german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by:
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

in cooperation with:
BIRN BALKAN INVESTIGATIVE REPORTING NETWORK SERBIA

OPEN SOURCE INTELLIGENCE

navigator for investigative journalists

The GIZ Global Program Combating Illicit Financial Flows (GP IFF) supports investigative journalists from the region of Western Balkans to aid to the global agenda of fighting illicit financial flows. In cooperation with the Balkan Investigative Reporting Network (BIRN) we have addressed the Open Source Intelligence techniques applicable for journalists.

We hope that this "Navigator" will assist journalists in their research and investigations and will prove to be a useful tool which arose as an output of previously conducted training with the BIRN journalists "*Open Source Intelligence for Journalists"* that was held in May 2019 in Skopje, North Macedonia.

Developers of the Navigator: Ludo Block and Andrej Petrovski

Skopje, 2019

Deutsche Gesellschaft für
Internationale Zusammenarbeit (GIZ) GmbH

Registered offices
Bonn and Eschborn, Germany

Antonie Grubisic 5
1000 Skopje, North Macedonia
T +389 2 3103-560
F +389 2 3109-586
E giz-mazedonien@giz.de
I www.giz.de

german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by:
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Contents

1. OSINT basics

Introduction

The internet is no longer only a "super highway" but has become a place for the digital archiving and storage of massive amounts of data. Each year, people share an increasing amount of data in the form of tweets, pictures posted to Instagram, videos on YouTube, messages and photo/video on Facebook. Repositories and databases get automatically filled with vast volumes of data and old archives are digitalised and become accessible. In sum, there is a tremendous outflow of data from our electronic devices to a more-public place.

The "What Happens Online in 60 Seconds?" info graphic[1] further illustrates the situation, especially when you look back to the same infographics from previous years.

This guide is designed to help you navigate the vast amount of data in open sources and understand how to collect and analyse it.

To that end, after the introduction in this chapter, we will first discuss how to prepare your system, what tools to use, how to document to results and of course how to stay operationally secure. You will learn why concepts as 'pivoting' and 'context' are important in online research and of course we will discuss different (types of) sources available out there.

We hope that this guide will help you to understand the overall methodology and help you in your work.

This guide has been written with journalists in the Balkan region as the main audience. This means that this guide, for example, devote less time to legal matters such as chain of custody and less time to resources in faraway regions like Asia, Afrika and the Americas. On the other hand, where relevant, we have provided more coverage of Balkan related data sources and data sources more relevant for investigative journalists.

This guide could not have been written without use of many other sources. Use has been made for example of the Bellingcat guides and tutorials[2], the OSINT book by Michael Bazzell[3] and the SEC487 course guide[4] and many other sources available.[5] Where needed we will reference the sources used and attribute those behind it.

Background

Often OSINT, the acronym for Open Source INTelligence, is used as a synonym for 'open sources'. Technically that is not correct as data itself is not the same as 'intelligence'. Data, like the ownership record of an offshore entity, is just data and is meaningless without the

[1] https://www.visualcapitalist.com/internet-minute-2018/

[2] https://www.bellingcat.com/category/resources/how-tos/

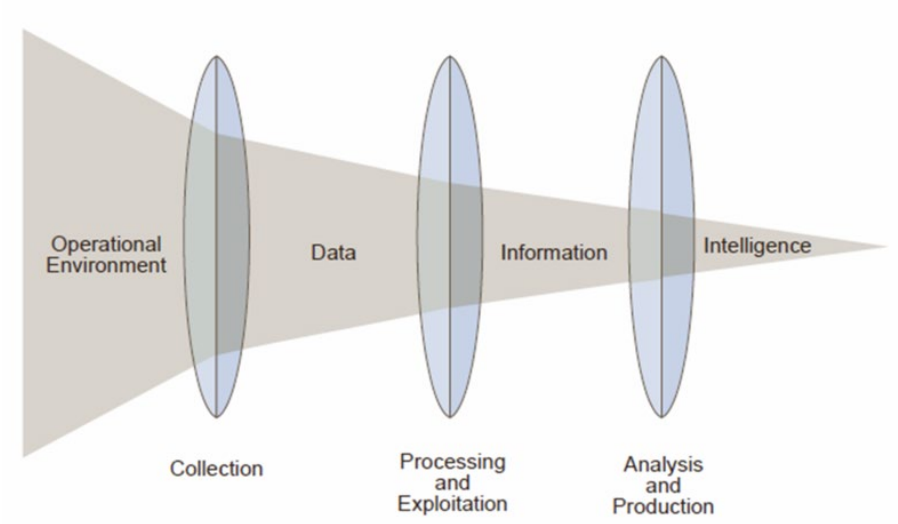
[3] https://inteltechniques.com/book1.html

[4] https://www.sans.org/course/open-source-intelligence-gathering

[5] Such as the UNESCO Handbook for Journalism Education and Training

context. Only in the context of a story which you're writing that data could be meaningful and could be the piece of an interesting puzzle. Also, to become meaningful, the data needs to be validated and analysed. And often used graphic to show the relation of data, information and intelligence can be found in the NATO open sources handbook:

Relationship of Data, Information and Intelligence



Source: Joint Intelligence / Joint Publication 2-0 (Joint Chiefs of Staff)

And while we will use the acronym OSINT in many instances in this guide, we would like to emphasise that the validation and analysis of all data you may collect so you understand it in its proper context, is perhaps the most important point.

There is a wealth of data available out there. In the more classic sense, collecting open source data mainly meant that data was taken from media, archives and perhaps some government registries if open to the public. The arrival of the digital age has profoundly changed that. Not only has the amount of data produced, transmitted, stored worldwide increased exponentially to unimaginable proportions, also, and probably more importantly, the nature of the data and available types of data have changed. At least five elements of that change are relevant.[6]

First, today exponentially more data is produced and stored than a few decades ago, and the amount of produced data keeps rising.[7] The chances that relevant data for any type of problem is publicly available have increased significantly.

Significantly, data produced nowadays is digital in nature instead of analogue, and an important consequence is that digital data is easy to index and to search. Compare for example your current University library full text search access with the ancient catalogue systems.

Third, the interconnectedness of data sources (i.e. the internet) and digitalisation of analogue datasets makes data from all over the world instantly accessible from our desktop. There is

[6] See https://www.leidensafetyandsecurityblog.nl/articles/solving-the-mh17-and-the-skripal-case-how-bellingcat-demonstrates-the-power

[7] https://www.visualcapitalist.com/how-much-data-is-generated-each-day/

hardly any reason to undertake painstaking research in damp archives anymore to obtain data, other than it could be an interesting experience of course.

Also, new types of data have emerged. For decades OSINT was dominated by content from traditional media, however, the internet gave rise to many new data types. A key example is of course the user created content in Social Media, including linkages, locations, sentiments as well as user-generated photo and video.

But also think of the open (Internet of Things / government) data and 'data-breach data'. In particular this latter type, data-breach data, which includes data on the activity of users across the internet (e.g. leaked passwords, phone numbers and credentials) and leaked government registry data, has been leveraged by Bellingcat in their Skripal research.

A last important element of change which reinforces the power of OSINT is the wide availability of computational power and digital tools to the general public. While collecting and processing large data sets used to be the prerogative of state (and academic) institutions which had access to large mainframes, nowadays there is an abundance of inexpensive tools available to the general public that allow for the collection, processing and analysis of large data sets. For example, the scraping of personal data of all citizens of the Kyrgyz Republic is actually not that hard.[8]

Investigation and validation

Research in open sources generally has two aims. First is of course to investigate something of interest on which there may be data available somewhere. Collection of unknown data is the primary aim in this case. Questions that come up in investigations include:

- What can we find on a certain event, person or company?
- Who is connected to who, who is behind a certain company?
- Where has this person be at a certain moment in time?

Another aim is often the verification of presented data, or in other words 'fact checking', which may be even more important in our present time. Questions that come up in verifications include:

- Has this person have been in that place at that time?
- Are the data they present correct and complete?
- Do they post a lot of content like that, are they knowledgeable?
- Can you find their other data and cross reference?

The amount of information in open sources that can be accessed through the internet is enormous. However, it is good to remember that from your screen you see only a virtual world. The real world may be different. A call to a friendly journalist with local knowledge and/or speaking the local language can often be very clarifying. And sometimes even that is not enough, you just need to go out there, be in the field and verify the facts yourself.

Legal and ethical considerations

Although the data we will discuss in this guide is 'open' in the sense that it is obtainable for everyone without hacking, there may be legal restrictions and requirements depending on the country where you operate, or the organisation you work for. An there are certainly ethical considerations connected to collecting and using data from open sources.

[8] https://www.bellingcat.com/resources/how-tos/2019/02/14/creating-your-own-citizen-database/

An important legal consideration may be data protection regulations, to be specific the EU General Data Protection Regulation (GDPR)[9] which came into force on 25 May 2018. Article 85 of the GDPR exempts the processing of personal data for journalistic purposes ('in the public interest') from many of the limitations and rules under the GDPR. However, the implementation of this article may differ between countries. Even though article 85 gives firm instructions to the Member States, some Member States have formulated conditions that have to be met before the processing falls under the journalistic purpose exemption.

We can only stress that you make sure that you understand the legal situation in the country where you publish / collect your data. There are cases, for example un Romania and Hungary, where authorities have (ab)used the GDPR to harass journalists.

In additional to potential legal consideration, the use of data from open sources may also need ethical considerations. What would you do with collateral findings which may harm bystanders? Are you proportional in divulging information in your publications, does it really serve the public interest? You may already have an organisation policy on these matters, make sure it also applies to data from open sources.

Whereas we proceed from the principle data in open source data collection we do not hack, steal or lie to obtain data, there is of course a grey area. For example, the following situations may be in a grey area:

- Using fake social media accounts ('a persona') to view profiles / connect to profiles of those you are researching. This usually violates the terms of the social media platform, but is for your own protection. Is that an accepted practice?;
- Using 'leaked' data (Wikileaks, Offshore Leaks, copied official databases, 'data breach' data). Technically you did not steal that data, but it was stolen somewhere. Does your organisation have a policy on using that type of data?
- Google 'dorking' and using Shodan to access unsecured devices or devices with default passwords. Technically you are not 'breaking in' into a system but at the same time, you're not supposed to be there either.[10]

Make sure that you understand what your personal position is in such grey areas and make sure what the position of your organisation is. We assume that you will use the five core principles of ethical journalism[11] as guidance:

- Accuracy – no deceptive handling of facts;
- Independence – not on behalf of anyone else – transparency about what you do;
- Fairness and Impartiality – recognize that there are more sides to a story;
- Humanity – be aware of the consequences of what you publish;
- Accountability – own and correct your mistakes

Note

A last note in this introduction. There are many sources referenced in this guide. However, nothing is subjected to change as much as the internet. All links to the sources were working at the time of writing this guide in April/May 2019, however may have changed and disappeared since then.

[9] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[10] In the US under the Computer Fraud & Abuse Act (CFAA) these actions may actually be qualified as a crime.

[11] https://ethicaljournalismnetwork.org/

2. Preparing your system and tools

The methods and techniques compiled in this guide are mostly based on free or inexpensive tools which everyone with a laptop of desktop already has available or can easily obtain. While we acknowledge that there are various specialised tools available for the collection and analysis of data from open sources, there are two reason why we will not discuss these.

The first reason is simply the price. Many of these tools are quite expensive and realistically cannot be afforded by most media outlets and journalists. Secondly, we feel that learning the tradecraft by using the simple tools provides a much more solid foundation of skills and experience. While tools can be helpful by harvesting larger datasets, these cannot replace the skills and tenacity of an experienced open source investigator.

Therefore, in this guide we will make use of mostly free basic tools. We will discuss setting up your system, bowser, additional tools and how to create a collection of sources for your work.

System

The choice for a system is probably the most fundamental choice you have to make. For ease of use, many choose to use their own laptop for their OSINT work. While to some point that is understandable, it opens you up to significant risks. Researching open sources means that you click on stuff more than average. Therefore, you may accidentally run into malware that will infect your laptop with potentially disastrous consequences depending on how much data is on it. Also, especially if you're an investigative journalist, you may be in the cross hairs of those that are subject of your writing. The malware you attract may not be random.

There are several alternatives, some of which require extra funding, some require technical knowledge. The most used options are:

- Using a separate laptop for OSINT work and only export the relevant findings to your personal system or the organisation system/repository. If the laptop is infected, you wipe it and reinstall it. This solution takes additional funding however;
- Using a Virtual Machine (VM) and only export the relevant findings to your personal system or the organisation system/repository. The VM can be any operating system (OSX, Windows, Linux) but mostly Linux and OSX can be preferred over Windows. There are different VMs for OSINT work available such as Buscador[12] which have all tools pre-installed. You can easily start each research project with a clean version of the VM so you also avoid cross-contamination of research findings. You do need a bit extra technical knowledge and you will need to obtain some Linux skills;
- Using the Silo browser of Athentic8 which gives you a secure cloud instance every time you start up Silo.[13] You can browse and save as much as you want and only download the actual relevant findings to your laptop. The bowser also hides your IP and location to the sites you visit. There are some costs associated, about 150 USD / Year. Also, the Silo browser is not as customizable as other browsers.

[12] https://inteltechniques.com/buscador/

[13] https://www.authentic8.com/

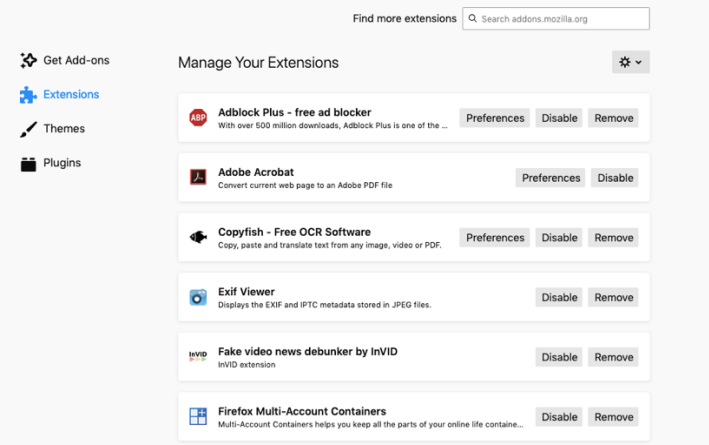
Take the time to do some thinking about the requirements. Do you have budget? Time to learn new skills? Are you the only one working or should you be looking for a solution that fits a team?

Browser

Whatever system you choose, your main tool, your window to the online world of open sources is your browser so choosing and customising it, requires and equal amount of attention. There are dozens of browsers available and discussing all of them would be unfeasible. Two browsers stand out as being most used for open source data collection: Firefox[14] and Chrome[15].

Make sure that whichever browser you choose, that you understand it. Take the time to learn and adjust the different privacy and security settings of your browser as well as the configuration of the screen.

Both Firefox and Chrome have the ability to ad many extensions or add-ons that each give functionality to the browser. You can search for the extensions in the browser,



Much used extensions that are useful for OSINT work include:

- Exif viewer – find quickly if there in exif information in an online photo;
- Https everywhere – less and less relevant but forces all connections you make over https (secure) instead of over http;
- User agent switcher – allows you to choose how sites you visit see your system and browser. This may be helpful in obfuscating who you are, but also sites react different to different systems and browsers;
- Location guard – allows you to choose a location from which you appear to be visiting from
- RevEye – connect to four reverse search engines with one right-click
- CopyFish – OCR and translate text from photo and video
- Privacy badger – block ads/trackers
- uBlock origin – block scripts
- IP & DNS info – show IP and DNS info on a domain with one right-click;

[14] https://www.mozilla.org

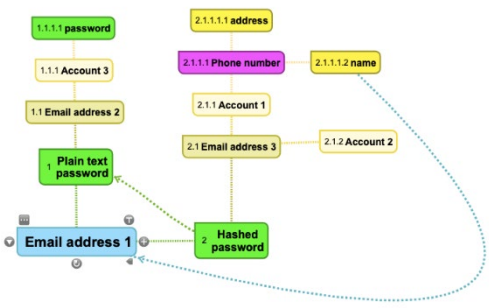
[15] https://www.google.com/chrome/

- Video downloaders – there are several extensions for downloading video from multiple sources. Handy to have one, however which one works best depends on system;

Additional tools

In addition to your browser, there are multiple tools that can be useful for research in open sources and analysing and presenting the data. Below we list briefly a number of tools that we suggest.

- TweetDeck – the free desktop Twitter application that helps you to structure the monitoring of your Twitter lists and or #keywords;
- Google Earth Pro – Finding and reviewing locations all around the globe with Google Earth Pro is easier than navigating Google maps, you can choose between different image dates (so see changes over time) and the interface is faster than Google Maps;
- KeePass(XC) – a password manager is a must to keep track of you passwords in a secure manner. See chapter 4.
- OpenRefine – When you collect load of quantitative data, the data cleaning usually takes most effort. Excel helps but also has limitations. OpenRefine works through your browser and has multiple smart options to clean and organise your data. See chapter 12.
- VPN – using a VPN add to your privacy (anonymity) and security. There are many providers and That One Privacy Guy compares most of them.[16]
- Maltego CE – The Community Edition (CE) of Maltego is a free but still quite functional version of Maltego and can be used both to collect data and make link-analysis schemes.
- Notepad – for many this small text-editor which comes standard with the Windows operating system (Linux and Mac have their own versions) is an unknown app. However, it has two very important functions we can use for OSINT work. First the app has a standard date/time stamp under the F5 key. Try it. This function helps you to very precise document what you did and when. Perhaps more important for crime investigators who have to show their whole chain of evidence, but still very useful for any other investigator. The second function is that all text pasted in notepad will immediately be completely stripped from all mark-up code. So, if you can to copy text from a document and use it as a selector in your searches, fist paste it in note pad, then copy it from there to the search box.
- Another inexpensive tool that can help driving jour investigations is a Mind map. Mind mapping is a creative and logical way of note-taking and that literally "maps out" your ideas.[17] Originally used for creative processes (‘brainstorming’) it has many uses and in online research it can help to not only easily record the steps you took but also to get the ‘overall’ picture where you are and what research angles still need attention.



[16] https://thatoneprivacysite.net/

[17] https://www.mindmapping.com/

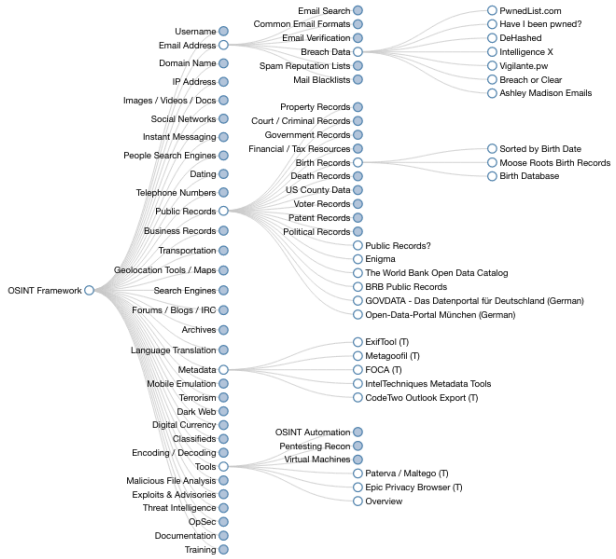
A great example source to see how a Mind Map could assist an investigation is Micah Hoffmann's OSINT Mind Map.[18]

Micah is the developer and main trainer of the 6 day live SANS SEC 487 OSINT course, at the moment the only structured and high level OSINT course available.

Building your link repository

As you go forward in online research you will probably compile your own link repository. This repository will contain the tried and tested links. Note that it takes time and effort to compile a good repository and just as much to keep it up to date. The links in this guide have been compiled and tested at the moment of writing in May 2019. However, probably in June 2019 the first links will already have been outdated. The sources may have altered their website and search syntax or may have disappeared altogether.

There is a number of link repositories that may for a good starting point and are regularly updated. We will discuss three.



The link repository which you can find at www.osintframework.com looks like a tree of categories. Every blue dot can be opened into a number of subordinate sources or further (sub)categories.

Once you have arrived at a white dot it has a link to the actual source which opens in a new tab.

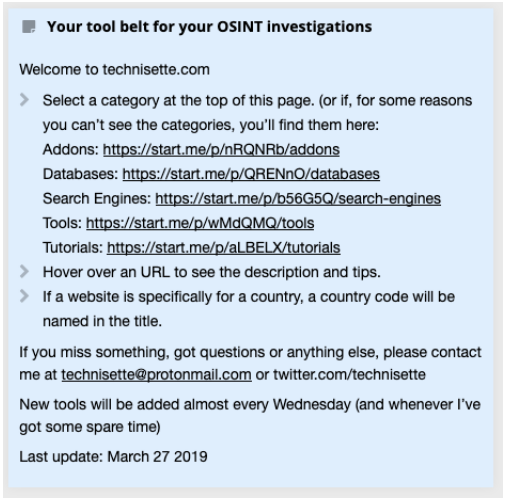
The OSINT framework is originally a GitHub project and is maintained by Justine Nordine (@jnordine)

[18] https://webbreacher.com/

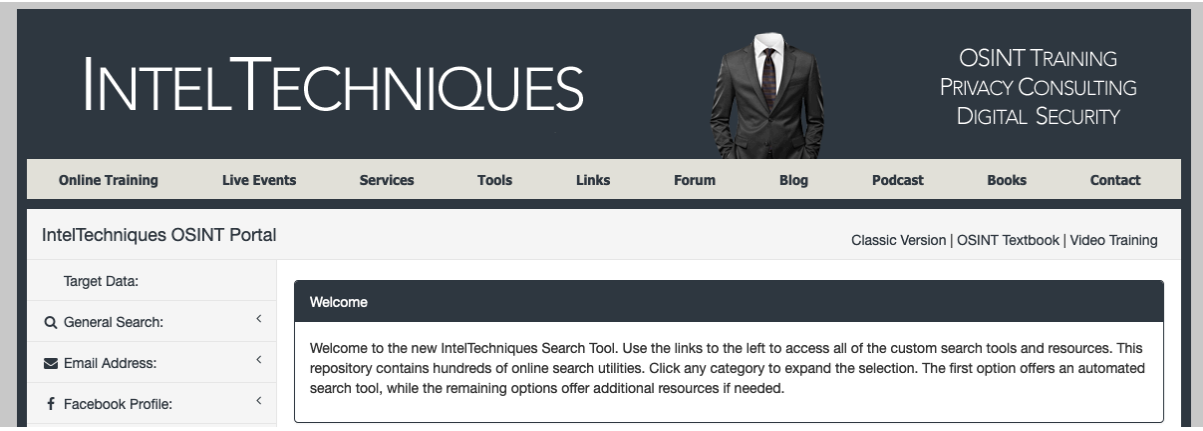
Another great repository is the repository available at http://technisette.com Technisette is a Dutch OSINT investigation, you can follow her on Twitter @technisette

Technisette keeps her repository at a start.me page with a number of subpages. Just take the time to browse through her pages and test some of the many, many links she presents.

Technisette is also one of the bloggers on https://osintcurio.us/ which is an interesting blog to follow if you want to stay up-to-date with the latest developments in the OSINT community.



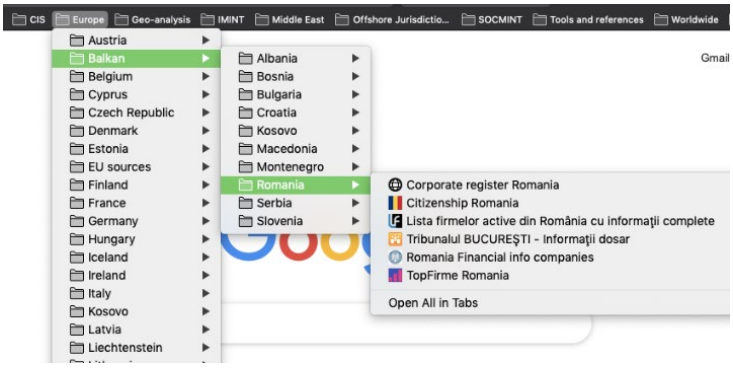
Lastly, we want to point out Michael Bazzell's IntelTechniques tool pages.



Actually, this page is much more than just a link repository, the tools on the page are preconfigured to be directly used in an online investigation. The tools are organised by type of data you're searching and each type has its sub-page.

Generally, the most used way to build your own repository of link is to keep them as bookmarks. Firefox uses a json database file to organize the links which can be exported as a single file. This file can then be easily exchanged and becked-up. Another advantage is that tags can be added to bookmarks in Firefox which given another way to organize your bookmarks. Note though that if you import a json bookmark file in Firefox that it will overwrite all older links.

How to organize links and bookmarks is usually a matter of personal preference. Some use geographic categorization, other functional or a combination of those. This will inevitably a trial-error process until you have what works best for you.



3. Documenting and archiving

In the previous chapter we already discussed some of the tools you can use to document the findings of your online research. In this chapter we look more at the methodology and in particular at archiving options. The reason behind that is that online resources have the nasty habit to disappear. Webpages are being taken down, media articles removed and database records altered.

The disappearance of information can have many reasons, some legitimate, others not, however if you are working as a journalist on a story for a few months or longer, you may want to make sure that the information you found is still available at the time you start to write. And of course, if your story is challenged you want to have solid proof that you indeed encountered these facts during your research

In this chapter therefore, we look at three documenting and archiving options: 1) in your own repository, 2) online and 3) in a shared cloud environment.

Documenting offline

The simplest way of documenting information found is to print is to a pdf file from your browser. This provides you with a simple indexable file that contains the relevant information. Make sure that in the printing settings of your browser you add the URL and a date/time stamp to the footer or header for further proof of the origin of the data.

Sometimes the printing to pdf does not result in a good representation of what your screen shows. In that case, take a screenshot instead. The downside of a screenshot is that it does not contain indexable text. However, if you have either the full version of Acrobat or one of the alternatives[19], you can create a pdf from your screenshot and perform Optical Character Recognition (OCR).

There are two aspects of off-line documenting that you somehow have to solve. The first is the security and availability of the documents. We assume that your laptop has full disk encryption. But do you perform regular back-ups? With regular we mean daily when you're working. And are the back-ups encrypted as well? Where are the back-ups? Not in the same location as your laptop, are they?

The second aspect that has to be solved are naming conventions. Do you have a standard convention on how you name your documents? For example, starting the name of every documents with date and initials of the creator like '2019XX0531" If you have no agreed naming convention, it can often be cumbersome to retrieve your findings. There are full-text search options in the different operating systems, however that does not always work.

Archive.org and other external archives

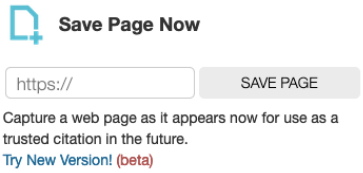
Even if you have the pages and data you found properly documented, you always run the risk that the authenticity of your documentation is disputed. Also when pages are not deleted on purpose, for example, studies have shown that over 10% of Internet references in scholarly articles were inactive after only 27 months. Having a copy of a webpage at an independent third party would solve this potential problem.

There are a few options for that, with the oldest being the so-called Internet Archive or 'WayBack Machine'. It was founded in May 1996 when it began to archive and preserve

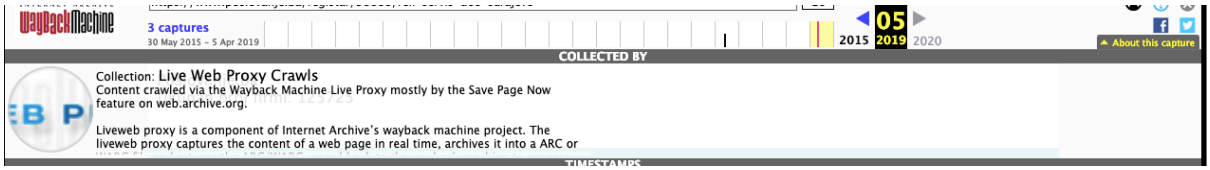
[19] https://pdf.wondershare.com/pdf-software-comparison/adobe-acrobat-alternative.html

World Wide Web. The archived content wasn't available to the general public until 2001, when it developed the WayBack Machine. Just like 'normal' search engines it crawls the internet and saves pages it encounters. We will discuss in chapter 5 its use for searching.

Because it is not feasible to crawl all pages, the WayBack machine has also an option to manually capture a page. There is a 'save page now' option[20] in which you can paste an URL which is then saved. After saving you are directed to the URL of the saved page which you can use for any future publication as a trusted citation.



Note that when you save a page in the Wayback Machine, the 'about this capture' information at archive.org shows that the page was not crawled but archived via the 'Save Page Now' feature. Hence showing human intervention which could be an Opsec issue depending on the circumstances.



There are other options to preserved the content of a webpage. One of these is WebCite[21] which is mainly aimed ad content preservation for academic use, however works for every URL. Simply go there add the URL and if you want, add some metadata as well. A short direct link to the content will be provided after preservation.

There have been other archives where you could manually add sources to be preserved (like archive.is, later renamed to archive.today) however none of them have been as consistently reliable as archive.org and websitation.org.

There is an extension for both Firefox and Chrome, named 'the archiver'[22] which lets you add URLs to the different archives with a right mouse click.

DocumentCloud

Another option is to document your findings in a shared online repository. Your news organisation may have one, however also consider DocumentCloud.[23] DocumentCloud is a service primarily for journalists where you can upload and annotate your documents, choose to make them public so you can share them with a link in your publication. In this way the authenticity of documents is less easy to dispute.

DocumentCloud also helps in linking your documents to other potentially relevant sources so for an investigative journalist this an indispensable tool to use. You have to apply for access and only journalists are accepted.

[20] https://web.archive.org/

[21] https://www.webcitation.org/archive

[22] https://www.cathalmcnally.com/tools/the-archiver/

[23] https://www.documentcloud.org/

4. Operational Security

Keeping your data and yourself secure is something that requires your continuous attentions. There are whole books available on that topic, in this guide we will touch upon some of the very basic stuff for online security.

Pseudo accounts

One of the first rules for open source investigations is that you never, ever, use your personal social media accounts. For most social media platforms an account is needed to (better) access the content so creating pseudo accounts is usually necessary. So, you create a 'pseudo account', also called 'sock puppet' or 'persona'.

Although there may be some legal issues with the creation of pseudo accounts, your security and the security of your data and investigation prevails over complying with the conditions for the use of platforms.

The most important question to ask yourself when you create a pseudo account is what are you going to do with it? This is because there generally are three types of sock puppets: generic, alter ego and investigation specific. We discuss them in more detail.

The first is the generic pseudo account and you use it to access (some) social media platforms only. For such accounts you set up a generic non-personalized email such as fggthht768@gmail.com and you fill the profile with only the minimum of information needed to satisfy registration requirements. So, no pictures, no posts, no engagement. These accounts are easy to set-up, and easy to drop if compromised.

The second is called the 'alter ego' and you use it to sign up for groups, databases, social media to research methods, obtain database records etc. It's you but with a different name so choose age and a background that you know about. Make sure you have matching e-mail addresses (multiple) and a burner phone so you can create a credible account that shows credible activity.

With this account you can engage at group level (e.g. FB & Linkedin Groups) but perhaps not at a personal level as you aim to have this account for the long run. Pro tip: do not engage in discussions about politics, religion etc., stay on the safe side.

Lastly, you can create investigative specific pseudo accounts to be used to engage with subjects in a specific investigation. Consult your friendly lawyer whether this is legal, proportional and acceptable in the situation you want to use it for. Know what you do before choosing this option.

Choose your persona well according to the context of the investigation and prepare a data sheet with all personal information of your persona before starting to create any accounts. Again, make sure you have matching e-mail addresses (multiple) and a burner phone. Create a credible account that shows credible activity but also make sure that the history checks out ('backstopping').

Engage at group level (e.g. FB & LinkedIn Groups) and establish 'active' account before engaging with the subject. It usually takes time before you have a persona that can be effectively used in an investigation. When you succeed, document every contact with subject.

Some general tips for the creation of an alter ego or investigation specific persona:

- Use http://www.fakenamegenerator.com/ for inspiration (but it is very US oriented);
- Prepare (non)distinct email address in advance, best to have two;
- Have a prepaid mobile number (burner phone) available in case verification is needed, switch to 2FA with an app asap;
- Photos are an issue -> you do not want to be impersonating someone, stock photos are a risk, so use photos from thispersondoesnotexist.com;
- Work from a clean environment so avoid mistakes;
- Construct a plausible resume / life history where age matches position/role/status, the working history is realistic and consistent, the studies fit career and really exist and in general information is consistent. Prepare to be quizzed about things.
- Carefully review the privacy setting of the profiles: expect to be looked up, what do you want them to see?
- Go slow on getting connections and watch out for LinkedIn and Facebook security algorithms -> too many invites / rejections may result in a block;
- Consider to upload 'your' address book at some point to appear are relevant for your target;
- Start with following, joining in with conversations, joining groups, play games and finally start collecting friends;
- And: use a password manager for the different (!!) passwords.

Password use and 2FA

In OSINT we make lots of use of data from data breaches (see chapter 7) which shows us how careless people are by choosing simple passwords, recycling their passwords over many sites and with multiple logins.

If you take OSINT work serious, you have different strong password for each platform /account /website. This also applies to the password for your persona's because you do not want to appear in data breaches with easy to track usernames and passwords.

A password manager like KeePassXC[24] allows you to store all needed credentials in a secure way (so NOT in your browser…..). And it allows you to easily generate unique long passwords which you do not have to remember.

Further where possible enable two-factor-authentication (2FA) for example with Authy[25] so no one can lock you out of your account with an easy password reset. Try to avoid 2FA by SMS as this is less secure due to the risk of SIM swapping.[26] Although, 2FA via SMS is still better than no 2FA at all.

A clean environment

When we discussed what system you can choose for your OSINT work, we already touched upon Virtual Machines. Most professional OSINT and security researchers that collect a lot of data online, do so from a VM.

Basically, a VM is a software programme which emulates a full computer system. VM's come in all sizes and shapes, much used in larger IT environments and often completely configured for a single purpose. However, also on your laptop / desktop you can run a VM

[24] https://keepassxc.org/

[25] https://authy.com/

[26] https://medium.com/@nickbilogorskiy/sim-swapping-7f1725ae0d23

and performing your OSINT work from within that 'machine' which has three main advantages:

- You work from a different environment and also when you connect to the internet from, for example, your Linux VM on your Mac, for the outside world it looks like you're really on a Linux machine;
- There is an extra layer of security between your OSINT environment in the VM and the data on your actual machine, and;
- If your VM gets infected with malware you can just delete it and start from a fresh copy. Or in fact if your results may be needed for court procedures, you can start from a fresh VM for every new investigation in order to avoid any type of contamination of the results.

The best known VM for OSINT work is Buscador[27] but any VM will do. The costs are limited to the license for VM player.[28]

IRL operational security

Many information leaks still happen 'in real life' (IRL) meaning not online, but in public spaces. The best firewalls and computer security suites will not help, if information is continuously disclosed outside of hardened IT-environments by careless employees, who loose USB flash drives. And this happens all the time.[29] Or information is lost in very mundane way, when a computer crashes and no (recent) backup is available, something that also happens daily.

No matter how good your cyber security measures are, the most important aspect is to security serious at all times, also when not behind a screen.

5. Search engines

General notes on searching

Before we dive into how to make best use of search engines, we first discuss a number of tips and tricks on searching in general.

One of the important things you need to remember is that every little detail can be a lead for a new search from a different perspective. In open sources research these are called 'pivot points'. Suppose we are looking for information on the assets of 'John Smith' but we cannot find anything relevant. We have found an address that he used but it turns out that he does not own the place. However, that address can be a pivot point because we can look for anything related to that address which may have a link to John but not directly. For example, we can look at (other) inhabitants of the address and see if we can link them to John. Or we look at any companies (previously) registered to the address to see if these maybe are owned by John. When 'pivoting', the only limitation is your lack of creativity.

Another important concept is 'context'. If you are really searching for a 'John Smith', you may never find 'your' John Smith if you do not have some more context which you can use to narrow your searches and corroborate findings. Context can be anything like place of birth, current place of residence, name of spouse and kids, unit of military services, high school and

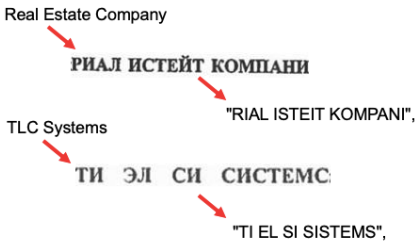
[27] https://inteltechniques.com/buscador/

[28] https://www.vmware.com/products/personal-desktop-virtualization.html

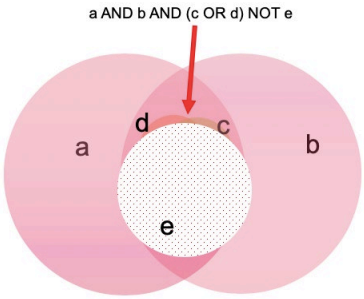
[29] https://keyfindings.blog/2019/03/26/intelligence-collection-on-the-train/

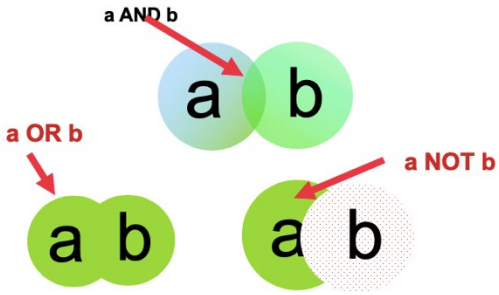
college where he went to, hobbies, previous employers, anything that helps you to identify the correct 'John' and to understand the data you retrieve on him.

Then language issues often complicate searches and the validation of findings. In chapter 2 we discussed CopyFish as a tool, but also Google Translate[30] or DeepL[31] are useful tools. Of course, translations between different language scripts (Latin, Cyrillic, Arabic, Chinese, etc.) may cause confusion and errors. Look at how these company names were transcribed from English to Russian, back to English. This is a real example from a company register and it took some time to find the actual company.



A last important aspect of searching is the understanding of basic Boolean operators, AND, OR and NOT. The AND operator narrows the search by retrieving only records containing both (or more) keywords used in the search statement. The OR operator broadens your search by retrieving either one or more of the keywords used in the search statement. The NOT operator excludes records containing the second (or third etc.) keyword in your search statement. We have attempted to show this in the following graphs.





Google

The key feature of Google for OSINT work is that it indexes websites so that we can search a specific term in the Google index and get as a result a link to the website(s) where this term has been found by Google. In theory that index should be able to be queried using Boolean operators. However, with Google (and other search engines) it does not work that straightforward.

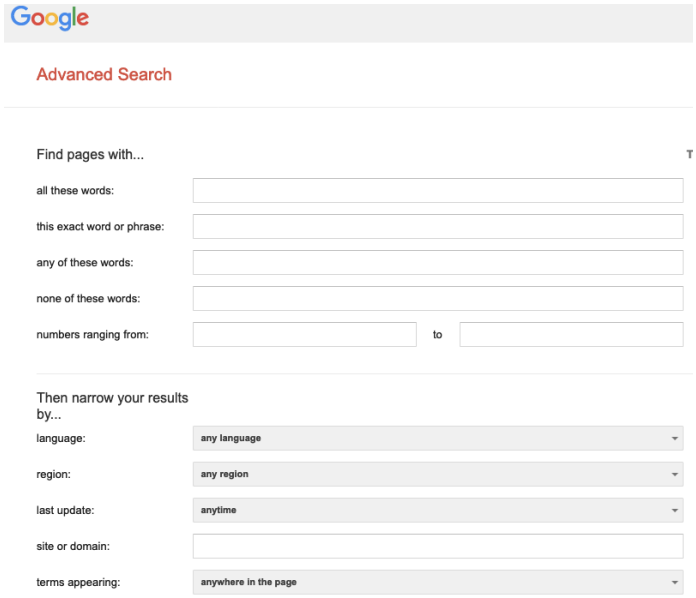
Simply said, Google is not a search engine built for investigators searching specific items on the internet. Google is a marketing machine and many decisions on the functionalities of the search engine are related to how the average user (consumer) is using the machine, how he can be fed advertisements and how his search history data can be sold to highest bidder.

You, as a consumer, are the product of Google because your browsing habits are relevant for the Google advertisers. And the results you see in the index made specific for you. Your results depend on, amongst other things, whether you have your cookies visible for Google, on whether you are logged in or not, from what country and IP you searched. Even the order of words may cause differences in the results.

[30] https://translate.google.com/?

[31] https://www.deepl.com/translator

As a professional searcher you have to take these limitations of Google into account and understand that Google does not strictly apply Boolean logic, it is more 'Booleish' logic. However, Google has created its own strong other search features. A very easy approach is using the Google Advances Search box.[32]



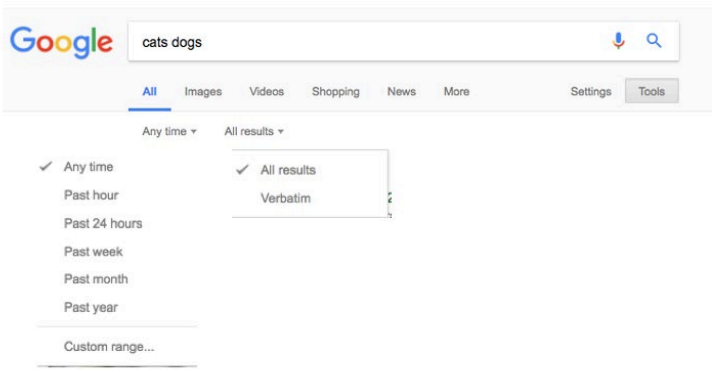
Basically, this search box helps you build a search query in a much easier way than writing it in Boolean language.

The selection possibilities also show that Google has a number of specific operators that you can use to narrow your results. The most used are:

- site: - limits your results to a specific domain, like 'site:linkedin.com'
- filetype: - limits your results to specific filetypes like 'filetype:pdf'
- intitle: - limits your results to the items where your keyword is in the page title
- inurl: - limits your results to the items where your keyword is in the ulr, for example when you are searching profiles, you could use 'inurl:profile'

Using smart combinations of the different search operators in Google is called 'Google dorking'[33] and can result in very specific searches which sometimes results in unexpected results.

In your results Google has additional possibilities to narrow down search results though the menu that appears under your search box:



Here you can choose to further filter on time, type or verbatim (=without synonyms).

One note on the number of hits you see when you have executed your search. Generally, this number does not mean a thing and can be different every time you search. To see the actual number of potentially relevant hits on your search, go to the very last pages of you results until you cannot go further. There you will see the actual number of hits.

[32] https://www.google.com/advanced_search

[33] See for an overview of many possibilities https://nfsec.pl/media/ghdb.pdf

So, although Google in essence is not a search engine built for professional use, if you understand its limitations and Google wisely, using the advanced search box and the Google operators, a lot of data can be found with Google.

Other search engines

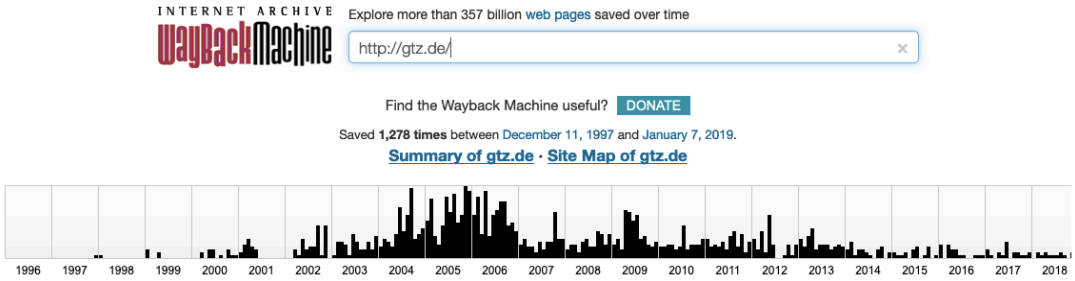
There are many other search engines besides Google such as Yandex, Bing, Yahoo and many more.[34] For general searches we advise always to look at least at two sources in addition to Google especially when Google gives no results at all.

In addition to these giants under the search engines, many smaller search engines exist. These are sometimes regionally focussed, sometimes privacy focussed like DuckDuckGo[35] or with a specific format. For personal use you may prefer something else than Google, however for OSINT work, Google remains the search engine to go to, especially if you understand its limitations.

Also, a number of very specific search engines exist, for example Proisk[36] which has indexed the content of open ftp servers and Shodan[37] which searches the Internet of Things (IoT) devices, such as webcams and databases connected to the internet. A discussion on the application of these search engines however goes beyond the purpose of this guide.

Wayback Machine

We discussed the Internet Archive in chapter 3 under documenting your searches. Obviously, the WayBack machine also functions as an important research tool. In essence it is a search engine, an archive of the entire internet. It visits, in theory, every website and crawls it while taking screenshots and logging the data to a database. These endpoints can then be queried to pull down every instant the WayBack machine has ever crawled in time as shown below:



You can search for the root of the domain or for any pages or subdomain. Just add the URL which appears to be missing in the search box and see if it has been captured at some point in time.

The basis search the WayBack machine can be found at https://web.archive.org/ where you can see if a specific domain or URL was captured. There is also an advanced, lesser known, option which can be entered via https://archive.org/search.php Here you can search through the archive by key word. Note however that this search is not reliable (yet).

---

[34] See https://marcodiversi.com/blog/alternative-search-engines/ for an overview.

[35] https://duckduckgo.com/

[36] https://proisk.com/

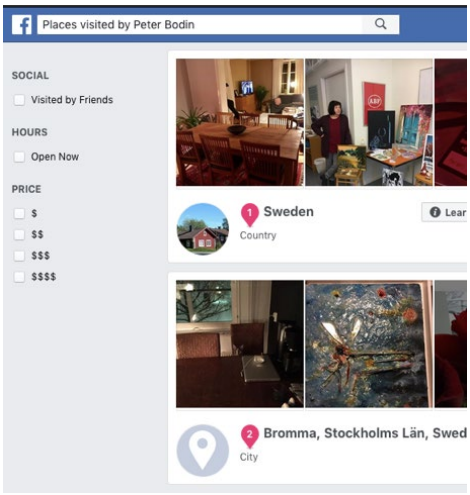
[37] https://www.shodan.io/

6. Social media

Before we discuss some of the OSINT applications of the main social media platforms, first a (repeated) word of caution. Social Media platforms generally require an account to be able to use the (full) search functionality. For your own security, never ever use your personal social media accounts to search anything work related. Even though your subject may not (directly) be able to see that you searched for him/her, the platform will know as all activity on these platforms is logged and somehow used for their (commercial) purposes. So please always use a pseudonym ('persona') as explained elsewhere in this guide.

Facebook

Facebook has been the go-to source for many OSINT researchers over the past decade mainly because people spill so much personal data on Facebook that the platform could be considered as a treasure trove. Although to some extent that is still the case, there have been some changes over the past few years that have limited results to be expected from searches on Facebook.



Not only have people become much more privacy conscious and set their pages to private, also Facebook has changed (and is constantly changing) its search syntax and the way you are able to perform searches. Previously you could just type in what you wanted to know, like 'places visited by John Smith'.

That was a very powerful search but it does not work anymore for the normal user. It is still possible if you know the exact URL and by manipulating the URL you still can get the results.

The key then is to understand how these search URLs are being built by Facebook. The problem with that is that the syntax is changing very often. Currently, if you want to know the places visited by a Facebook user (assuming that the user 'checked in' there) the syntax would be: https://www.facebook.com/search/usernumber/places-visited/

In all searches the user number or location number is key. Users and locations are given a number based in which the search can be built. For example, if you know that 111980668819694 is the location number of Ljubljana, the following search string gives you all photos with the location tagged as Ljubljana posted between 1 January and 19 April 2019:

https://www.facebook.com/search/111980668819694/photos-of/1/jan/2019/date-3/19/apr/2019/date-3/photos-2/intersect

NB: the Facebook settings should be on US English to make these searches work.

As you can see understanding the exact syntax takes a lot of practices. Fortunately, there is a number of pages that can be used to find the user numbers and to build the searches. The most useful at the moment are:

https://inteltechniques.com/OSINT/facebook.html

http://graph.tips/beta/

https://whopostedwhat.com/

https://stalkscan.com/

Each of these pages have a bit of a different approach but eventually they use the same way to build the search syntax. Because Facebook constantly changes the search syntax and the authors of these pages may not immediately update their page when a syntax isn't working anymore. Therefore, always try multiple options if one of the pages does not give any results.

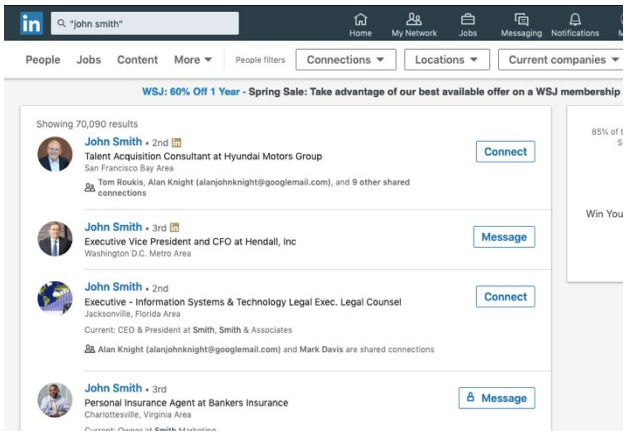
And what if your target does not have a Facebook profile? Well, still you may find him or her in photos by his (close) circle of friends and family. That takes of course more time to research, however it can certainly pays-off.[38]

LinkedIn

Another much used social media platform is LinkedIn which from an OSINT perspective mostly will be relevant to establish connections between individuals and companies and less relevant for the actual posts.

Searching on LinkedIn provides varying results and is by design limited in order to seduce you to take a Premium subscription which gives you much more search options, flexibility and results.

With a Premium account there are more search filters and operators to use although recently also the search operators on LinkedIn stopped working.[39]



LinkedIn has a search box in the top left and once you have searched for something (usually a persons or company name) you are provided with the results and some filters appear.

You can then filter the results for only those connected, on certain locations or in certain companies.

One important way around the limitation of LinkedIn, is what recruiters call 'X-Raying'.[40]

X-Raying is using Google to search one specific domain, in this case LinkedIn. In Google, your search would look something like:

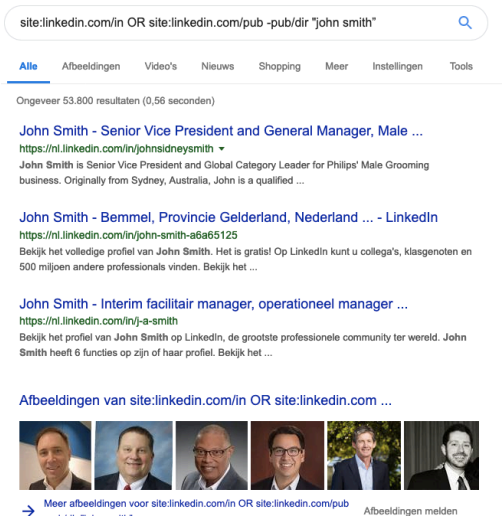
"site:linkedin.com/in OR site:linkedin.com/pub -pub/dir [name]"

[38] See https://keyfindings.blog/2019/04/12/red-flags-on-other-peoples-facebook/

[39] See https://booleanstrings.com/2019/04/19/update-on-linkedin-people-search/

[40] See https://booleanstrings.com/2018/09/04/did-you-notice-new-way-to-x-ray-linkedin/

In normal language: you would be searching in the linkedin.com domain only and in the directories /in and /pub with exception of the subdirectory pub/dir and you would be looking for 'name'.



The added benefit of X-Raying, is that you can add unlimited additional selectors and Google search operators in you search.

This search provides you with all the results on profile pages (as indexed by Google) where the name appears, also when it was in the column 'people also viewed'.

Those pages would not be the profile of your target but shows anyone else who also looked at the profile you just found. That is not evidence of anything and may be completely irrelevant, however there could be some link in a way you may not have thought about previously.

Once you have the desired result, you can see a profile overview, shared connections and recent activity of your target and then visit the profile.

OPSEC warning: Note that if you visit a profile on LinkedIn, your target can see that someone visited his profile. So never use your own account but your investigative pseudo and for general use make sure the account settings for the LinkedIn account of your pseudo are set on completely private (in certain cases you may want to attract the attention of the target).

Instagram

Mostly popular by the generations younger than Facebook users, Instagram has grown from a purely photo sharing platform to a full-fledged social media app with chats, advertising, video etc. Instagram has been bought by Facebook in 2012 and slowly the platforms are integrating.

Still Instagram remains primarily a photo / video sharing platform where the shared photos and video make up the wall of the users. A different feature of Instagram is the so-called stories option where a user can share videos or selection of photos for 24 hours and more recently there is also an option to do live video on Instagram.

Instagram has been an interesting source especially for asset tracing, with the OCCRPs' 'The Secret of the St. Princess Olga' story[41] as an interesting showcase of how Instagram pictures can be used in a (journalistic) investigation. Like on most social media platforms, less and less people show their location when posting something, but still people always share more than the realize.

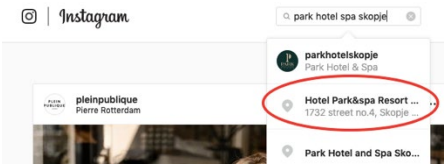
When you want to use Instagram via the platform website[42] you need to have an account. Sock puppet rules apply (see chapter 4), however Instagram is less strict in their account creation policy.

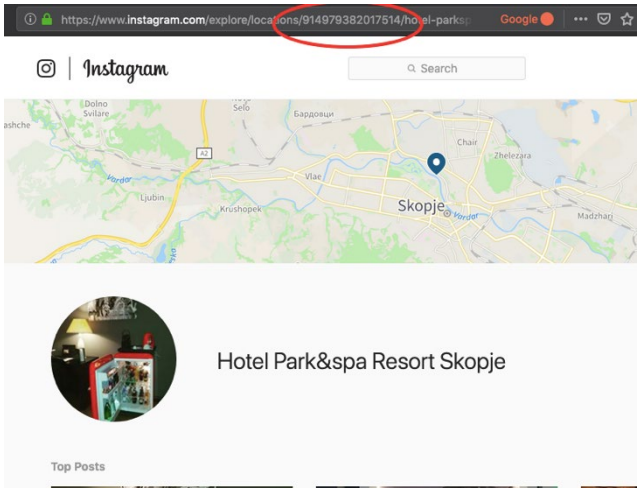
The key elements in Instagram are accounts with usernames starting with an '@', tags starting with a '#' and locations. On Instagram, every location has a fixed formula that

[41] See https://www.occrp.org/en/investigations/5523-the-secret-of-the-st-princess-olga

[42] https://www.instagram.com/

begins with: https://www.instagram.com/explore/locations/ and is followed by a number. You can find the location by searching for it and click on the result with the location icon. As a result, you will get all posts tagged with that location. Remember, the location is what the user added as location, it is not a verified location.





The URL of the result will give the location ID. Note that location ID is the same on Facebook location ID.

The downside of the Instagram website is that the search possibilities are limited. There is no way to combine search parameters or even use more than one hashtag.

There is however a number of websites that provide easier search facilities into the Instagram data. An easy web source to search and view Instagram accounts is Pikram.[43] The downside of this source it that it does show a lot of annoying advertisements so you may want to activate the adblocker for a moment. A similar tool with less annoying advertisements (there still are some) is the Instagram Online Web Viewer.[44] If you enter a keyword in the search bar you will get users, tags and locations that contain that keyword which you can easily select to drill down.

Two other useful sources are Search My Bio[45] through which you can search in Instagram bio's and the Instagram Online Web Viewer[46] lets you search for a topic and retrieves a nice overview of tags and accounts.

When you research the interactions via Instagram, Gramfly[47] is a helpful tool. After typing in the username, it provides the most recent and top mentions, locations and hashtags. Although it provides an overview of the users with whom the account had recent interactions, it does not provide any actual comments; this is some that just takes painstaking research.

Downloading of full-size photos from Instagram, also the profile pictures, can be done with InstaDP[48] and for any research into images this is highly recommended.

In contrast to what many think, also stories - even when not visible anymore on the users' profile - can often be retrieved. The tool, Storiesig[49] lets you type in the user name (no need

[43] https://pikram.com
[44] https://www.inst4gram.com/
[45] https://www.searchmy.bio/
[46] https://www.inst4gram.com/
[47] https://gramfly.com
[48] https://instadp.org/
[49] https://storiesig.com

to log in or be logged in) and then scroll through the stories and download what you need. Tests we did show that the success rate is not 100%

Finally, there is a paid source, Picodash[50] that for a fee lets you download large amounts of Instagram metadata into csv files for further analysis. Generally, this service is more relevant for marketing professional and less for journalists and investigators.

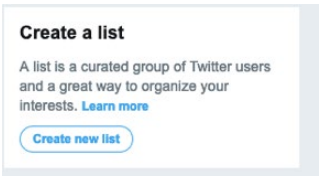
Twitter

Twitter is a 280 character (was 140 character) 'micro blogging' platform where the posts are called 'tweets'. Although the content of tweets generally tends to be more opinion than fact, some topics are dominated by bots (automated Twitter accounts) and include a lot of fabrications, there are some good investigative research use-cases for Twitter.

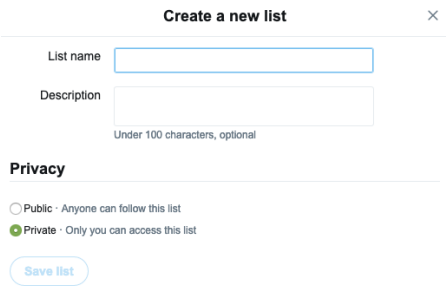
You can use all kind of search operators in the search box of Twitter, just like in Google to find specific tweets on a subject, from a specific user, in a date range or even from a specific location. An overview of the operators can be found on the Twitter support pages[51], however, Twitter has fairly decent advanced search facilities.[52]

Remember to always verify any tweets you find possibly relevant. Is the content re-tweeted or a screenshot? Who tweeted first, who retweeted, check profile and handle, is it all verifiable and consistent? Be careful with opening links from tweets, preferably in a safe environment.

To analyse accounts there is number of websites available, Quick analysis can be done with FollerMe[53] and more deep analysis of an account with Tweetbeaver.[54] If you look at a specific tweet and relations, Social Bearing[55] can be useful. For visual searching on location use One Million Tweet Map[56] or QTRTweets.[57]



Since it is largely an opinion voicing tool, following certain politicians or certain subjects over time can give an insight in political developments and connections between people on the same subject. NodeXL[58] can help retrieve and analyse larger amounts of tweets.



For monitoring of groups of people, the list tool of Twitter itself is very useful and you can create a private list so people do not get notified that you are following them. There is a good tutorial on how to research the most relevant people for a specific topic here.[59]

---

[50] https://www.picodash.com

[51] https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/standard-operators.html

[52] https://twitter.com/search-advanced

[53] https://foller.me/

[54] https://tweetbeaver.com

[55] https://socialbearing.com

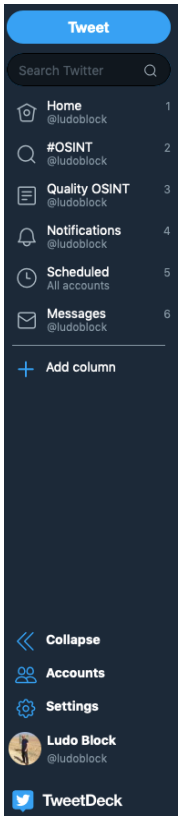
[56] https://onemilliontweetmap.com

[57] http://qtrtweets.com/twitter/

[58] https://www.smrfoundation.org/nodexl/

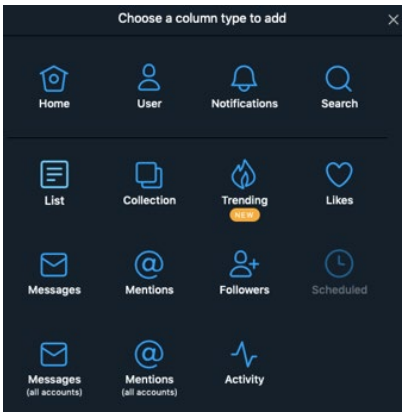
[59] See https://www.linkedin.com/pulse/20140702064329-435117-how-to-track-a-company-or-subject-using-a-private-twitter-list/

The list can be created on the Twitter website in the platform or in the TweetDeck application on your computer.



TweetDeck is the native Twitter application for your computer and by far the easiest application to organise your Twitter feed in. You can create dozens of columns for specific subjects so in fact you will have multiple timelines you can choose from depending on what you are researching.

If you want to define one column for a list, add a column and then select 'List' after which you can choose to either have an existing list in that new column or create a new list. From there the creation of a new list works exactly as on the Twitter website.

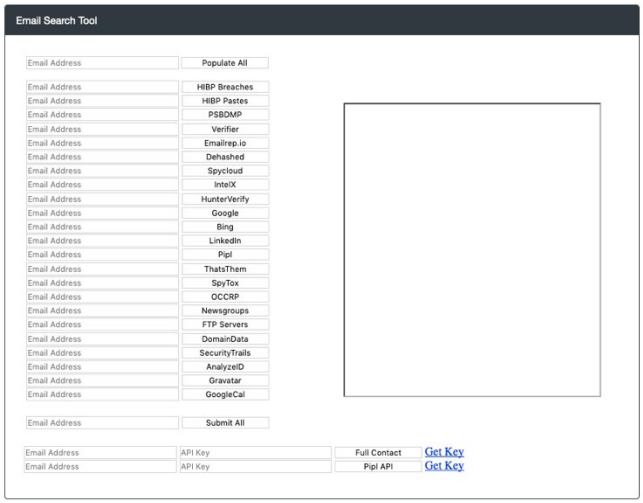


7. People searching

When searching for specific individuals, the best starting points are the 'identifiers' that people use for communication and when connecting to online services. These are email addresses, user names and phone numbers. The reason why we use identifiers as a starting point is that names of individuals are usually not unique, however, email addresses, phone numbers and (to a lesser extent) usernames are.

In this chapter we will therefore focus on how to validate these identifiers and how to link them to the actual individual.

Email



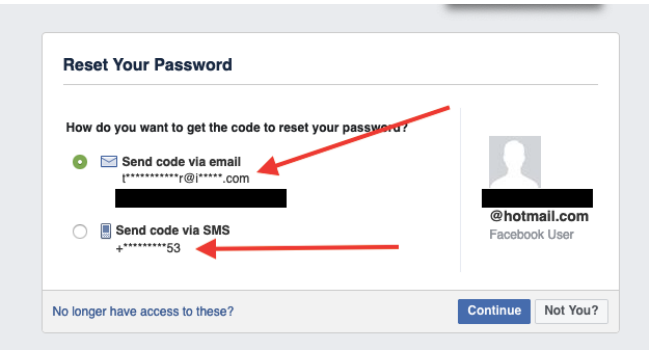
Email addresses are usually the best starting point for investigating individuals. If I know that my subject's name is John Wilson that does not get me very far given the number of people with the name 'John Wilson' out there. However, if I know he uses the email address john.wilson123@gmail.com that information provides me with a solid starting point.

The website inteltechniques.com has

most of the tools you need under tools Email Search Tool page. By simply pasting the email address into one (or all) of the search boxes, the tools help you to reach these pages and see if there is an account associated with the email address you entered.

If there is an account you can use the content of the account to verify if this is really the subject of your investigation. If it is, you have verified the email address and may have gained additional information in the process.

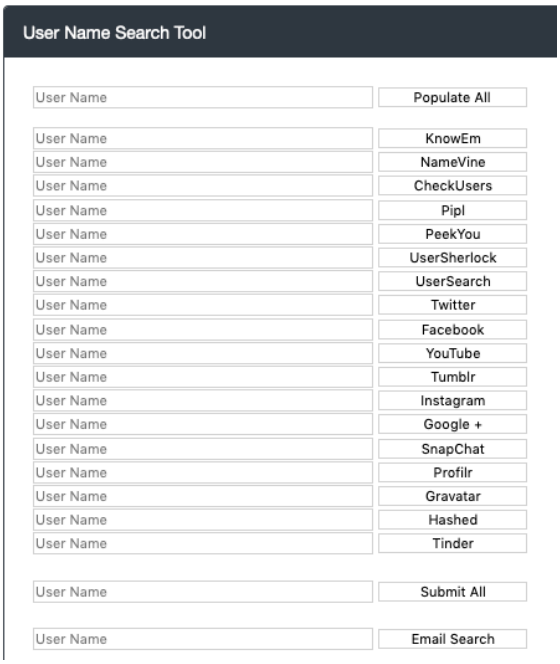
Note that the page does not have an option for checking the existence of Facebook profiles based on email addresses. This is (currently still) possible via the URL https://www.facebook.com/login/identify/ If you enter an email address you either get 'No Search Results' which means that the address is unknown at Facebook.



However, when the address is known you are prompted whether you want to reset your password and parts of a mail address and/or phone number are revealed. In the example depicted right, you see that the query for a Hotmail address returned another address (although not complete) and the last two digits of a phone number. Depending on the privacy setting of the account, you could also see a photo. Not only have you validated the first email address as being used for Facebook, also the new information are again pivot points for further investigation.

Make sure that after making a screenshot you select 'Not You?' you avoid that your target receives an alert!

Usernames



Very often email addresses, or the first part of an email address, are used as the user name for many social media platforms and login for other sites. Sometimes it is different and we find separate user names. There are some good tools verify usernames and see if these are in use on any platforms. An overview of the tools is again available at the Inteltechniques tools page.

Know'em is in fact a tool to find available social sites for a certain username. It is in essence a marketing tool that checks whether a username is still available. We use it reversely: if the tool shows that it is still available, that is not interesting for us. On the contrary, we want the platforms where it is not available anymore because that is an indication that our target may have an account there.

There are some more tools but the most relevant can all be found on the IntelTechniques page.

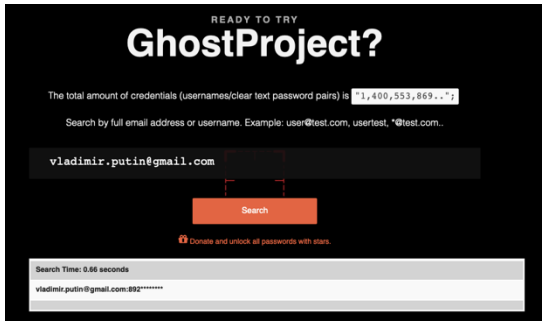
A specific interesting one is Pipl.com. This is a so-called 'people search' provider or 'people search engine'. While most of these people search engines usually are US oriented, Pipl.com has datasets from all over the world and it combines social media with other type of data. Usually it pays off to check a username (or name, or email address) in Pipl.com

Breach data

A last option we discuss on email research is breach data. As you know a lot of data breaches occur where personal information such as user names, email addresses and password are being stolen and made public after some time. The data from these breaches in the end finds its way to open sites like Pastebin.[60]

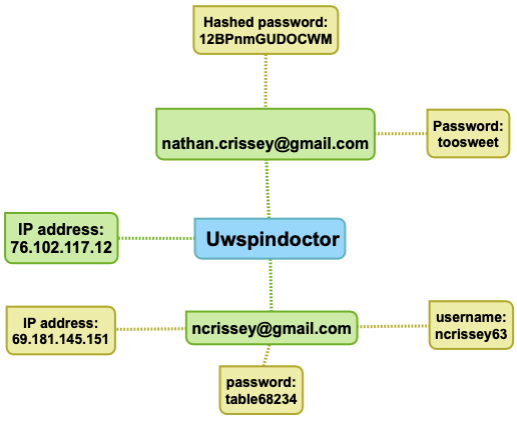
This 'breach data' is incredibly useful for OSINT investigations in many ways:

- it helps to validate email addresses -> has an email indeed been in use;
- it helps to identify platforms where your target has been active on;
- it could help to connect user names to email addresses;
- the passwords (or the hashes of the password) which can be searched as well, help to link different email addressed together.



While it is entirely possible to look for these breaches yourself, collect the data and build a database, for basic OSINT use it is much more efficient to use the services offered through for example GhostProject[61] and Dehashed[62] which are the two most used.[63]

Both offer a free search services but for the full results a small fee needs to be paid. At a live demonstration during the training in Skpje in May, we started with the username 'Uwspindoctor' randomly obtained from untapped.com and we were able, based on breach data alone to find the information as shown in the adjacent Mind Map.



If we would 'pivot' to Pipl.com with this data we would very quickly be able to obtain even more information.

In case passwords are not yet de-hased, on the Inteltechniques website also a hash/dehash tool is available. We can use this to quickly either hash a password and search in Dehashed with the hash instead of the password, or see whether a hash found in Dehashed may be turned into the plaintext password.

---

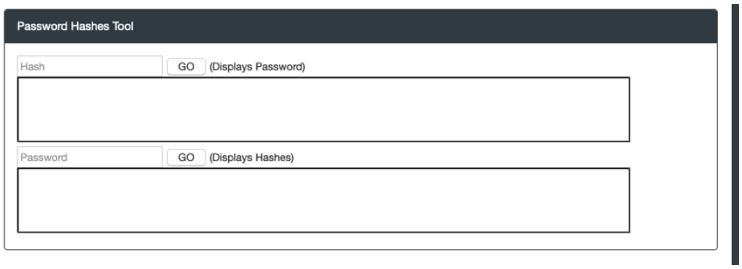
[60] https://pastebin.com/

[61] https://ghostproject.fr/

[62] https://dehashed.com/

[63] SpyCloud is the professional provides with much more data available but an investigative subscription is extremely expensive.

As said, of course, you could go and look for yourself to find the data from data-breaches online. Often these are freely available if you know where to search. However, there are security risks involved, and you need to have some background in working with databases and of course lots of diskspace. There are good tutorials available[64] but beware that this is way beyond the 'basics' of OSINT.

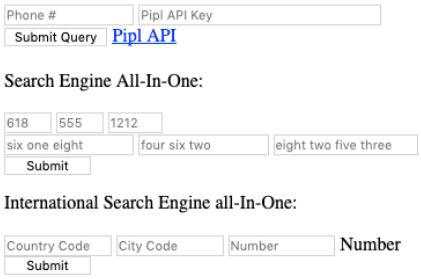


Phone numbers

A last identifier which can be used to search for individuals are phone numbers. Like with email addresses, people tend not to switch much from phone number. This is also the reason that marketing companies use phone numbers as the key identifying piece of data and all apps and platforms ask for your phone number.

Searching by phone number in a search engine not often results in any findings. One of the reasons is that there are many ways to note a phone number, including spaces (most used in Europe) or dashes (often used in the US).

The Inteltechniques pages has a number of sources, mostly US oriented but some international. In practice Pipl.com turns out to be a very good source. You can get an API key and then pay 10 to 40 US cent per match. As with the other data, Pipl often has relevant data outside the US as well.



A somewhat tricky way of researching phone numbers is the use of Phone contact book apps such as Get Contact[65]and TrueCaller.[66] The way these apps work is that the need access to your contact book and then send all this information to their database. Of course, this violates the privacy of yourself and all your contacts but there are still people dumb enough to do so.

You can use this technique but with a burner phone only. On the Bellingcat site there is a good tutorial on how it works, what the differences between the apps are and what results you get.[67]

truecaller

[64] See for example https://osintcurio.us/2019/05/21/basics-of-breach-data/

[65] https://www.getcontact.com/

[66] https://www.truecaller.com/

[67] https://www.bellingcat.com/resources/how-tos/2019/04/08/using-phone-contact-book-apps-for-digital-research/

8. Image verification and geolocation

Often images of (current) events are available via social media. While the availability of this so-called 'User Generated Content' or 'UGC' can be very beneficial for journalists, verification of these images is of course always needed in order to identify and exclude fake or falsified content.

At the same time holiday pictures of your target on an undisclosed location could be used to piece together a good story as well. Social Media platforms nowadays scrub all uploaded pictures from the exif data so the location indicated (if any) is what the user wanted to show, not necessarily the real location. This chapter therefore aims at providing the basic tools and methodology to analyse and verify images and to geolocate where thee have been taken.

First inspection

The very first task when verifying imagery is to visually inspect the photo. What do you see, describe it in words. What landmarks are visible? What text is visible? What is out of place? What does the weather look like[68]?

If there is text in the image or video in a language that you do not understand, try the CopyFish OCR and translation plugin which is available for Firefox and Chrome.[69]

Very often a thorough visual inspection provides enough clues from further investigation especially when combined with the context in which the image was obtained, like for example accompanying text.

In addition to a visual inspection, also subject each image to a technical inspection. This should include checking the meta data (Exif) data which could show date, lime, location, devices) and can be checked online with Jeffreys Exif tool.[70] There are also apps available for on your system.

Also examine whether the image was altered for example with Image Edited?[71] or FotoForensics.[72]

Google, Yandex, and Bing reverse image searching

All major search engines have the option to perform a reverse image search and they provide different results.

Google[73] is the most used has a number of additional option that Yandex and Bing lack. This is the possibility to add additional keywords to the search could give immediately different results.

[68] Which you can match at https://www.wolframalpha.com/examples/science-and-technology/weather-and-meteorology/ where you can find the weather at that date and location.

[69] For Firefox: https://addons.mozilla.org/en-US/firefox/addon/copyfish-ocr-software/

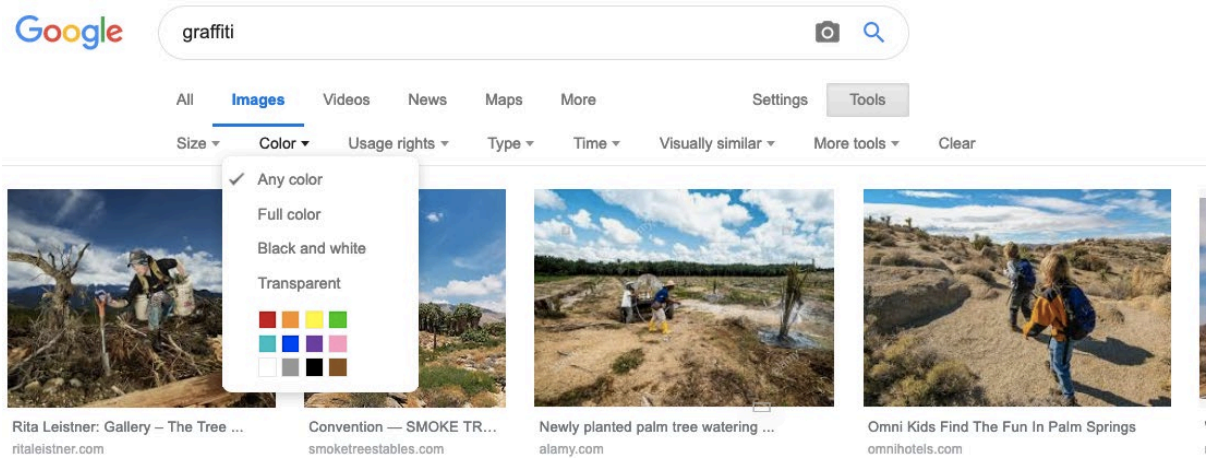
[70] http://exif.regex.info/exif.cgi

[71] http://imageedited.com/

[72] https://fotoforensics.com/

[73] https://www.google.nl/imghp

This feature suggests that Google tries to translate the image into words and then performs a sematic search. Also, it is possible to filter on colour if you encounter many similar images.



Be sure to use the image search facility of Google by searching just on keywords. Very often this gives relevant results.

Yandex[74] reverse image searching does not have such filters, however it very often is better at recognising locations that Google. Bing[75] again shows a different approach and currently has the least applicability (but that can change). Then there is TinEye[76], the site that more of less started reverse image searching but also this site seems much less effective than Google and Yandex.

If you cannot find the location but if there are specific landmarks or features on an image, crop the image to that landmark and search again.

If you still cannot find it, remember that reverse searching only works if the picture, or one that very much looks like it, is already indexed by one of the search giants. If not, you may not find it (although Yandex is scary effective) and you need another approach.

That means going back to the visual inspection and start making search hypotheses based on the clues that the picture provides. What do you see what can help you locate it? Architecture, vegetation, humans, climate? Try to imagine in which country you are? Can you find some characteristics? Can you compare physical landmarks, structures, terrain and vegetation within the image against known constants such as satellite imagery?

This could be painstaking work and once you have a good idea in which country it should be, try to obtain assistance from someone local who often very quickly knows where to look.

Meanwhile there is a number of tools that can help determine the exact location.

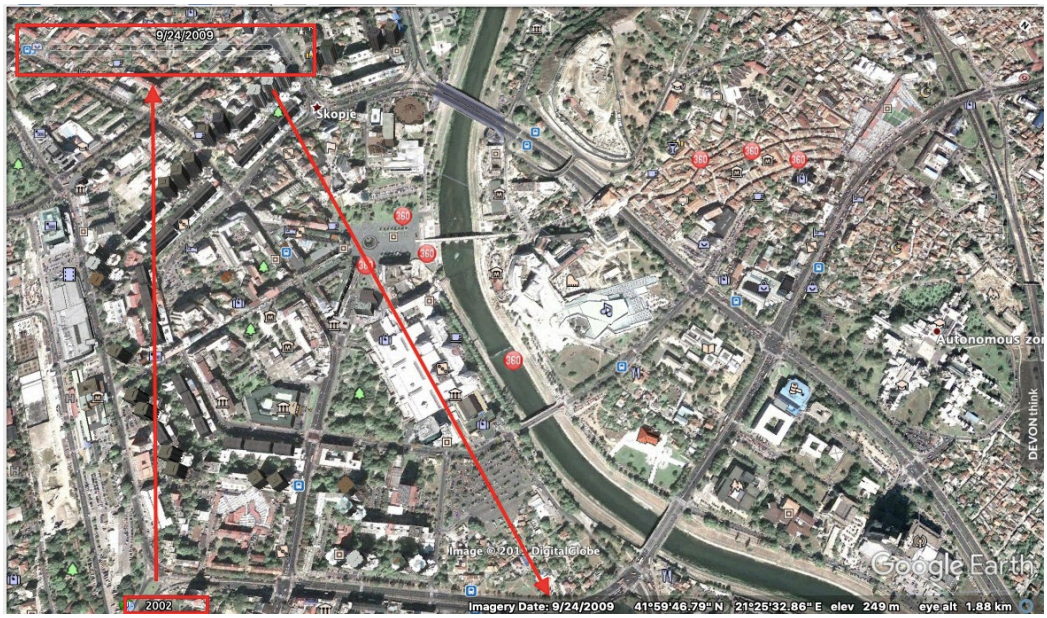
Google maps and auxiliary tools

While Google Maps (online) already has most useful features, there are two additional tools which are very useful. The first is the Google Earth desktop app which not only often provides newer imagery, but also has a date slider which allows easy switching between the different dates of the available satellite imagery.

[74] https://yandex.com/images/

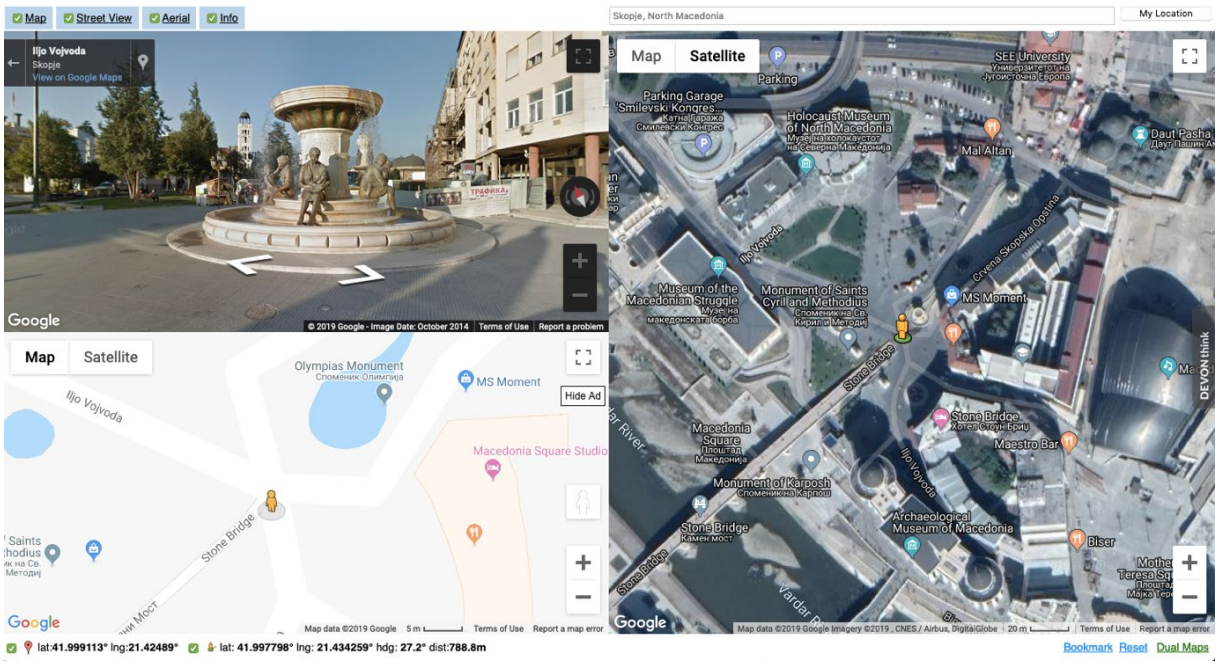
[75] https://www.bing.com/?scope=images

[76] https://www.tineye.com



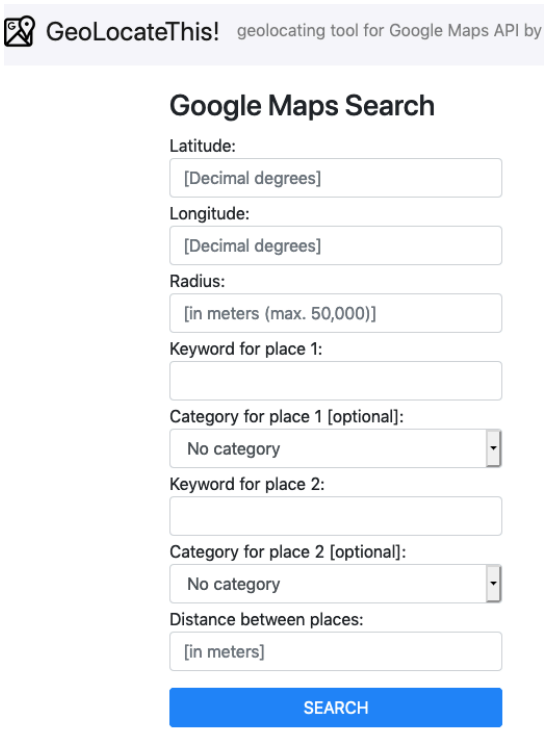
Select the year in the lower left corner after which a slider in the upper left corner opens. There you can slide through the years with the exact date of the imagery noted in the lower middle.

Another tool is DualMaps[77] Based on the Google data, this site provides the map, the aerial photo and street view images in one screen. If you move the perspective in one of the three windows, the other two will adapt accordingly.



[77] http://data.mashedworld.com/dualmaps/map.htm

Another tool that builds on Google maps is GeolocateThis[78] which actually uses the labels on Google maps to find instances where a combination of two places exists within a certain radius of a point.



But do not limit yourself to Google maps, there are dozens of other map and satellite tools like Bing[79], OpenStreetmap[80], ZoomEarth[81] and many, many more.

Geolocation of sites is something that you should do often to obtain a bit of routine in understanding what methods and steps usually get you good results. There is a nice recent blog post on the methods applied and the fact that you sometime have to from assumptions.[82]

If you want to train your abilities, go to Geoguesser[83] on a rainy Sunday afternoon and start practicing. Or you participate in the daily @quiztime geo-location challenges on Twitter.

9. Corporate registries

Because a large part of commercial activities and holding of assets is organised via legal entities, corporate registers – also known as company registers – form a very important source of information for any (online) investigation.

In the basis these registers are organised by (part of) the country where the entities are formally registered. However, there are also (commercial) databases with a wider reach where corporate data from different countries is combined. And there are other types of databases where relevant information of legal entities could be available. We discuss these all three.

Local registers

The primary source for data on legal entities is formed by the local official registers which mostly are maintained by governments. The registry could be maintained by a separate government entity that is tasked with company registrations (e.g. chamber of commerce), it could be a task of the Ministry of Justice or a task of the courts. The registers vary from detailed relational databases to a non-linked collection of government announcements on registrations and changes of legal entities. Data available to the public varies from only the entity name, number and whether it is active, till a complete repository of all filed documents.

[78] http://www.geolocatethis.site/
[79] https://www.bing.com/maps
[80] https://www.openstreetmap.org/#map=18/42.00645/21.42727
[81] https://zoom.earth/
[82] https://nixintel.info/osint/quiztime-9th-may-2019-the-french-prison-system/
[83] https://geoguessr.com/

In some countries there is no (online) digital company register but, in most countries, nowadays the corporate register is accessible via the internet even though content and modes of access vary wildly. Often it takes some time to identify the relevant register, to obtain access and to learn how to navigate it.

Examples of very extensive registers are Companies House[84] in the UK and the Danish register[85] where all relevant documents on the legal entities can be downloaded for free, whereas the register on Bermuda shows only the name and number[86] of legal entities. When searching for a register, note that there are cases where a register only covers a part of the country and there are multiple registers. Examples include Bosnia of course, but also Canada where there are provincial registrars as well as a federal registrar and no centralised index.

An overview of company registers can be found for example on the website of RBA Information Services[87] and on the Investigative Dashboard.[88]

Aggregated (commercial) registers

Various data brokers maintain aggregated databases with regional or worldwide company information. Most often these registers are in the English language, with user friendly interfaces for easy search and have linked data from different jurisdictions together which is indexed, translated and normalised. These databases are of course a secondary source and generally do not provide original filings while also the data may not be completely up-to-date. Examples include the global data providers BvD Orbis[89] and LexisNexis[90] and the regional Russia/CIS database SPARK[91].

These commercial solutions can be rather expensive ranging from several thousand euro for a subscription to a few hundred thousand euro per year.

Free sources

Fortunately, there are also many free alternatives and the most comprehensive free company information source is Open Corporates[92] which provides quite good searching and filtering options and a link to the original source. In most OSINT research that would be the use case for these aggregated data sources: use them to locate the original sources and then find there the most up to date, complete and accurate data on the entity you are researching.

Note that a number of the aggregated database can be access in a ‘pay-per-view’ mode where searching is free, however, viewing the content requires a payment. These are very useful resources to locate where relevant data may be. Examples include:

- Cedar Rose[93] which is very strong in the Middle East and Africa
- ClarifiedBy[94] also strong in the Middle East

[84] https://beta.companieshouse.gov.uk/

[85] https://datacvr.virk.dk

[86] http://companysearch.bz/public_search/

[87] http://www.rba.co.uk/sources/registers.htm

[88] https://investigativedashboard.org/databases/

[89] https://orbis.bvdinfo.com

[90] https://risk.lexisnexis.com/

[91] https://www.spark-interfax.ru/

[92] https://opencorporates.com/

[93] https://www.cedar-rose.com/

- Orbis[95] worldwide
- Info-Clipper[96] worldwide (use more to identify companies, their reports are not the best, check what you get before you buy it)

A most interesting free source is of course the Offshore Leaks website[97] maintained by the International Consortium of Investigative Journalists (ICIJ). Note that journalists can get access to the totality of the leaked files by sending an email to data@icij.org (applications are vetted)

We can recommend always to check multiple sources especially if you are looking for rare data.

Alternative sources

There is a number of alternative sources that you can use to identify legal entities and obtain information. We will discuss a few.

One alternative source could be tax registers. For example, in Europe a VAT number search tool[98] is a great source to quickly search though the available VAT repositories and find company data. Also, the Kazakh government maintains a website[99] which you can use for such searches.

Another source of data could be the Legal Entity Identifier (LEI) database. In 2011 the Financial Stability Board, which was established by the G20 decided to develop the LEI as a global 'company registration number', a 20-character, alpha-numeric code. The Global Legal Entity Identifier Foundation (GLEIF) was established and tasked to support the implementation and use of the LEI. It maintains a database[100] of all entities that applied for such a registration.

One other and often overlooked data source are the credit agencies. Credit agencies collect information about the paying habits of companies and individuals and turn that into credit ratings. In this process they often come across information which is not available in corporate registers. Information is collected from legal proceedings, from companies that outsource or insure their receivables and sometimes from the target companies themselves who provide detailed information in the hope of obtaining a better rating.

A credit agency which has currently the best coverage in the Balkan region is Bisnode[101]and Scoring in Serbia[102]. There is a charge per report however these often contain names and financial information such as bank account numbers which may not be found (anymore) in the official registers.

---

[94] https://www.clarifiedby.com/

[95] https://orbisdirectory.bvdinfo.com

[96] http://www.info-clipper.com/en/

[97] https://offshoreleaks.icij.org/

[98] https://tva-recherche.lu/

[99] http://kgd.gov.kz/en/services/taxpayer_search

[100] https://search.gleif.org/#/search/

[101] https://search.bisnode.ba/

[102] https://www.scoring.rs/

Regional corporate registries

We have collected the links to the regional registries in the Balkans (as of May 2019). Note that these links are subject to contact change so keep them updated!

Country	Corporate register	Remarks
Albania	http://www.qkr.gov.al/search/search-in-trade-register/search-for-subject/	Data only in Albanian
Federation of Bosnia and Herzegovina and Brčko district	https://bizreg.pravosudje.ba/pls/apex/f?p=183:20:3056199352154207::NO::P20_SEKCIJA_TIP,P20_POMOC:PRETRAGA,FALSE	
Republika Srpska	http://bizreg.esrpska.com/Home/PretragaPoslovnogSubjekta	
Croatia	https://sudreg.pravosudje.hr/registar/f?p=150:1	
Greece	https://www.businessregistry.gr/publicity/index	Only in Greek
Kosovo	https://arbk.rks-gov.net/page.aspx?id=3,1	
North Macedonia	https://www.crm.com.mk/DS/default.aspx?MainId=12	Poor search algorithm, very limited data. Use of BIFIDEX (Western Balkans) recommended
Montenegro	http://www.pretraga.crps.me:8083/	
BIFIDEX	https://www.bifidex.com/sr-latn/home	Regional platform established by Serbia and North Macedonia, i.e. the national Business registrars. It is directly connected to the national databases. Other countries from the region are expected to join
Serbia	http://pretraga2.apr.gov.rs/unifiedentitysearch	Only available in Serbian Cyrillic. Very comprehensive information and many documents available.
Slovenia	https://www.ajpes.si/prs/Default.asp?language=english	Mandatory registration to browse data, lots of data available

10. Meta data research

Website metadata

Every website, even the most basic one, has at least two components, a hosting provider, i.e. disk space on some server connected to the internet and a domain name, so that it can be

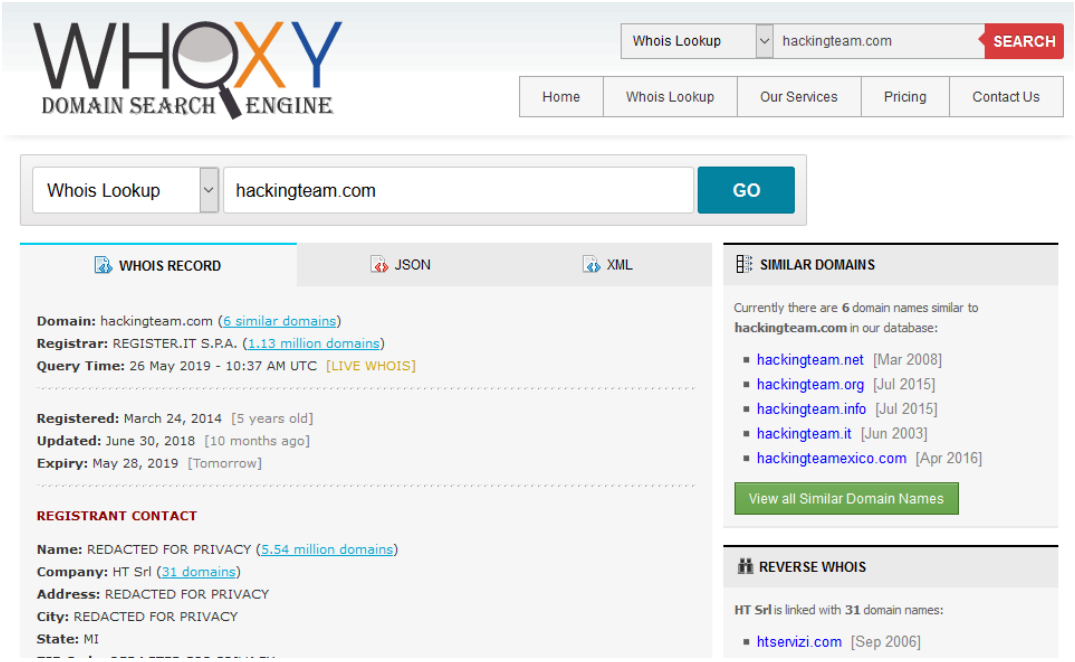
accessed from the internet[103]. Both these components require the use of intermediaries between the website owner and the internet. These intermediaries are the Internet service or Hosting provider and the Domain name registrar.

These intermediaries store some data on the website owner for different purposes, including billing, but also technical data. Even though through different changes in legislation like the GDPR and services like Domain [Whois] Privacy the data on website owners on the internet is limited, some useful information can be found and used alongside other information, or help establish stronger connections.

Domain data

The most common way to search for domain data is through *whois* searches. There are a few services online that can do the job quite well, such are who.is and whoxy.com as used in the example below.

As an example, we will look up a company that sells offensive cyber tools to governments, Hacking Team from Italy. Their domain name is hackingteam.com.



A simple search gives us very little information on the domain, since the company uses whois privacy. First of all, we can see that the domain is registered in Italy (REGISTER.IT S.P.A.). Additionally, there are six other domains that are similar to this one, which might be an indication that they have been registered by the same entity, since it is a common practice to register both .com and .net domains for example.

Another important item is the date the domain was registered, even though that points to the registration with the current registrar, the domain name might have previously be registered with other registrars and then transferred to the current one.

[103] A domain name is not actually a prerequisite for the existence of a website, since it can be accessed through the IP address of the hosting server. However, it is extremely common that a website is linked to a particular domain name.

A crucial piece of information can be found in the company field, *HT Srl (31 domains),* means the same company has registered 31 domains in total, and those can point to further clues in the investigation. A reverse whois on this entity shows the following information:

HT SRL

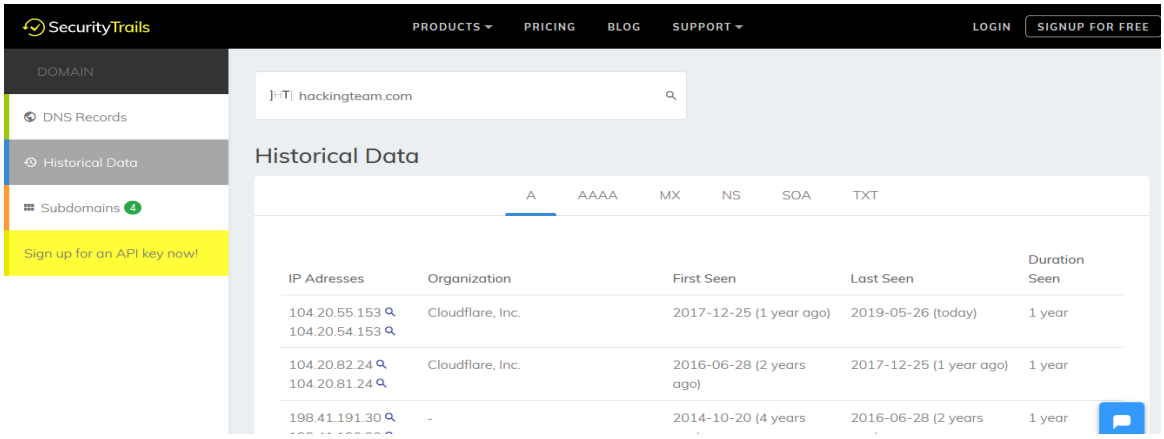
Reverse Whois » COMPANY [*HT SRL*] { **31 domain names** }

NUM	DOMAIN NAME	REGISTRAR	CREATED	UPDATED	EXPIRY
1	hackingteam.com	Register.it SPA	24 Mar 2014	30 Jun 2018	28 May 2019
2	rodrigoroyg.com	FastDomain Inc.	18 Nov 2013	18 Nov 2018	18 Nov 2019
3	madrematerіapadreprogettofigliopensiero.com	Register.it SPA	14 Apr 2018	14 Apr 2018	14 Apr 2019
4	artkitchenartbathroomartliving.com	Register.it SPA	14 Apr 2018	14 Apr 2018	14 Apr 2019
5	minimalismoitaliano.com	Register.it SPA	5 Feb 2018	5 Feb 2018	5 Feb 2019
6	minimalismomaterico.com	Register.it SPA	5 Feb 2018	5 Feb 2018	5 Feb 2019

The most important feature of this service compared to others is that it features the connections between different domains in a very simple manner, through links, which give the possibility to dig around quite a bit.

Historical and additional data

Since most services for whois privacy are paid, there is a chance that at some point the website didn't use this service and some of the details might still be available in historical domain name records. There are a few available services that allow you to look for *historic domain records (dns) data*. The one used in this example is SecurityTrails.[104]

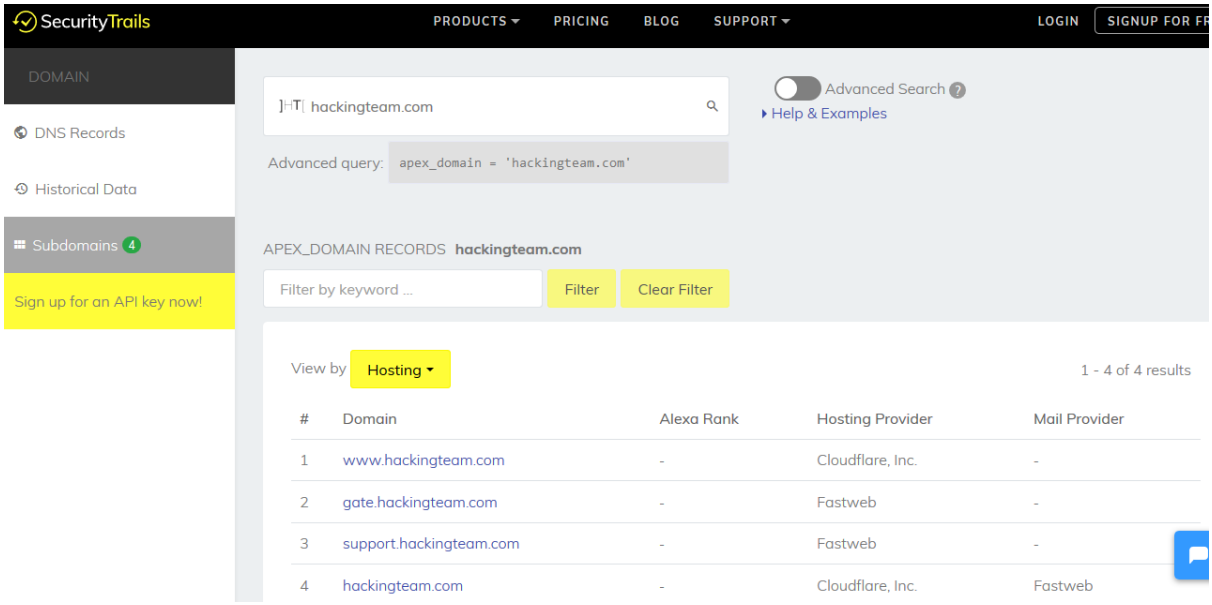


In the *Historical Data* segment there are historical data for different common DNS record types.[105] The records in this case go back some 10 years and shows how the hosting (A record) and email (MX records) have changed, meaning the domain has changed service providers.

[104] https://securitytrails.com/dns-trails

[105] https://simpledns.com/help/dns-record-types

Other important information regarding the infrastructure of the company from our example can be found in the *Subdomain* section.

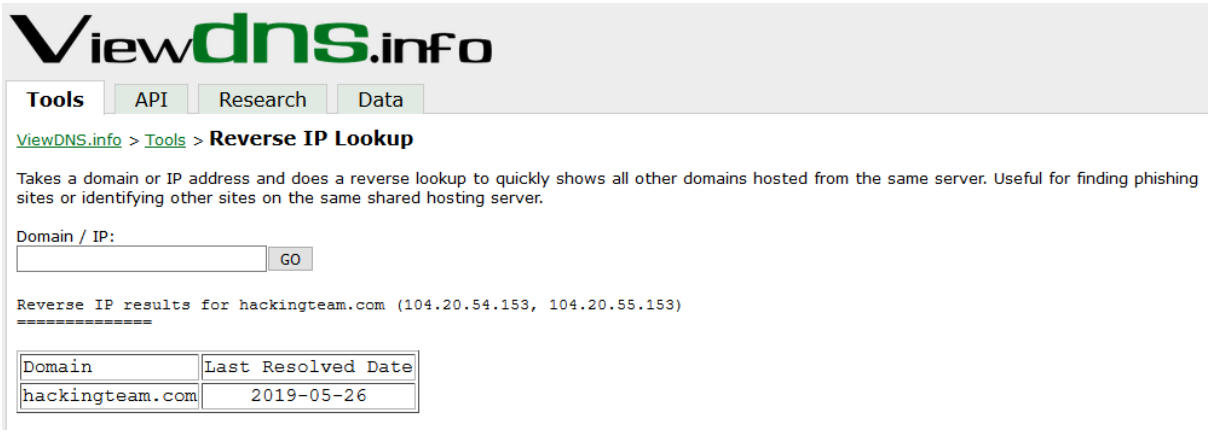


The records here show data about the hosting and email provider for the current domain.

Related hosted websites

In some cases, it might be necessary to check whether other websites are hosted on the same server as the primary website. This doesn't work well if a website uses *shared hosting,* as the other websites hosted on the same server might and are probably not related to it. A very simple service for *Reverse IP Lookup* is ViewDNS.[106]

In this case the search shows that there are no other websites hosted on the same server as hackingteam.com.



[106] https://viewdns.info/reverseip/

Google Analytics data

When trying to establish a connection between different websites and thus link a website to a particular organisation or a person, a useful piece of data might be a *Google Analytics ID* used on a particular website. While it is not that easy to map it directly to an exact company, this ID can show connections between various websites.

Google Analytics is a service offered by Google. It allows website owners to have a quantitative and qualitative overview of their website audiences, on the other hand it allows google to create precise profiles of people, tracking them across websites that implement this technology.

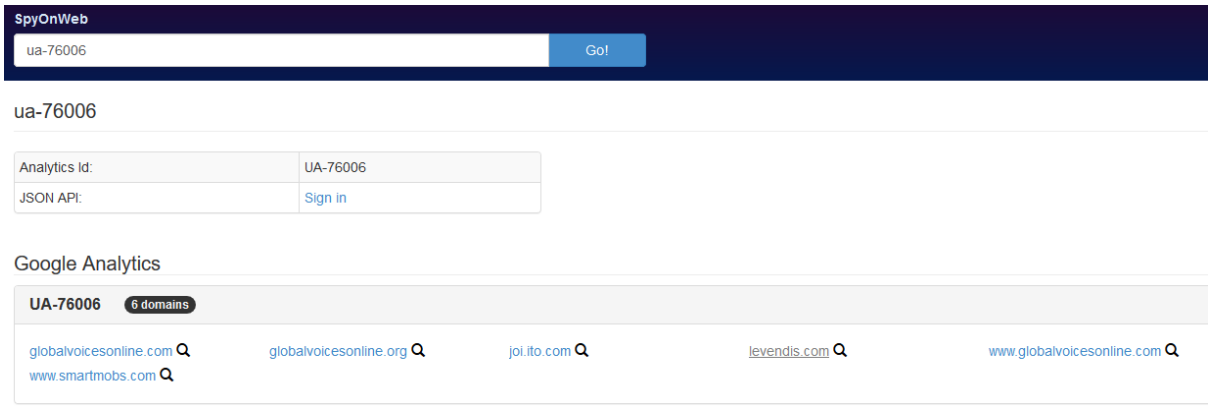
Every website owner who wants to embed Google Analytics in their websites has a specific GA ID number; the number looks like this: *UA-12345678*. There are a few ways to obtain this number.

In your browser:

1. Open a website
2. Right-click anywhere (On *Safari*, go to the *Page* menu on the top right)
3. Click *View page source*
4. Search (Ctr/Command + f)
5. Type 'UA-'
6. Find the part of the website code that looks something like this:

```
<script type="text/javascript">
    (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
    (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
    m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
    })(window,document,'script','//www.google-analytics.com/analytics.js','ga');
    ga('create', 'UA-76006-4', 'auto');
```

Once you find the ID number you can use tools like SpyOnWeb[107] to look for connected websites.

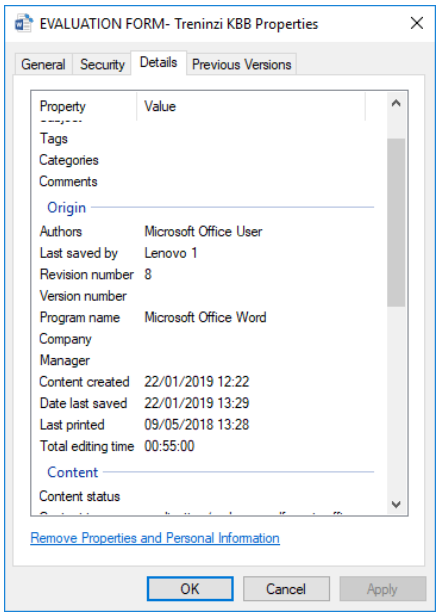


File metadata

[107] http://www.spyonweb.com/

Documents

Every document created on any device carries some amount of metadata with it. The amount may vary, but this information might be very useful when trying to understand the context and circumstances in which a document was produced.



The simplest way to look at metadata of practically any document is to right-click any document, then go to *Properties* and then to the *Details* tab.

Images

With images, the metadata can be used to verify the authenticity of the image. Whether tools like *Adobe Photoshop* were used to change something on the image or not, to find the location where a photo was taken or to see the make and model of the camera.

Metadata search is not useful for images that have been posted on *Social Media*, since platforms like *Facebook* and *Twitter* strip the images of their metadata.

There are various tools that can be used for looking at image file metadata, one of them being Jeffrey's Image Metadata Viewer.[108] Besides images, it can also extract metadata from *.pdf* files as well.

XMP

XMP Toolkit	Adobe XMP Core 5.3-c011 66.145661, 2012/02/06-14:56:27
Document ID	xmp.did:55F93EE8586211E991C799A990CFA35B
Instance ID	xmp.iid:55F93EE7586211E991C799A990CFA35B
Creator Tool	ILCE-6300 v2.01
Derived From Instance ID	30ED59743DFA1D9CBADF79B5547ADAAA
Derived From Document ID	30ED59743DFA1D9CBADF79B5547ADAAA

Photoshop

IPTC Digest	fce11f89c8b7c9782f346234075877eb

APP14

DCT Encode Version	100
APP14 Flags 0	[14], Encoded with Blend=1 downsampling
APP14 Flags 1	(none)
Color Transform	YCbCr

IPTC

Coded Character Set	UTF8
Application Record Version	2

[108] http://exif.regex.info/exif.cgi

11. Dark web

Deep, Dark, what is the difference?

As may have become clear throughout this guide, much of the internet's resources are outside of the view of search engines where the content and social media can be found that most people use on a daily basis. The analogy of an iceberg floating in the ocean is often used to illustrate this.[109] Only an estimated 10% of the information available via the internet is floating above the surface and indexed by search engines.

Most of the information can be found at deeper levels once you know where to search. Much of this information probably cannot be crawled and indexed by search engines because access requires some form of authentication. This is where the different databases we discussed reside.

Then, within that deep web, a space exists that is usually called the Dark Web. The resources in this part of the internet require additional knowledge and software to access. The Dark Web is a collection of thousands of resources that use anonymity tools like Tor and I2P to hide their IP address. While it's most famously been used for all kinds of illegal trade, the Dark Web also enables anonymous whistleblowing and protects users from surveillance and censorship.[110]

There are multiple services and nets that are considered to be part of the Dark Web. These include more known networks like Tor and Freenet operated by public organizations and individuals but also small, friend-to-friend peer-to-peer networks.

Overall, the Dark Web is a place for people to (more) anonymously communicate with others which may be important if you live in a country where freedom of speech is not allowed. For example, Tor allow users to move from the internet into the Dark Web, bounce from system to system, and then come back to the internet to access or provide content more-anonymously. That anonymity also has attracted criminal and there are marketplaces in the Dark Web where you can buy and sell all types of illegal items from guns and drugs to stolen information.

Finding relevant data in some of the Dark Web networks is usually challenging because these services are private or hidden. We will discuss here Freenet very briefly and Tor, which may be the most useful for OSINT, more extensively.

Why would you as a journalist want to visit the Dark Web? There could me multiple reasons:

- To hide your identity when accessing certain sources;
- To communicate (more) anonymously with your human sources;
- To access content not available on the Clearnet
- Curiosity, knowing what is out there.

[109] https://www.quora.com/What-is-deep-web-and-tor-How-to-browse-the-deep-web

[110] https://www.wired.com/2014/11/hacker-lexicon-whats-dark-web/

Please note that in the Dark Web, it is easy to find illegal, unethical, and disturbing content. If your work of your curiosity takes you into the Dark Web, you need to ensure that you are protecting yourself, your organization, and your computer systems against threats.

Are you prepared to see images that cannot be 'unseen' and can you deal with that? Are you considerate towards the feelings and beliefs of the colleagues working around you and that may involuntarily see what you have on your screen as well? Does your organisation have any rules in place on accessing certain types of information? You may want to clear this with your editor-in-chief before you start accessing the Dark Web. And finally, sites may contain harmful content for your system, using a VM is highly recommended when you are (regularly) visiting the Dark Web.

Freenet[111]

The Freenet is in essence a distributed file-sharing network that breaks up files into pieces and scatters those pieces in retrievable sections across its network. This makes it impossible to remove content that is uploaded to the Freenet.

Think of the Freenet project as a large storage device where content is spread across all the computers of a network, which are called 'nodes' in the Freenet project. Each connected node (computer) has a local data-store that contains encrypted fragments of popular files that it shares with the other nodes on the network. When you upload a file to Freenet, it is divided into pieces, encrypted, and then distributed to other nodes. You received a key to retrieve that file and you can provide that key to anyone else on Freenet and that person can also retrieve that file.

All nodes of the Freenet share pieces of the data uploaded to Freenet and because each file is divided into pieces and distributed to other nodes, each node could have content from any file. Any user can increase or decrease the encrypted storage space on their hard drive that they are willing to donate to the Freenet within the Freenet settings.

Be aware that, if someone uploads stolen or otherwise sensitive data to the Freenet, it will be encrypted and distributed to peer nodes. Your computer may hold that content, but you may never know it. That is the power of the Freenet. No user knows what data their system contains therefore they cannot delete content from the Freenet and authorities cannot shut down the Freenet by removing some of the nodes.

In addition of anonymously file sharing, Freenet offers several other mechanisms to exchange information such as email, forums, and you can event create your own web site. Visit https://www.deepdotweb.com/2016/10/22/introduction-freenet-censorship-resistant-network/ for further tutorials and explanations.

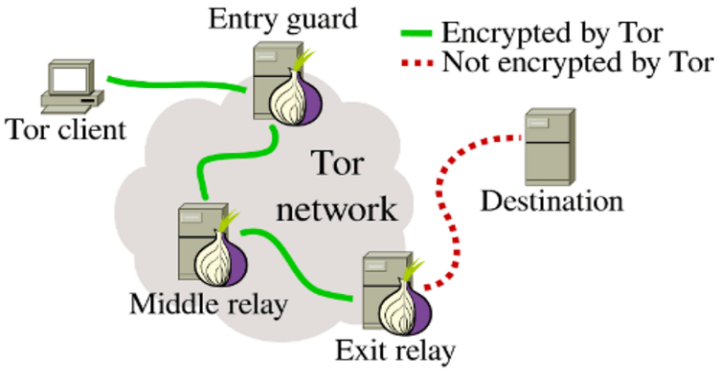
Tor

The Tor dark web may be referred to as 'onionland' a reference to the network's top-level domain suffix .onion and the traffic anonymization technique of onion routing. It is based on the use of TOR which is the abbreviation of ***The Onion Router***

This is a decentralised, volunteer-based overlay network. The network implements encryption of data embedded in various layers, much like an actual onion.

[111] https://freenetproject.org/

Through a client, i.e. the Tor browser [112], the user connects to the Tor network through an ***Entry Node,*** afterwards the traffic is routed through a few other points, or ***relays*** in the network in which only the previous node is visible. This makes it hard to trace a particular piece of data to its source while it's been transferred through the network. Finally, through an ***Exit Node***, the data is sent out of the Tor network to the destination host.



NB: The connection between the Exit Node and the final destination you are visiting, i.e. the server does NOT go through an encrypted channel. Thus, Tor is an anonymity, and not a security tool.

How to configure TOR to exit in a specific country?

Sometimes while researching you might need to configure TOR to exit in a specific country. This means that the website you will be visiting on the public internet will see you as a visitor from the country you have configured. The potential issue with this technique is that exit nodes are not present in every country in the world, and if you go through a country that has few exit nodes, the speed might be bad.

Anyway, the configuration is very simple, but in order for this to work, you need to run Tor browser at least once after you install it:

1. Navigate to the folder in which you installed Tor then go to: ***Browser/TorBrowser/Data/Tor***.
2. Find a file named ***torrc***
3. Copy it and name the copied file ***torrc_backup (we are going to use this in case something goes wrong)***
4. Open the ***torrc*** file using ***Notepad*** or ***Notepad++***
5. Add this line in the end of the file ***ExitNodes {xx} StrictNodes 1***
6. Replace ***xx*** with the appropriate country code from the list.[113]
7. You can add multiple countries
8. ***ExitNodes {xx}, {yy}, {zz} StrictNodes 1***

[112] https://www.torproject.org/

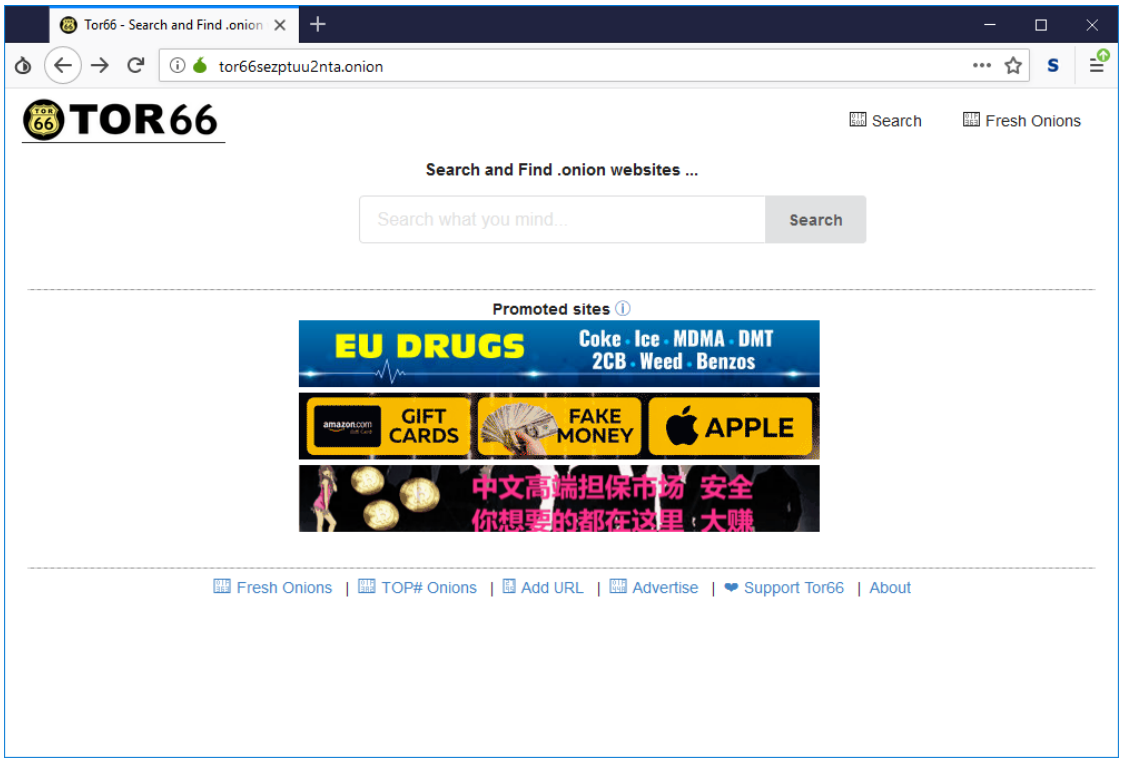
[113] http://archive.is/jiYA9

9. Change the ***StrictNodes*** attribute from 1 to 0 if you want to allow Tor to go through other countries if it is not possible to establish a connection through the countries you listed.
10. Save the file, restart Tor
11. Now your exit node is in the country you specified

Browsing the Tor network

The Tor network has its own ***TLD*** - ***Top Level Domain, .onion.*** Even though the Tor network is built over the internet, it is not possible to browse websites with an .onion TLD using conventional ***search engines*** like ***Google*** or ***DuckDuckGo.***[114] although there are services like Tor2Web[115] that provide a bridge between the .onion world and the surface web.

Special search engines like ***tor66sezptuu2nta.onion*** are used to browse websites in the Tor network. The way these search engines work is quite similar to their regular counterparts. The difference is that using the .onion websites you can browse for things that are illegal or unavailable on the regular internet.



Note that recent research has shown that the Dark web contains only very limited information[116] and searching onion site other than for very specific research will generally not give you much relevant results.

[114] DuckDuckGo has its .onion version, https://3g2upl4pq6kufc4m.onion/. However, it still browses through indexed websites on the public internet.
[115] https://www.tor2web.org/
[116] https://www.recordedfuture.com/dark-web-reality/

12. Data handling

Formats

Data handling is a set of processes and techniques that are used in large datasets, i.e. to improve the quality of large datasets and thus have more precise analysis of the data.

Human readable data are not always understandable for machines, i.e. software or an algorithm for data processing and analysis. Patterns and anomalies in data that are not visible for humans, or can easily be disregarded, strongly influence the way an algorithm would understand said dataset.

While humans need a lot of context and narrative data, machines prefer dry and clean data. A scanned 17 century document can show a lot to a human being, but for a machine it might be a little hard to understand.

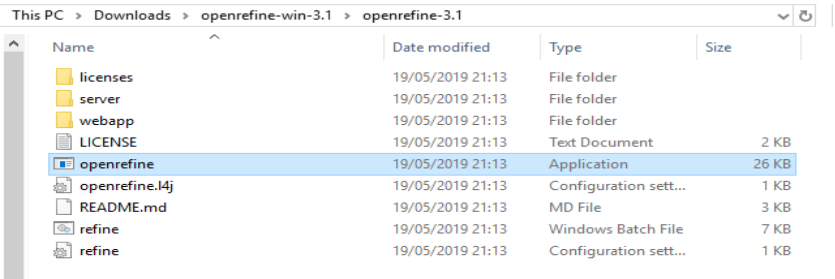
There are a few machine-readable formats, like *xml, json* and *csv.* All these formats are text-based formats. The simplest one of these is *csv,* i.e. Comma Separated Values. This is a format of a text file in which data is delimited by the use of characters like comma "," or semicolon ";".

Data cleaning

The data featured below are delimited by semicolon, but at the same time, another separator "-" is used. This will make it hard for programs like Excel to deal with this file unless it's properly cleaned first.

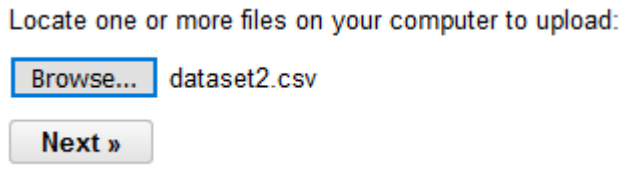
```
1  Country;;2012-2013-2014
2  Република Србија;CRIMINAL OFFENCES;123-25-25
3  Република Србија;;
4  Република Србија;;
5  Република Србија;CRIMINAL CHARGES;65-8-10
6  Република Србија;PERSONS;143-41-38
7  Република Србија;CRIMINAL CHARGES;65-8-10
8  Република Србија;PERSONS;143-41-38
9  Република Србија;CRIMINAL CHARGES;65-8-10
10 Република Србија;PERSONS;143-41-38
```

For that purpose, we will use OpenRefine[117], a common tool for cleaning data. After a successful download and unzipping at the desired location we open the application *Openrefine*.



[117] http://openrefine.org/

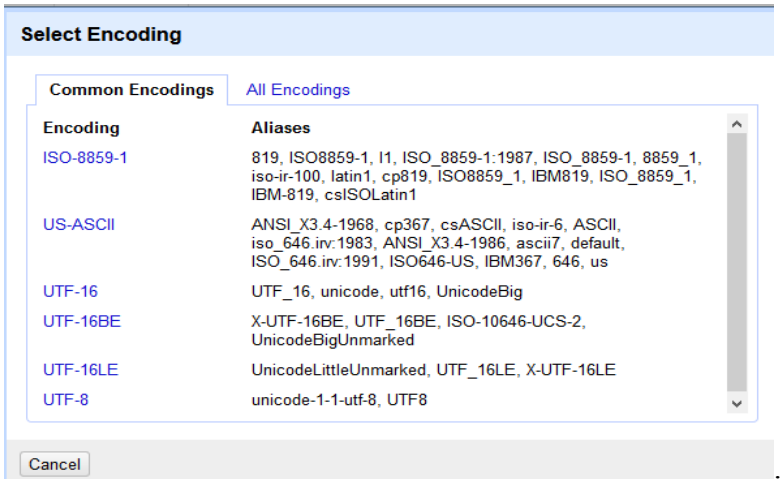
Once the application launches, the OpenRefine tab should appear in your default browser. Here, browse the input file and click next.



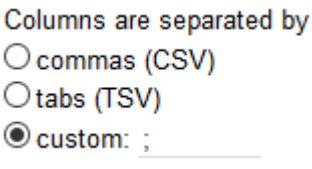
Since the first column of the dataset we use in this example is in Serbian Cyrillic, the application may have a problem reading it and might show something like this:

	Country	Column	2012-2013-2014
1.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°	CRIMINAL OFFENCES	123-25-25
2.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°		
3.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°		
4.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°	CRIMINAL CHARGES	65-8-10
5.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°	PERSONS	143-41-38
6.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°	CRIMINAL CHARGES	65-8-10
7.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°	PERSONS	143-41-38
8.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°	CRIMINAL CHARGES	65-8-10
9.	Đ ĐµĐ¿ÑƒĐ±Đ¸Đ°Đ° Đ¡Ñ€Đ±Đ¸Ñ˜Đ°	PERSONS	143-41-38

In order for OpenRefine to show the Cyrillic characters in a proper character encoding.[118] In this case we need to choose *UTF-8* in the encoding section

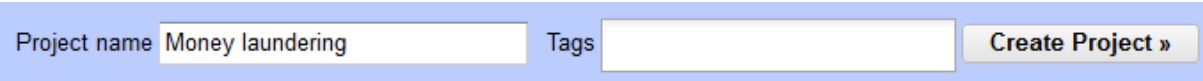


In some cases, OpenRefine captures the column separator automatically, but in case it doesn't, it can be chosen in the separator section. In this case custom ";".

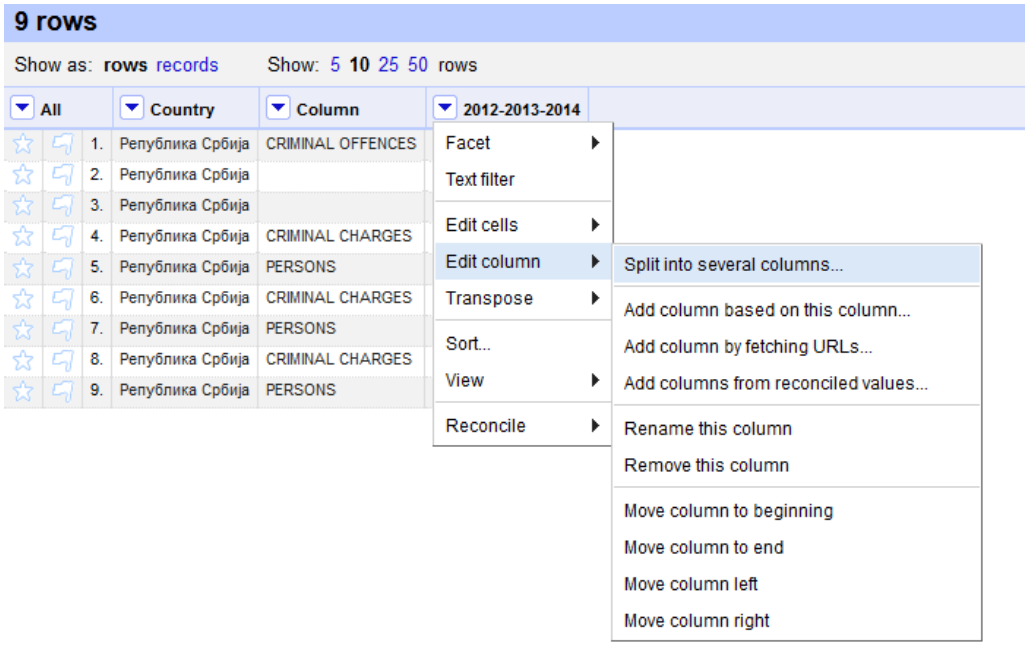


[118] https://www.w3.org/International/questions/qa-what-is-encoding

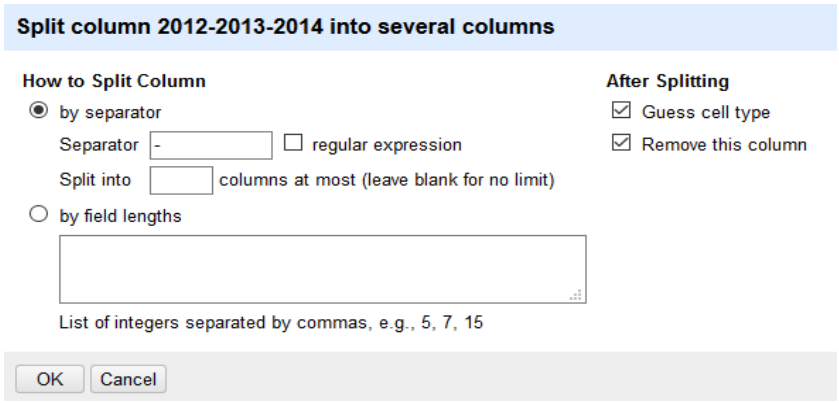
After these basic settings are configured, we can *name* the project and *create* it.



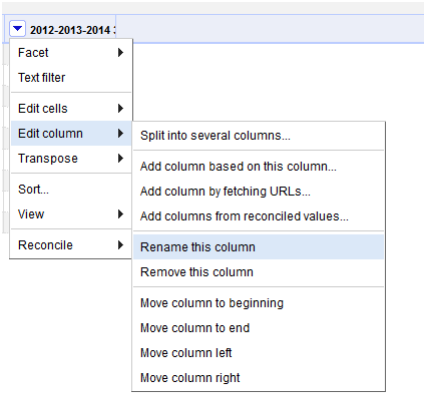
Once the project opens, we need to deal with the last column, which obviously doesn't show the data in a proper manner. In order to do that we click on the triangle in that column and under *Edit column* chose *Split into several columns...*



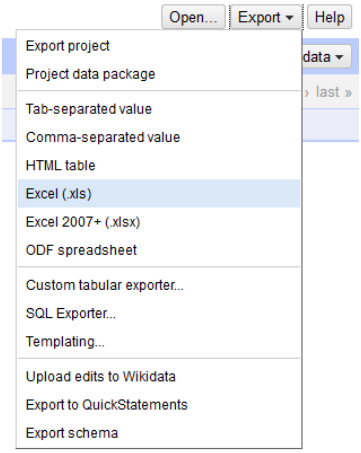
In the *Split column* window, we specify "-" as a separator and click OK.



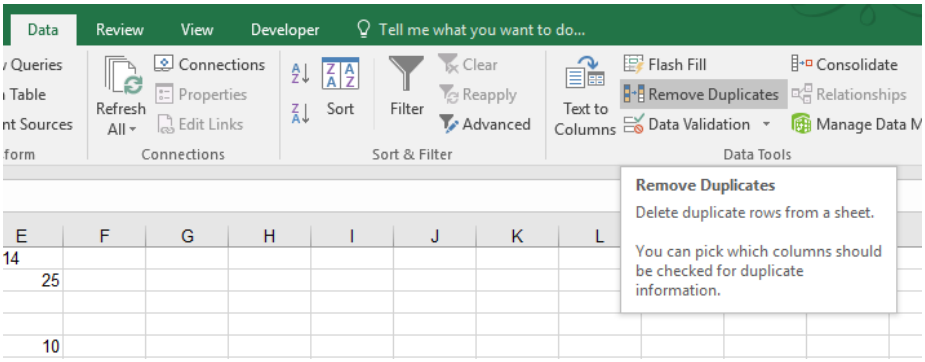
By clicking the triangle next to each of the three newly created column, then *Edit column,* then *Rename this column,* we change the names of the columns to 2012, 2013 and 2014 respectively.



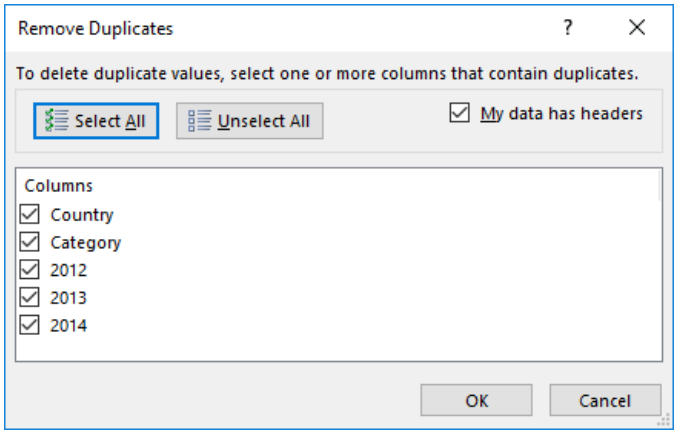
We notice that there are a few duplicate rows in the file. Since this can lead to false conclusions, we need to remove the duplicate rows. This is easier to do in Excel, so we export our data in Excel format.



Once the file is open in Excel, navigate to the *Data* tab, to the *Data tools* section, and chose *Remove duplicates*.



In the remove duplicates we choose which columns should be checked for similar values. Since we only want to remove rows that are identical to other rows, we check all the columns.



This operation will remove all the duplicates, but we are still left with one empty row (containing data only in the first column) which we don't need, so we delete it the conventional way, by *right-click* and *Delete.*

Now we have a more or less clean data.

	A	B	C	D	E
1	Country	Category	2012	2013	2014
2	Република Србија	CRIMINAL	123	25	25
3	Република Србија	CRIMINAL	65	8	10
4	Република Србија	PERSONS	143	41	38

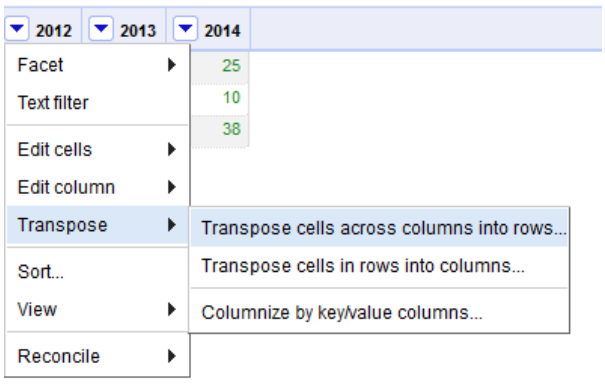
Unpivoting

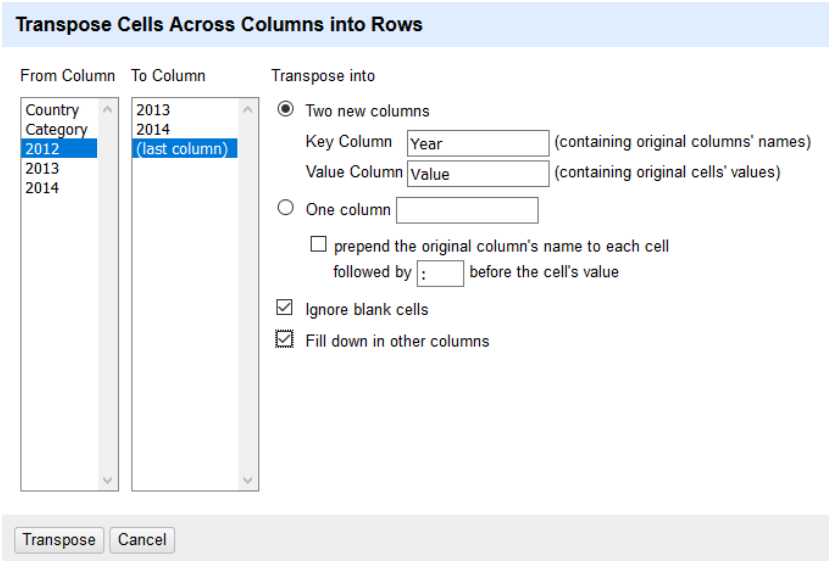
Even though this data is perfectly clean for human use, it still needs one more step to be fully machine readable. That step is *unpivoting* and for that we need to go back to OpenRefine.

Unpivoting in this example will be creating a new column *Year* which will take the three last columns as values and transpose their primary values into rows.

Once the new table is imported we click on the triangle of the first column that we want to unpivot, in this case *2012*, then *Transpose* and then *Transpose cells across columns into rows...*

In the *Transpose* window, we chose to *Transpose into Two new columns,* for which we chose the names *Year* for the column that will contain the original columns' names and *Values* for the column that will contain the original values. We also check the *Fill down in other columns* to have the values in the other columns filled in the appropriate manner. Then we click *Transpose.*





Thus, we get an unpivoted, machine readable dataset, which can be used with different data processing and visualisation tools.

9 rows

Show as: **rows** records Show: 5 **10** 25 50 rows

All		Country	Category	Year	Value
1.		Република Србија	CRIMINAL OFFENCES	2012	123
2.		Република Србија	CRIMINAL OFFENCES	2013	25
3.		Република Србија	CRIMINAL OFFENCES	2014	25
4.		Република Србија	CRIMINAL CHARGES	2012	65
5.		Република Србија	CRIMINAL CHARGES	2013	8
6.		Република Србија	CRIMINAL CHARGES	2014	10
7.		Република Србија	PERSONS	2012	143
8.		Република Србија	PERSONS	2013	41
9.		Република Србија	PERSONS	2014	38