. FTP servers are computers with a public IP address used to store files. While these can be secured with mandated access credentials, this is rarely the case. Most are public and can be accessed from within a web browser. The overall use of FTP to transfer files is minimal compared to a decade ago, but the servers still exist in abundance. I prefer the manual method of searching Google for FTP information. As mentioned earlier, Google and Bing index most publicly available data on FTP servers. A custom search string will be required in order to filter unwanted information. If I were looking for any files including the term "confidential" in the title, I would conduct the following search on Google and Bing.

inurl:ftp -inurl (http|https) "confidential"

The result will include only files from ftp servers (inurl:ftp); will exclude any web pages (-inurl: (http|https); and mandate that the term "confidential" is present (""). I have located many sensitive documents from target companies with this query. The above search yielded 107,000 FTP results. However, these specific hits are not the only valuable data to pursue. Consider the following example. I want to locate PDF documents stored on FTP servers that contain "cisco" within the title or content, and I conduct the following search on Google.

inurl:ftp -inurl (http|https) "cisco" filetype:pdf

This results in 20,000 options within multiple FTP servers hosted on numerous domains. The first result is hosted on the Southwest Cyberport FTP server and connects to a PDF document at the following address. It appears to be a chapter of a textbook.

ftp://ftp.swcp.com/pub/cisco/03chap01.pdf

Manually changing the last "01" to "02" loads the second chapter of the book. However, it is easier to eliminate the document name altogether and browse the directory titled "cisco". The first of the following addresses displays the contents of that folder, while the second displays the content of the "pub" folder. Copy these directly into a web browser to see the results.

ftp://ftp.swcp.com/pub/cisco/
ftp://ftp.swcp.com/pub/

This type of manual navigation will often reveal numerous publicly available documents that traditional searches withhold. I have located extremely sensitive files hosted by companies, government agencies, and the military. Most File Transfer Protocol (FTP) servers have been indexed by Google, but there are other third-party options that are worth exploring. At the end of each description, I identify the number of results included for the search "Cisco" "PDF".

Global File Search (globalfilesearch.com)

Global File Search provides one of the few web-based engines for searching files on these public servers. At the time of this writing, the site claims to have indexed 243 terabytes of files in public FTP servers. Anyone searching for intelligence on any business should take a look at this site. The results are usually minimal, but very reliable.

"Cisco" "PDF": 121

Napalm FTP (searchftps.org)

This FTP search engine often provides content that is very recent. After each result, it displays the date that the data was last confirmed at the disclosed location. This can help locate relevant information that is still present on a server. While it generated the most results of all four services, many of them were no longer available on the target FTP servers. Some could be reconstructed with cached copies, but not all.

"Cisco" "PDF": 3,384

Mamont (mmnt.ru)

This Russian FTP server allows you to isolate search results by the country that is hosting the content. This is likely determined by IP address. While most of the filtered results will be accurate, I recommend searching through the global results before dismissing any foreign options. My favorite feature of this engine is the "Search within results" option. After conducting my search, I checked this option and my search field was cleared. I entered "router" and clicked search again. I was prompted with the 436 results within my original hits that also included the word router. While this could have been replicated manually, I appreciate the option.

"Cisco" "PDF": 789

For comparison, Google found 19,600 results for inurl:ftp -inurl:(http|https) "Cisco" "PDF".

wenjian (s.wenjian.net)

This Chinese service is less robust than the previous options, but it should not be ignored. In 2020, I was searching for documents in reference to an international fraud investigation. This was the only service which presented contracts signed by my target.

Nerdy Data (nerdydata.com/search)

Google, Bing, and other search engines search the content of websites. They focus on the data that is visually present within a web page. Nerdy Data searches the programming code of a website. This code is often not visible to the end user and exists within the HTML code, JavaScript, and CSS files with which most users are not familiar. This code can be extremely valuable to research in some scenarios. Viewing the source code of a website can be done by right-clicking on the background of a page and selecting "View Source". The following two examples should explain a small portion of the possibilities with this service.

In later chapters, you will learn about free services that try to identify additional websites that may be associated with your target website. The backbone of these services relies on the indexing of programming data of websites. Nerdy Data may be the purest way of searching for this data. If you were to look at the source code of one of my previous websites (no longer online), you would have seen at the bottom that I used a service called Google Analytics. This service identifies the number of visitors to a website and the general area where they are located. The following is the actual code that was present.

```
<script type="text/javascript">
try {var pageTracker = _gat._getTracker("UA-8231004-3");
pageTracker._trackPageview();
} catch(err) {}</script>
```

The important data here is the "UA-8231004-3". That was my unique number for Google Analytics. Any website with which I used the service would have needed to have that number within the source code of the page. If you searched that number on Nerdy Data a few years prior, you would have received interesting results. Nerdy Data previously identified three websites that were using that number, including computercrimeinfo.com and two additional sites that I maintained for a law firm. You can often find valuable information within the source code of your target's website.

Many web designers and programmers steal code from other websites. In the past, this would be very difficult to identify without already knowing the suspect website. With Nerdy Data, you can perform a search of the code of concern and identify websites that possess the data within their own source code. In 2013, I located a custom search website at the YGN Ethical Hacker Group that inspired me to create my own similar search service. I was curious if there were any other search websites that possessed this basic code that might give me more ideas. I looked at the source code of the website and located a small piece of code that appeared fairly unique to that service. I conducted a search on Nerdy Data for the following code.

```
<li>http://yehg.net/q?[keyword]&c=[category]  (q?yehg.net&c=Recon)</li>
```

This code was within the JavaScript programming of the search website. The search results identified 13 websites that also possessed the same code. Two of these results were hosted on the creator's website, and offered no additional information. Three of the results linked to pages that were no longer available. Three of the results linked to pages that were only discussing the code within the target website and how to improve the functionality. However, four of the results identified similar search services that were also using the programming code searched. This revealed new search services that were related to the website in which I was interested.

This same technique could be used to identify websites that are stealing proprietary code; locate pages that were created to attempt to fool a victim into using a cloned site; or validate the popularity of a specific programming function being used on hacking websites globally.

IntelTechniques Search Engines Tool

At this point, you may be overwhelmed with the abundance of search options. I can relate to that, and I do not take advantage of every option during every investigation. During my initial search of a target, I like to rely on the basics. I first search Google, Bing, Yandex, and the smaller search engines. In order to assist with this initial search, I created a custom tool that will allow you to quickly get to the basics. Figure 9.12 displays the current state of this option, which is included in the search tools archive mentioned previously.

The search options will allow you to individually search directly through Google, Bing, Yahoo, Searx, Yandex, Baidu, Exalead, DuckDuckGo, Start Page, Google Newsgroups, Google Blogs, FTP Servers, data folders, Google Scholar, Google Patents, Google News, Google Newspapers, The Wayback Machine, and others. Across all options, each search that you conduct will open within a new tab within your browser. The search all takes place on your computer within your browser, directly to the sources.

The "Submit All" option will allow you to provide any search term that will be searched across all of the services listed. Each service will populate the results within a new tab in your internet browser. Regardless of the browser that you use, you must allow pop-ups in order for the tool to work. You can also use any of the search operators discussed previously within this tool, including quotation marks.

I present a similar search tool at the end of most chapters which summarizes and simplifies the query processes for the techniques explained. I encourage you to become familiar with each of these. Once proficient, you can query target data across all options within a few minutes. This saves me several hours every week.

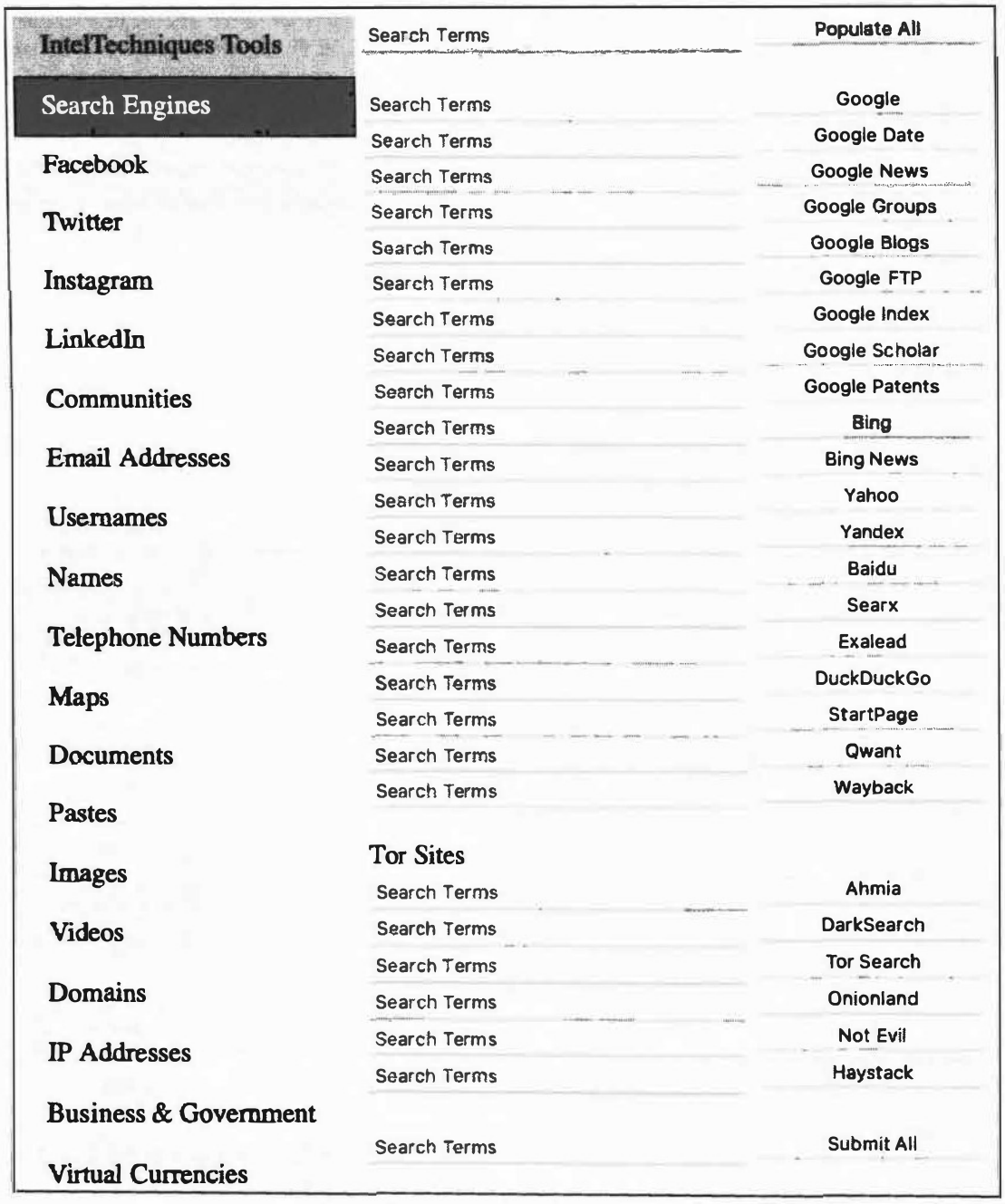

Figure 9.12: The IntelTechniques Search Engines Tool.