Ludo Block

Andrej Petrovski

OPEN SOURCE INTELLIGENCE NAVIGATOR FOR INVESTIGATIVE JOURNALISTS

Contents

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 reporters' network (BIRN) we have addressed the open-source data gathering 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 outcome of previously conducted training with the BIRN journalists *"Open Source Intelligence for Journalists"* that was held on 21-22 May 2019 in Skopje, North Macedonia.

This guide was updated in October 2021 to reflect the many changes in tools and sources available since 2019.

Developers of the navigator: Ludo Block and Andrej Petrov

1 | OSINT basics

Introduction

Over the past two decades internet has become a crucial resource, not only for communication but also for the access to unimaginable amounts of data. Each year, the amount of data accessible via the internet grows and in particular the amount of video shared every minute is staggering. That is not only due to the fact that an increasing number of databases are accessible via the internet, but also the amount of data shared on social media.

If we look at the What Happens Online in 60 Seconds1 which is published every year, we can get an impression of the high number of tweets, pictures posted to Instagram and so on, most of which eventually may become publicly available.

---

1   https://www.bondhighplus.com/2021/04/14/what-happen-in-an-internet-minute/

This guide is designed to help you navigate the vast amount of publicly available data and understand how to collect and analyse it to support your investigative journalistic work.

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', 'tenacity' 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 other regions like Asia, Afrika and the Americas. Also, we pay less attention to the analysis part which is important in OSINT. On the other hand, where relevant, we have provided more coverage of Balkan related data sources.

This guide has not been written without use of many sources. We would like to especially point out the Bellingcat guides and tutorials[2], the OSINT book by Michael Bazzell[3], the UNESCO Handbook for Journalism Education and Training[4] and many other sources available.[5] Where needed we will reference the sources used and attribute those behind it, usually by adding their Twitter handle which will allow you to quickly find them.

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 context.

---

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

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

4   https://en.unesco.org/fightfakenews

5   See https://www.blockint.nl/the-osint-library/ for a full library on OSINT publications.

Of course, in the context of a story which you're writing that data could be meaningful and could be the piece of an interesting puzzle. 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 a publication of the US Army:

(US Army Joint Publication 2-0, Joint Intelligence)
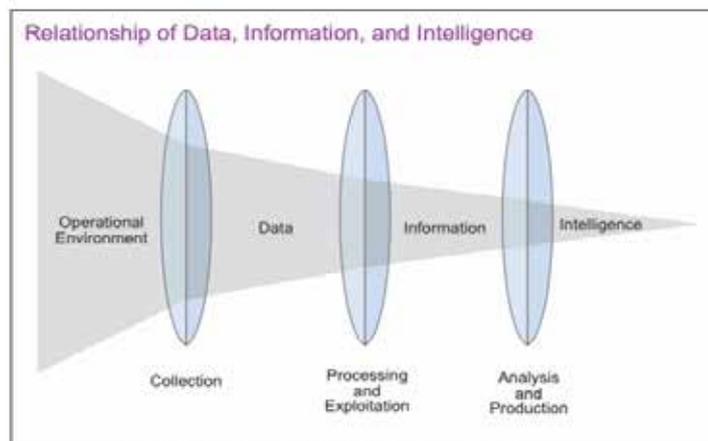
Figure I-1. Relationship of Data, Information, and Intelligence

To come from the real world to the final product, multiple steps are needed. 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 enormous proportions, also, and probably more importantly, the nature of the data and available types of data have changed. Five key changes are relevant:[6]

6   https://www.leidensecurityandglobalaffairs.nl/articles/solving-the-mh17-and-the-skripal-case-how-bellingcat-demonstrates-the-power

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 library 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 hardly any reason to undertake painstaking research in damp archives anymore to obtain data, other than it is an interesting academic 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]

---

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

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

Investigation and validation

Research in open sources for journalistic purposes generally has two aims. First is of course to investigate a subject of interest to see if there may be data available somewhere that could shed a light on the matter. Collection of previously unknown information and data on a certain subject is the primary aim in this case. Questions that come up in such investigations include:

▶ What happened and where?

▶ What more 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 of the investigation can be verification of know data, or in terms of journalists 'fact checking', which may be even more important in our present time. Questions that come up in verification investigations include:

▶ Did that really happen and in that location at that time?

▶ Has this person really 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 on the topic?

▶ 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. And there are certainly ethical considerations connected to collecting and using data from open sources.

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. While 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, 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 purposes' 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 in Romania and Hungary, where authorities have (ab)used the GDPR to harass journalists.

In additional to potential legal considerations, 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 that 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 social media accounts with another (non-existing) identity ('a research account') to view profiles / connect to profiles of those you are researching. This definitely violates the terms of the social media platform, however using a research account may be for your own protection. Is that an accepted practice?

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

▶ Using 'leaked' data (Wikileaks, Offshore Leaks, copied official databases, 'data breach' data). Technically you did not steal that data, but it was stolen somewhere by someone. Does your organisation have a policy on using this data?

▶ Google 'dorking' and using for example Shodan to access unsecured devices or devices with default passwords. Technically you are not breaking in' into a system, however at the same time you, 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 updating this guide in October 2021, however may have changed or disappeared since then.

---

10 Note that in the US under the Computer Fraud & Abuse Act (CFAA) these actions may even 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. While we acknowledge that there are various specialised tools available for the collection and analysis of data from open sources, there are two reasons 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. Additionally, 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 – like in the training – we will make use of mostly free basic tools although we at some moments point towards paid tools to show their existence. We will discuss setting up your system, browser, 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 links which you do not know (yet). Accidentally, or even on purpose such as in the NSO case[12], you may run into malware that will infect your laptop with potentially disastrous consequences depending on how much data is on it.

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 only and only export the relevant findings to your personal system or the organisation system/repository. If the laptop the becomes infected with malware for example, you wipe it and reinstall it. This solution takes additional funding for a separate laptop;

---

12 See 'Private Israeli spyware used to hack cellphones of journalists, activists worldwide', The Washington Post, 18 July 2021: https://www.washingtonpost.com/investigations/interactive/2021/nso-spyware-pegasus-cellphones/

▶ 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. Michael Bazzell explains in his Open Source Intelligence Techniques book how to set up a custom Linux VM. 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 however 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, which are about 150 USD / year. Also, the Silo browser is not as customizable as other browsers.

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.

---

13 https://www.authentic8.com/

14 https://www.mozilla.org

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

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

**Extensions
or add-ons**
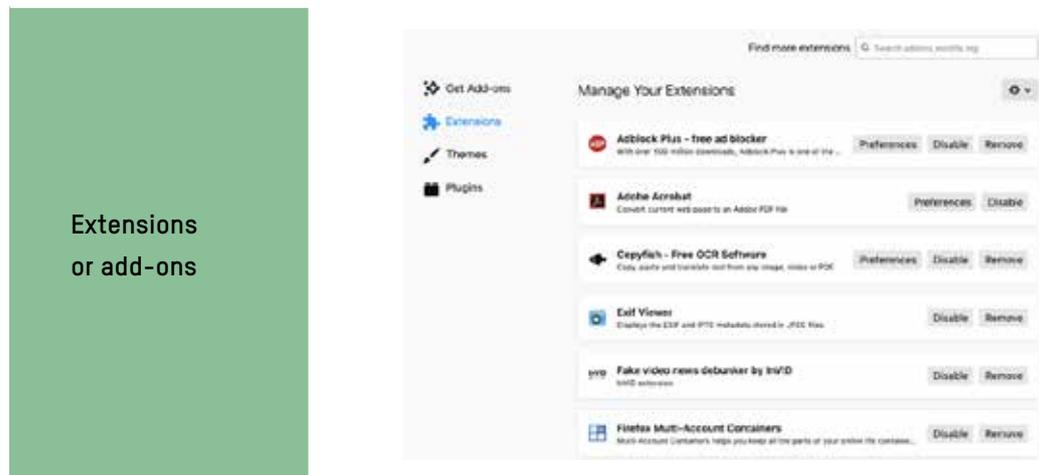
Much used extensions that are useful for OSINT work include:

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;
Instant Data Scraper	scrapes data from websites and experts the data in csv;
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
Privacy badger	blocks ads/trackers
uBlock origin	blocks scripts (this may 'break' webpages so not always good for OSINT)

IP & DNS info	show IP and DNS info on a domain with one right-click;
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 to use:

|	
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.
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. Does need some investment of your time to make it work;

Notepad

for many this smallest 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 form 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 investigators who have to show their whole chain of evidence, but still very useful for journalists as well. 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 Notepad, then copy it from there to the search box.

Notepad
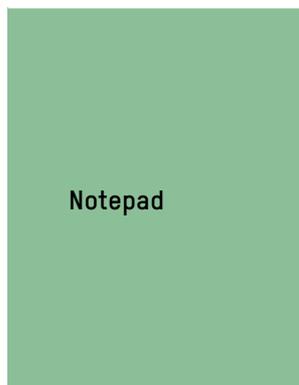
Mind map

Another free (or inexpensive) tool that can help driving your investigations is a Mind map. Mind mapping is a creative and logical way of note-taking and that literally "maps out" your ideas.[16] Originally used for creative processes ('brain-storming') 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. There are various free and inexpensive mind map software available. You can create a mind map on the go, just using it as a kind of note book.

16 https://www.mindmapping.com/

Building your link repository

As you go forward in online research you will probably compile your own link repository. This repository will probably 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 updating in October 2021. However, probably in November 2021 the first links may already have been outdated. The source 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. We will discuss four

**Link
repositories**

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)

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

<table>
<tr><td>

http://
technisette.com

</td><td>

Welcome to Technisette.com!

Here you'll find my collected tutorials, tools, databases, addons, search engines and more to help you with your Open Source Intelligence (OSINT) research.
Select a category at top to browse or use the search bar to find what might need.
These resources were previously stored at https://start.me/p/m6XQ08/osint

Can't find what you're looking for? Or do you have something you'd like to add? Contact me via the 'Contact' page.

Latest resource added on July 19th 2021

P.S. Interested in a 4-hour workshop on how to search Facebook? Sign up for my online webinar on May 21th or October 18th here!
Wanting to learn more OSINT? Sign up for the live online Module 1 here!

</td></tr>
</table>

She 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.

<table>
<tr><td>

start.me page

</td><td>

DIRECTORY | QUICK ACCESS TO OTHER OSINT FRAMEWORK PAGES

Start.me OSINT DIRECTORY, use this to jump off to other pages
there are two tabs, click on the Tab called DIRECTORY

DIRECTORY   TRAINING PARTICIPANTS

- INVESTIGATIVE DUE DILIGENCE | AML | CRYPTOCURRENCY
- SEARCH ENGINES
- SOCMINT SOCIAL MEDIA INTELLIGENCE
- GEOINT | ASSET TRACKING | TRAVELRISK
- IMINT | VIINT IMAGE INTELLIGENCE
- INTERNET INFRASTRUCTURE
- KEYWORD DISCOVERY
- MONITORING ALERTS | RSS
- OSINT TOOLKIT
- PERSON RELATED DATA | EMAIL NAME PHONE LOCATION
- SECURITY BASICS | VPN VM OPSEC
- SOURCES | NEWS | TOR
- COUNTRY WORLDWIDE (Except MENAT, China, Russia and USA)
- COUNTRY USA
- INVESTIGATIVE IDENTITIES | INVESTIGATIVE ACCOUNTS | OPSEC

</td></tr>
</table>

Technisette is also one of the members of the OSINT Curious project[17] which is an interesting blog and vlog community to follow if you want to stay up-to-date with the latest developments in the OSINT community.

17 https://osintcurio.us/

A last example of a link repository is the link collection of Bruno Mortier (@digintelosint) which can be found at https://start.me/p/ZME8nR/osint Bruno has organised his resources in different pages and provides a directory on the first page.

https://start.
me/p/ZME8nR/
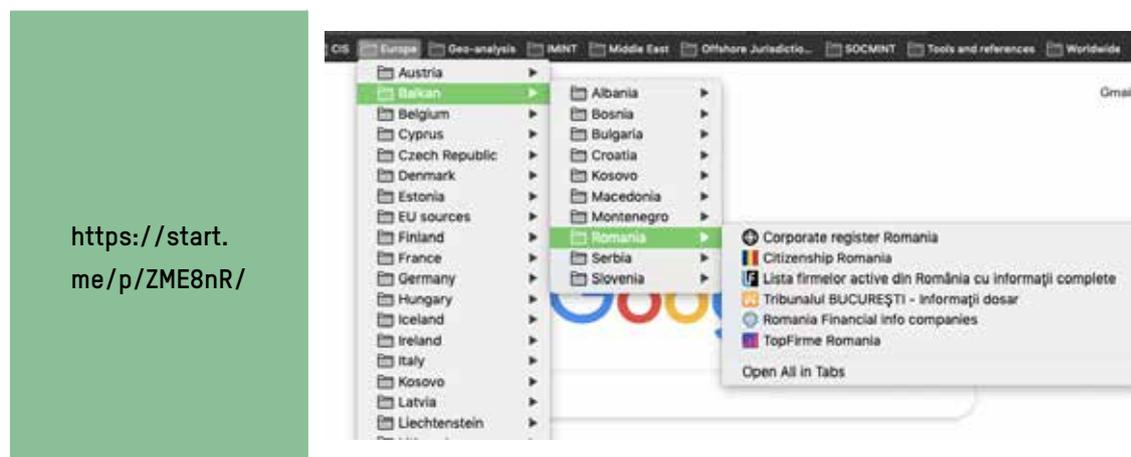
Eventually of course you will build a link repository of your own, or shared within the team. Generally, the most used way to build your own repository of links, is to keep them as bookmarks. Firefox uses a json database file to organize the links which can be exported as a single file. This file can then be easily exchanged and backed-up. Another advantage is that tags can be added to bookmarks in Firefox which given another way to organize your bookmarks. Note however 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 may be altered.

As this guide is not for (law enforcement) investigators, we will not focus on the need for documenting how you found certain information. For an investigator, and certainly if the work is challenged in court, that would be very important. Here we focus on making sure that information hasn't disappeared by the time your story is challenged.

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 you add the URL and a date/time stamp to the footer or header for further proof of the origin of the information.

Sometimes the printing to pdf does not result in a good representation of what your screen shows. In that case, take a screenshot instead. 'Go Full Page' is a Chrome extension which works quite well. 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[18], you can create a pdf from your screenshot and perform Optical Character Recognition (OCR).

18 https://pdf.wondershare.com/pdf-software-comparison/adobe-acrobat-alternative.html

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 document with date and initials of the creator like '2021LB0930-subject-screenshot" If you have not, 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 locally, you always run the risk that the authenticity of your documentation is disputed. Remember that even if pages are not deleted on purpose, studies have shown that one in five articles suffer 'reference rot'. Having a copy of the webpage you found your data on 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 decades ago and attempted to 'archive' the whole internet. In 2001 WayBack Machine was developed and the data became available for everyone.19 Just like "normal" search engines it crawls the internet and saves pages it encounters. We will discuss in chapter 5 its use for searching.

19 https://en.wikipedia.org/wiki/Wayback_Machine

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. At https://web.archive.org/ there is a 'save page now' option 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 then can use for any future publication as a trusted citation.

There are other options to preserved the content of a webpage such as Archive.is/Archive.today or Perma.cc. Perma.cc is a paid source used mostly in the legal sector, however a free account allows 10 references to be saved for free.

There is an extension for both Firefox and Chrome, named 'the archiver'[20] 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.[21] 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 while it is also hosted outside your organisation protecting it from orders to take content down.

Finally DocumentCloud can help in linking your documents to other potentially relevant sources so for an investigative journalist this an indispensable tool to use.

---

20 https://www.cathalmcnally.com/tools/the-archiver/

21 https://www.documentcloud.org/

4 | Operations Security

Especially as an investigative journalist, it is important to keep your data and yourself secure is. This is something that requires your continuous attention when connecting to the internet. There are many books available on that topic so in this guide we will discuss some of the basic stuff for your online research. Please understand that proper operations security entails much more than what we discuss here.

VPN

A VPN, or Virtual Private Network, is, very simply put, a service that extends your private network to a remote server so you will access the 'open' internet only from there. The connection between your computer and the VPN server is encrypted and can be seen as a tunnel.

Using a VPN when connecting to the internet can have three purposes:

▶ shielding your location: the logs on any visited webpages will show the IP address of the VPN server, not of your home or work place/internet service provider. This can be an important security feature.

▶ Keeping your traffic secure which could be handy if you make use of public Wifi. The VPN connection puts an extra layer of encryption over your data which stops others from snooping on the traffic. Also, an important security element.

▶ Changing how websites react to you, for example search engines will show you different results depending on where you search from. This is a very practical side effect of using VPN, you can call it 'virtual travelling'.[22]

One of the most secure and perfectly affordable VPN services is offered by Proton.[23] They have a free version however if you or your organisation can afford 4 euro/month it is recommendable to take a paid subscription. Proton has exit points all over the world and you can easily choose 'where you're from' today.

There are a few downsides to using a VPN. First is that you run more often into captcha's where you have to solve a stupid puzzle before you can access the website. Second is that some websites

---

22 See some more tricks at: https://booleanstrings.com/page/4/

23 https://protonvpn.com/

even completely block access when you come from an IP address which is associated with a VPN provided. Thirdly, your speed may be lower than with a direct connection, however for normal browsing that is usually not a problem.

Research accounts

In open-source investigations we would like to strongly warn against using your personal social media accounts. Most social media platforms in some way or another, 'leak' the identity and details of users who have look at a certain page or profile. So, your interest in a certain subject could be quickly exposed, long before you are ready to write your story.

For most social media platforms an account is needed to obtain (full) access the content so creating research accounts is usually necessary. To avoid that your identity is leaked, you create a research account under a pseudonym, sometimes also called 'sock puppet' or 'persona'.

There may be some legal issues with the creation of research accounts especially in relation to the conditions of the platform. However, your security and the security of your data and investigation prevails over complying with the conditions for the use of the platform. This may of course be different if you are a (law enforcement) investigator as then more legal condition apply to your work.

The most important question when you create a research 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 research account and you use it to access (some) social media platforms only. For such accounts you set up a generic non-personalized email e.g., fggthht768@gmx.us 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 and other sensitive subjects if you want to be able to use this research account for a long time.

Lastly you can create investigative specific research 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 research account well according to the context of the investigation and prepare a data sheet with all personal information of your research account 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 and 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 research account 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 a generic name for the region where you are from and make sure it is consistent also with the age of the research account (some first names are more popular in certain time periods);

▶ Prepare (non)distinct email address in advance, best to have two, if possible, find a local or regional provider although Gmail will do as well. Secure the access to these with 2FA;

▶ Obtain a prepaid mobile number for use in creation of social media accounts;

▶ Photos are an issue -> you do not want to be impersonating someone, stock photos are a risk, so use for example thispersondoesnotexist.com and slightly alter these photo's to avoid detection of the fact that it a computer generated photo;

▶ 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 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 of the social media accounts.

- ▶ Keep a log on the use of the research account and make sure that you 'activate' the research account from time to time.

Of course, when you are using a research account, make sure you're logged out from any other account and your browsers are clear. Even better, use the research accounts only from another laptop or VM.

Password use and 2FA

As we will see in chapter 7, when we discuss the use of data from data breaches, many people are careless by choosing simple passwords and recycle their passwords over many platforms and websites.

If you take OSINT work serious, you have different strong passwords for each platform /account / website. This also applies to the password for your research accounts 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 it 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 (VMs). Most professional OSINT and security researchers that collect of lot of data online, use a VM for their work.

Basically, a VM is a software programme which emulates a full computer system. VM's come in all sizes and shapes and often are used in larger IT environments. However, also on your laptop / desktop you can run a VM and perform your OSINT work from within that 'machine'. This 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.

---

24 https://keepassxc.org/

25 https://authy.com/

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

The best description on how to set up and customise a VM with all kind of tools for OSINT purposes is provided by Michael Bazzell in his OSINT book.[27] The costs are limited to the license for VM player if you use VMWare[28] however there are also free alternatives.

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. Loosing USB flash drives or discussing sensitive matters in public spaces is something that happens all the time.[29]

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

---

27 https://inteltechniques.com/book1.html

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

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

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 form 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 John 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' on a specific data point, 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 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.**

Real Estate Company
→
РИАЛ ИСТЕЙТ КОМПАНИ
→
"RIAL ISTEIT KOMPANI",

TLC Systems
→
ТИ ЭЛ СИ СИСТЕМС
→
"TI EL SI SISTEMS",
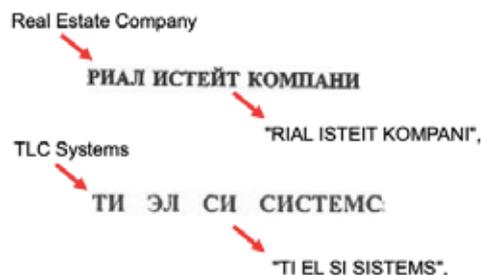
Then language issues often complicate searches and the validation of findings. Google Translate[30] or DeepL[31] are useful tools but take into account that, translations between different language scripts (Latin, Cyrillic, Arabic, Chinese, etc.) may cause confusion and errors. Look at how these company 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.

Another 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.

Basic Boolean operators, AND, OR and NOT.

30 https://translate.google.com/

31 https://www.deepl.com/translator
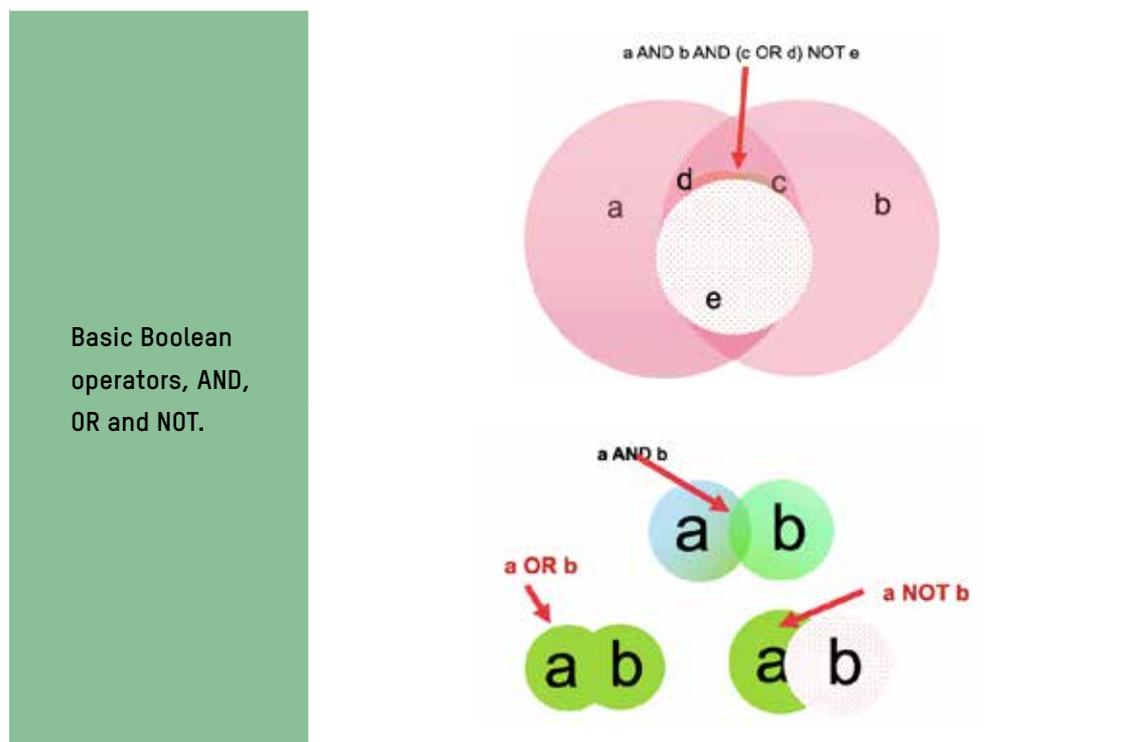
Google

The key feature of Google, or any search engine for that matter, for OSINT work is that it indexes website 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.

The results of a Google search are often unpredictable, unreproducible, and incomprehensible. For consumers, who are generally only looking at the first 5 results, that is not really a problem. However, for researchers and journalists it is different.

Simply said, Google is in essence not a search engine built for people that want to search specific items, it is a marketing machine aiming at the average user which is the average consumer. Advertisers pay to have a link to their site shown based on the search and Google also collects data on its users which is then be sold.

Therefore, you as the user of the search engine, are the product of Google because your data is relevant for the Google advertisers. And the results you see in the index are displayed specific for you as these depend on, amongst other things, whether you have your cookies visible for Google, whether you are logged in or not, from what country and IP you searched. Also, the order of words may cause differences in the results.

Google

Advanced Search

As a professional searcher however, this is not what you want. You want comprehensive and certainly reproduceable results.

However, you have to work with these limitations and understand that Google does not strictly apply Boolean logic, it is more 'Booleish' logic. On the other hand, 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. 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

---

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

33 See for an overview of many possibilities h

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. [34]

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. One good source for all kinds of Google searches is the Boolean Strings blog[35] and another interesting source to watch is a recent webinar by Henk Van Ess.[36]

Other search engines

There are many other search engines besides Google such as Yandex, Bing, Yahoo and many more.[37] 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[38] 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[39] which has indexed the content of open ftp servers and Shodan[40] which searches the Internet of Things (IoT) devices such as webcams and databases connected to the internet. The application of these search engines goes beyond this guide.

---

34 See also https://www.blockint.nl/methods/how-less-is-more-advanced-google-searching/

35 https://booleanstrings.com/2021/02/21/ten-google-tips-for-osint-research/

36 https://www.youtube.com/watch?v=k_LPGCOatJo

37 See https://marcodiversi.com/blog/alternative-search-engines/ for an overview.

38 https://duckduckgo.com/

39 https://proisk.com/

40 https://www.shodan.io/

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 (in theory) the entire internet. Like a normal search engine, it visits websites and crawls it and add this with a date and timestamp to the data base.

These instances can then be queried as shown below:
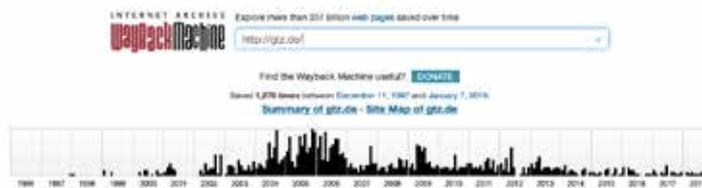
You can search for the root of the domain or for any pages or subdomain. Just add the URL of a website from which you want to see a historical version in the search box and see if it has been captured at some point in time.

The basic 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).

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, please never 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 better use a research account ('persona') as explained elsewhere in this guide.

Facebook page

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.

Searches on
Facebook
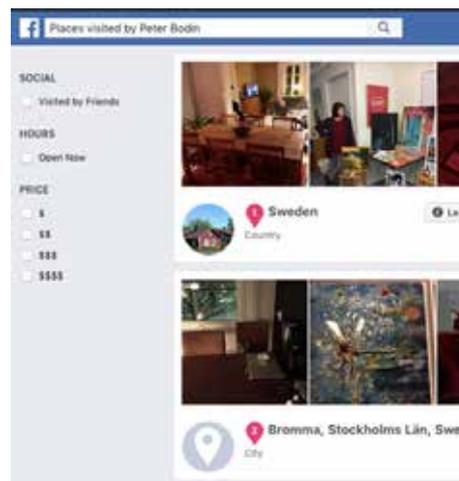
Not only have people become much more privacy conscious and have set their pages to (more) private, also Facebook has changed (and is constantly changing) its search syntax and the way you are able to perform searches.

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 would be to understand how these search URLs are being built by Facebook. The problem with that is that the syntax is changing very often and Facebook masks how the query is built. For example, if you search for photos of bridges in Ljubljana from 1 January 2021 onwards, the search query is:

https://www.facebook.com/search/photos/?q=bridge&epa=FILTERS&filters=eyJycF9sb2Nh
dGlvbiI6IntcIm5hbWVcIjpcImxvY2F0aW9uXCIsXCJhcmdzXCI6XCIxMTE5ODA2Njg4MTk2OTQgXCJ9In0

Therefore, if you need searches beyond what the normal menu of Facebook offers, look at these pages:

http://graph.tips/beta/

https://whopostedwhat.com/

Each of these pages have a bit of a different approach but eventually they use the same way to build the search syntax. For more background, see the Plessas website where this is well explained and kept up to date:

https://plessas.net/facebookmatrix

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 pay-off.[41]

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.

Where Facebook over the past years has decreased the search possibilities, LinkedIn has made it easier. With the following link you can search the whole database:

https://www.linkedin.com/search/results/people/?firstName=&origin=FACETED_SEARCH

---

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

Search database	
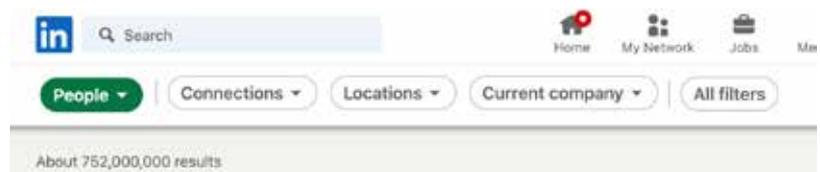
You can search for a name and then filter the results in many ways.

Another of searching LinkedIn, is what recruiters call 'X-Raying'.[42] 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]"

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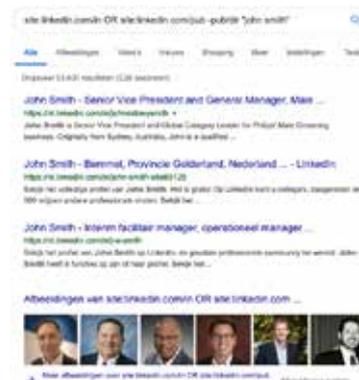
In normal language: you would be searching in the linkedin.com domain only and in the directories / in and /pub with exception of the subdirectory pub/dir and you would be looking for 'name'.

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'.

---

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

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 it could provide an angle 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 make sure the account settings for the LinkedIn account of your research account are set on completely private.

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 the platforms are quite integrated, which we all witnessed when early October 2021 Facebook, Instagram and WhatsApp were simultaneously down for more than 6 hours.

Instagram is 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[43] 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 (https://www.instagram.com/) you need to have an account. Sock puppet rules apply (see chapter 4), however Instagram is less strict in their account creation policy than Facebook.

---

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

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 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. Also there can be multiple numbers all referring to the same location.
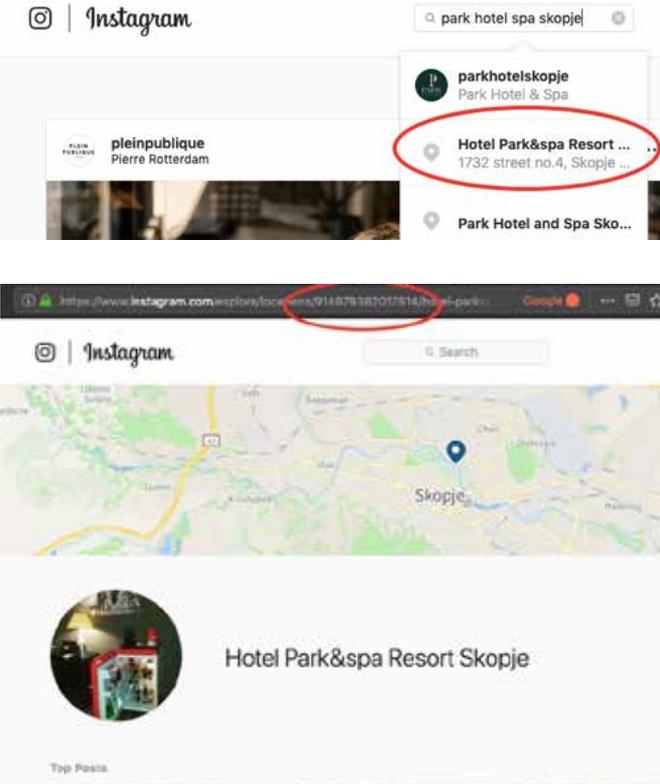
**https://www.
instagram.
com/explore/
locations/**

The URL of the result will give the location ID. Note that location ID in Instagram, can also be used on Facebook.

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 (https://pikram.com) 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.

Another useful source is https://www.searchmy.bio/ through which you can search in Instagram bio's

Downloading of full-size photos from Instagram, also the profile pictures, can be done with https://instadp.org/ and for any research into images this is highly recommended. Alternatively, you can open the 'developer tools' in your browser and see under 'images'.

Finally, there is a paid source, Picodash (https://www.picodash.com) 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 the best known 'micro blogging' platform and although the content of tweets generally tends to be more opinion than fact, some topics are dominated by bots (automated Twitter accounts) and also could include fabrications, there are some good investigative research use-cases for Twitter.

You can use the Boolean 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[44], however, Twitter has fairly decent advanced search facilities available at https://twitter.com/search-advanced
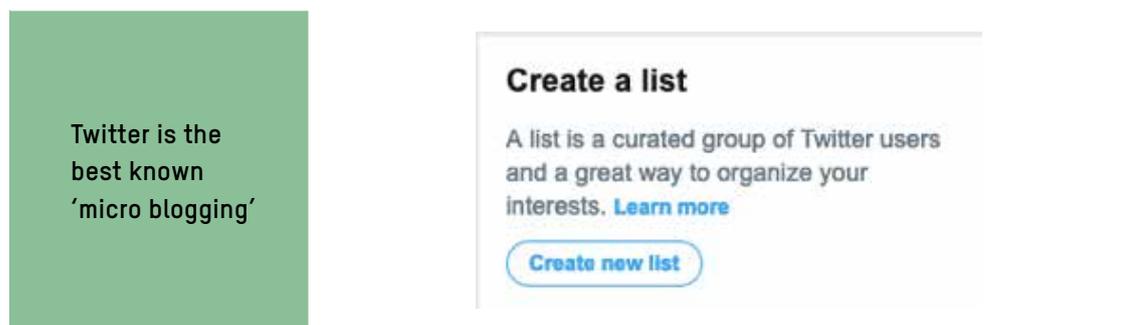
Remember to always verify any tweets you find possibly relevant. Is the content re-tweeted or a screenshot? Who tweeted first, who retweeted, check profile and handle, is it all verifiable and consistent? Be careful with opening links from tweets, preferably in a safe environment.

To analyse accounts there is number of websites available, Quick analysis can be done with https://foller.me/ and more deep analysis of an account with https://tweetbeaver.com If you look

---

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

at a specific tweet and relations, https://socialbearing.com can be useful. For visual searching on location use https://onemilliontweetmap.com

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 (https://www.smrfoundation.org/nodexl/) can help retrieve and analyse larger amounts of tweets.

Twitter is the best known 'micro blogging'

Create a list

A list is a curated group of Twitter users and a great way to organize your interests. **Learn more**

Create new list

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.[45]

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

List tool

Create a new list ✕

List name

Description

Under 100 characters, optional

Privacy

○ Public · Anyone can follow this list
◉ Private · Only you can access this list

Save list

---

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

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.

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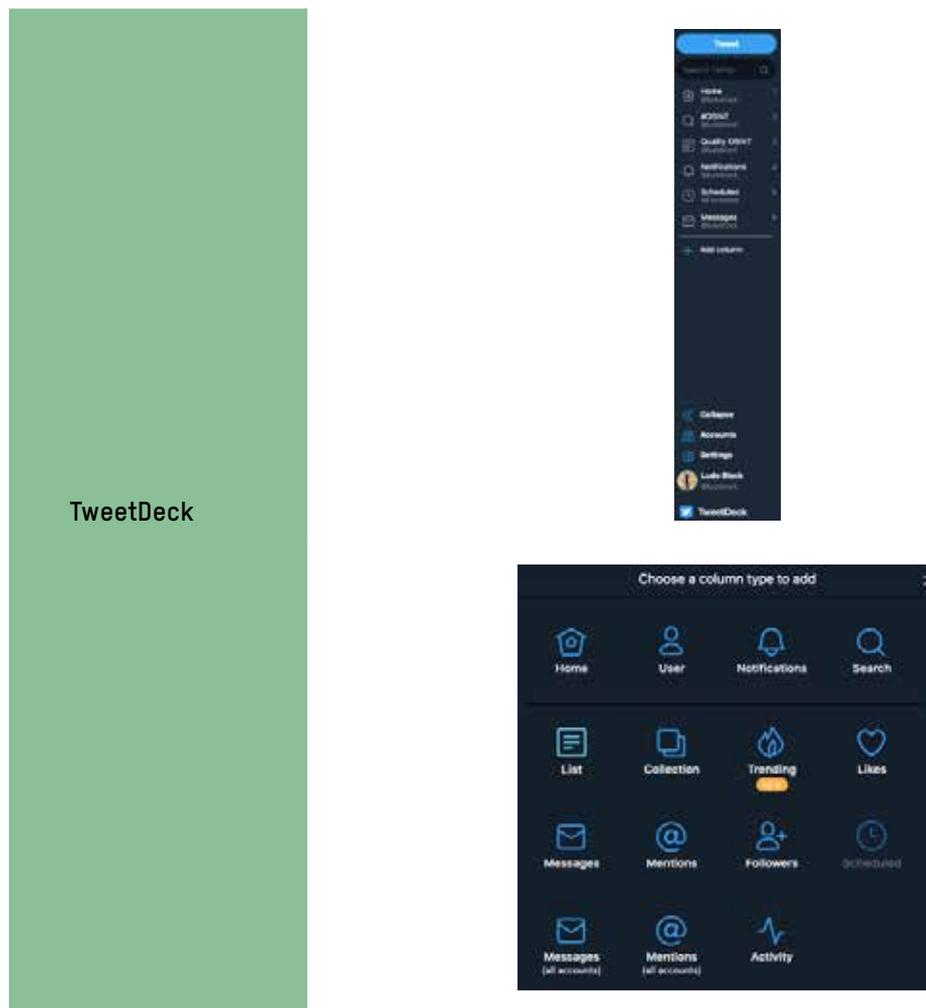
If you want to define one column for a list, add a column and then select 'List' after which you can choose to either have an existing list in that new column or create a new list. From there the creation of a new list works exactly as on the Twitter website.

For those with a bit more technical skills, you can also use the Twint command line tool to collect large amounts of data from Twitter. (https://github.com/twintproject/twint).

7 | People searching

When searching for specific people, the best starting points are usually the identifiers that people use for communication and when connecting to online services. The three most relevant are:

▶ email addresses

▶ user names

▶ 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. Also, people often themselves give away already some information when they create their username, like for example, part of their name, year or day of birth, hobby, etc.

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

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.

There is a large amount of different free email search tools such as Verifalia[46], EmailHippo[47], Epios[48] and ManyContacts[49]. Do not forget to do a search in the search engines. The 'old' Facebook trick to check the existence of Facebook profiles based on email addresses does not work anymore[50].

---

46 https://verifalia.com/validate-email

47 https://tools.emailhippo.com/

48 https://tools.epieos.com/email.php

49 https://www.manycontacts.com/en/mail-check

50 Facebook now requires a phone number to use https://www.facebook.com/login/identify/

Usernames

Very often email addresses, or the first part of an email address, are used as the username for many social media platforms and login for other sites. However, often users (also) have a username or screenname which provides a starting point for research if we want to obtain the full identity of the subject. Interestingly, people usually use the same username across multiple platforms and with proper research that could lead to uncovering their identity. Basically, that research contains a number of steps:

1. Obtain all data from the (social) media) account in use by your target. This includes the ID number if available, the bio and photo or avatar;

2. Do a search in the three main search engines on the "username" and see what results are already available;

3. Check the username search websites such as Instantusername[51], Knowem[52], Namecheckup[53], and Whatsmyname[54]. Basically, these are marketing tools that check whether a username is still available. In OSINT we use them 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.

4. Do a reverse image search on the bio photo of the user in a search engine or use the facial recognition website Pimeyes.[55]

5. Check breach data (see below)

---

51 https://instantusername.com/#/

52 https://knowem.com/

53 https://namecheckup.com/

54 https://whatsmyname.app/

55 https://pimeyes.com/en

6 | Check a commercial database such as Pipl.[56]

7 | Do a content analysis: check the posts of the user to find some personal identifiable information such as addresses, date of birth and any other data you can find which could be relevant later;

8 | Perform a Social Network Analysis.

These steps start from the least effort (2) and go on to the most effort (8) and are best followed in this order.

Breach data

A last option we discuss on email and username research is breach data. As you might know, a lot of data breaches happen where personal information such as user names, email addresses and password are being stolen and made public after some time.
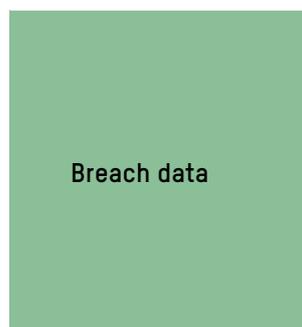
Breach data

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;

56 https://pipl.com/

▶ the passwords (or the hashes of the password) which can be searched as well, help to link different email addressed together.

> **the passwords (or the hashes of the password) can be searched as well**
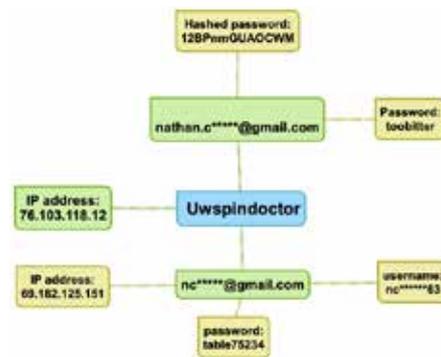
While it is entirely possible to look for these breaches yourself, collect the data and build a database, for basic OSINT use it is much more efficient to use the services offered through for example GhostProject[57] and De-hashed[58] which are the two most used.[59]

Both offer a free search service but for the full results a small fee needs to be paid. At a live demonstration during the training, we started with the username 'Uwspindoctor' randomly obtained from a website and we were able, based on breach data alone to find the information as depicted here on the right.[60]

In case passwords are not yet de-hashed, Dehash.me website[61] a hash/de-hash tool is available. We can use this to quickly either hash a password and search in De-hashed with the hash instead of the password, or see whether a hash found in De-hashed may be turned into the plaintext password.

---

57 https://ghostproject.fr/

58 https://dehashed.com/

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

60 Data depicted here in a redacted and changed slightly form to protect the privacy of the subject.

61 https://dehash.me/

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. There are good tutorials available[62] 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 email addresses people tend not to switch much from phone number which is also the reason that marketing companies use phone numbers as the key identifying piece of data.

Searching by phone number in a search engine not often results in any relevant data. 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). However, always give it a try. Also, try the white pages of the country is available.

If you want to dive in further, please see "Phone numbers investigation, the open-source way"[63] for some tips and examples.

A somewhat tricky way of researching phone numbers is the use of Phone contact book apps such as Get Contact[64] and TrueCaller.[65] 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 should 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.[66]

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

63 https://www.secjuice.com/phone-numbers-investigation-the-open-source-way/

64 https://www.getcontact.com/

65 https://www.truecaller.com/

66 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 used wanted to be, not necessarily the real location. This chapter therefore aims at providing the basic tools and methodology to analyse and verify images and to geolocate these to where 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 is the weather and what was the weather at the alleged date on the alleged location[67], does that match?

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.[68]

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.[69] There are also apps available for on your system.

Also examine whether the image was altered for example with Image Edited?[70] or FotoForensics.[71]

---

67 Use the examples at https://www.wolframalpha.com/examples/science-and-technology/weather-and-meteorology/ to find the weather at that date and location.

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

69 http://exif.regex.info/exif.cgi

70 http://imageedited.com/

71 https://fotoforensics.com/

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[72] 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. 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.
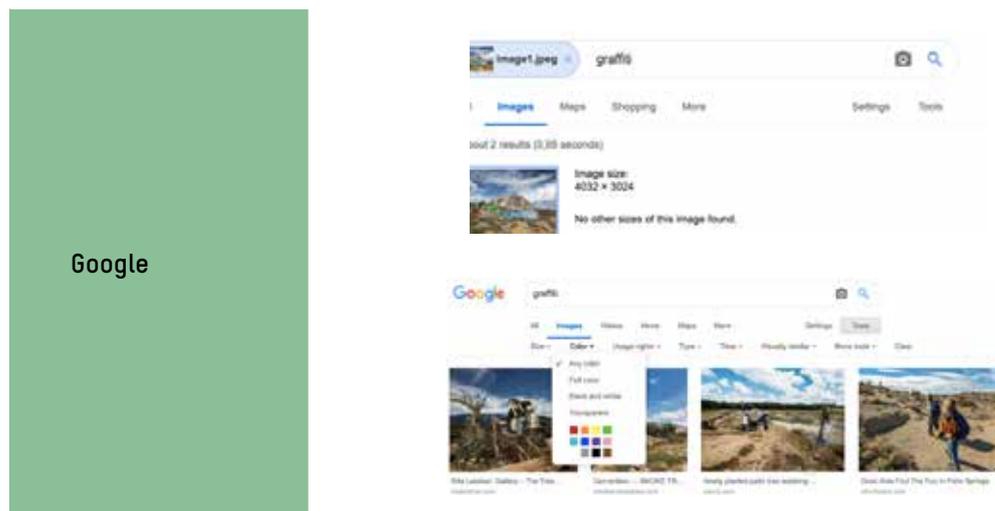
Google

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

Yandex[73] reverse image searching does not have such filters however it very often is better at recognising locations. Bing[74] again shows a different approach and currently has the least applicability (but that can change). Then there is TinEye[75], the site that more of less started reverse image searching but also this site seems much less effective than Google and Yandex.

---

72 https://www.google.nl/imghp

73 https://yandex.com/images/

74 https://www.bing.com/?scope=images

75 https://www.tineye.com

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.

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 relevant tool is DualMaps.[76] 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.

Google maps

But do not limit yourself to Google maps, there are dozens of other map and satellite tools like Bing[77], OpenStreetmap[78], ZoomEarth[79] 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.[80]

If you want to train your abilities, go to Geoguesser[81] on a rainy Sunday afternoon and start practicing.

76 http://data.mashedworld.com/dualmaps/map.htm

77 https://www.bing.com/maps

78 https://www.openstreetmap.org/#map=18/42.00645/21.42727

79 https://zoom.earth/

80 https://nixintel.info/osint/quiztime-9th-may-2019-the-french-prison-system/

81 https://geoguessr.com/

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 investigation. In the basis these registers are organised by (part of) the country where the entities are formally registered. However, there are also (usually 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.

Local registers

The primary source for data on legal entities is formed by the local official registers which mostly are maintained by the government of the country. The registry could be maintained by a separate government entity that is tasked with company registrations e.g., a Companies House like in the UK, it could be a task of the Ministry of Justice or a task of the courts. The registers vary from detailed databases with actual documents available, to a collection of government announcements on registrations and changes of legal entities. Data available varies from only the entity name, number and whether it is active, till a complete repository of all filed documents.

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[82] in the UK and the Danish register[83] where all relevant documents on the legal entities can be downloaded for free, whereas the register on Bermuda shows only the name and number[84]. 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 and Herzegovina, but also Canada where there are provincial registrars as well as a federal registrar.

---

82 https://find-and-update.company-information.service.gov.uk

83 https://datacvr.virk.dk

84 http://companysearch.bz/public_search/

An overview of company registers can be found for example on the website of RBA Information Services[85] and on the Investigative Dashboard.[86]

Aggregated (commercial) registers

Various data brokers maintain aggregated databases with regional or worldwide company information. Most often these registers are in English, 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[87] and LexisNexis[88] and the regional Russia/CIS database SPARK[89].

These commercial solutions can be rather expensive ranging from several thousand euro tot a few hundred thousand euro per year. There is also a whole industry of commercial companies that offer al kind of information. Be sure to check what information is actually available in a report before you buy it.

Free sources

Fortunately, there are also many free alternatives and the most comprehensive free company information source is Open Corporates[90] 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.

---

85 http://www.rba.co.uk/sources/registers.htm

86 https://id.occrp.org/databases/

87 https://orbis.bvdinfo.com

88 https://risk.lexisnexis.com/

89 https://www.spark-interfax.ru/

90 https://opencorporates.com/

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[91] which is very strong in the Middle East and Africa

▶ | ClarifiedBy[92] also strong in the Middle East

▶ | Info-Clipper[93] worldwide (use more to identify companies, their reports are not the best)

A most interesting free source is of course the Offshore Leaks website[94] 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[95] is a great source to quickly search though the available VAT repositories and find company data. Also, the Kazakh government maintains a website[96] which you can use for such searches. Note that you may need to do your own research to identify all relevant sources in a certain geographical area.

---

91 https://www.cedar-rose.com/

92 https://clarifiedby.diligenciagroup.com

93 http://www.info-clipper.com/en/

94 https://offshoreleaks.icij.org/

95 https://tva-recherche.lu/

96 http://kgd.gov.kz/en/services/taxpayer_search

Andrej Petrovski

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[97] 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[98]and Scoring in Serbia[99]. 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.

---

97 https://search.gleif.org/#/search/

98 https://search.bisnode.ba/

99 https://www.scoring.rs/

Regional corporate registries

We have collected the links to the regional registries in the Balkans (as of October 2021). 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 the hosted data can be found and accessed via the internet.[100] Both these components require the use of an intermediary 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.

Hacking
Team
Italy

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.

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 (31 domains)

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.[101]

Historical
Data
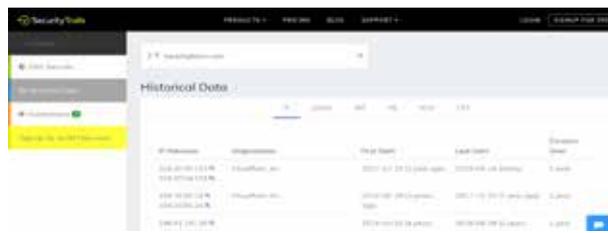
In the *Historical Data* segment there are historical data for different common DNS record types.[102] The records in this case go back some 10 years and shows how the hosting (A record) and email (MX records) have changed, meaning the domain has changed service providers.

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

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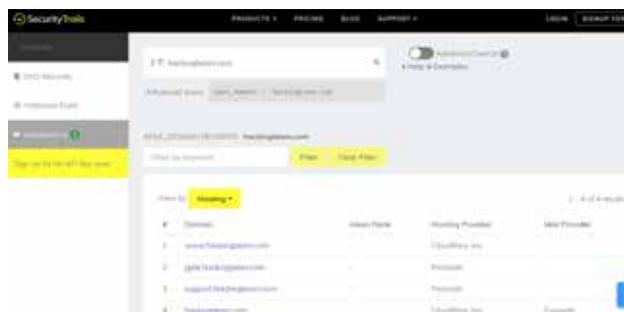
The records here show data about the hosting and email provider for the current domain.

---

101 https://securitytrails.com/dns-trails

102 https://simpledns.com/help/dns-record-types

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.[103]

A very simple service for Reverse IP Lookup is ViewDNS

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

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.

103 https://viewdns.info/reverseip/

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:

website code

```
<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[104] to look for connected websites.

SpyOnWeb

104 http://www.spyonweb.com/

File metadata

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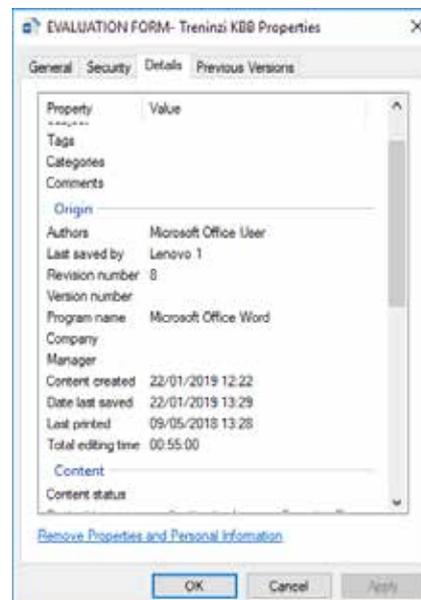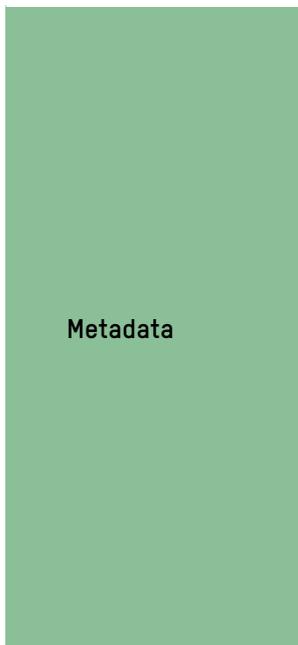
Every document created on any device carries some amount of metadata with it. The amount may vary, but these information might be very useful when trying to understand the context and circumstances in which a document was produced.

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.[105] Besides images, it can also extract metadata from *.pdf* files as well.

Jeffrey's Image
Metadata Viewer

105 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 resources accessible via the internet are not indexed by search engines. The analogy of an iceberg floating in the ocean is mostly used to illustrate this.[106] Only an estimated 5-10% of the information available via the internet is the part which is floating above the surface and is indexed by search engines.

Most of the information however 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 data is not available in a static form however is presented based on a search only. Also, in many cases for access to these sources, some form of authentication is needed.

Then, within that deep web, a space exists that is usually called the Dark Web. The Dark Web is a collection of thousands of resources that use specific software and 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.[107]

There are multiple services and nets that are considered to be part of the Dark Web. These include more known networks like Tor which is operated by many public organizations and by individuals but also small, friend-to-friend and peer-to-peer networks.[108]

The anonymity provided by the Dark Web enables people to (more) anonymously communicate with others either for good or for bad. It may be important and legitimate for activists, journalists or for people living in a country where freedom of speech is not allowed. But at the same time, criminals widely use the anonymity for their illegitimate purposes, for example selling drugs and guns on Dark Web market places.

---

106 https://data-ox.com/web-dark-web-and-deep-web/

107 https://www.wired.com/2014/11/hacker-lexicon-whats-dark-web/

108 https://en.wikipedia.org/wiki/Dark_web

The trouble for investigators (and journalists) is that the data in the Dark Web is not indexed in the same way as in the surface web, and mostly isn't indexed at all. While there are multiple types of services available which are part of the Dark Web, such as Freenet[109] and the 'Invisible Internet Project' (I2P)[110] we will focus on Tor which may be the most useful for OSINT.

Some words of caution. First, please note that in the Dark Web, it is easy to find illegal and disturbing content. 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.

Second, the sites on the Dark Web may contain harmful content for your system. Therefore, using a VM is highly recommended when you are (regularly) visiting the Dark Web, especially when you interact by clicking links and downloading content.

ToR

ToR stands for 'The Onion Router' which allows you to connect (almost completely) anonymously with different services online. This is a decentralised, volunteer-based network and implements encryption of data embedded in various layers, much like an actual onion.

We first discuss the use of the special ToR browser which is the easiest way of connecting to ToR and which you can use for connect to any website. Thereafter we will dive into the Onion Service (previously called 'Hidden Service'), sometimes also called 'Onionland' a reference to the network's top-level domain suffix '.onion'. That part of the internet is only accessible via ToR.

---

109 https://en.wikipedia.org/wiki/Freenet

110 https://en.wikipedia.org/wiki/I2P
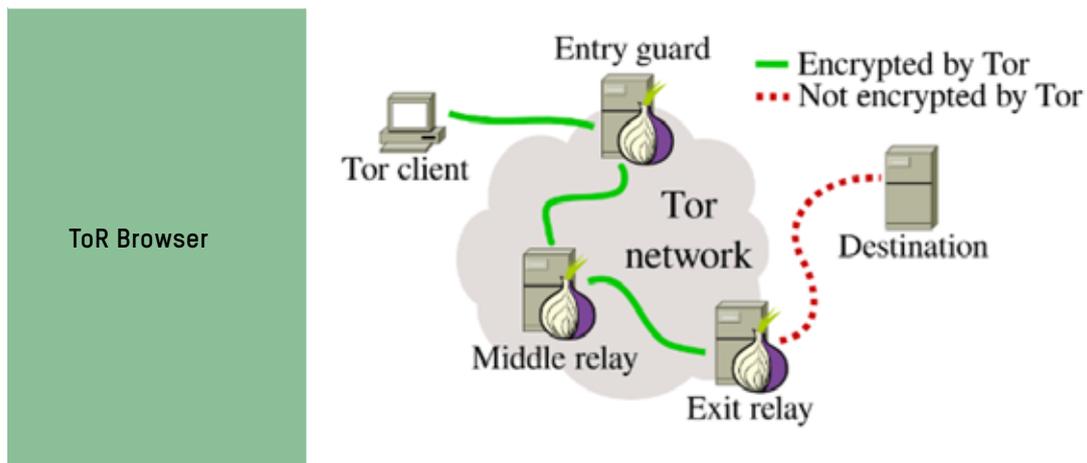
The ToR browser

The ToR Browser can be easily downloaded [111] and installed. The browser is based on Mozilla Firefox but connects to the internet via ToR.

In its' simplest form ToR can be explained as follows. The user connects to the Tor network through an *Entry Node after* which 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.

Once set up, you can visit any website you want via the ToR browser. The speed generally is slower and you may run into (many) captchas. However, you are surfing with much more anonymity![112] You can check if you are using ToR by visiting https://check.torproject.org/ (tip: make it your default Home page in the browser)

The exit node is chosen randomly and will be different every time you start the ToR Browser or when you click 'New Identity' in the ToR Browser. In other words, you can visit any site through the Tor

---

111 https://www.torproject.org/

112 Note that some caveats apply here. Your real IP could be compromised by a DNS leak for example. Also staying completely anonymous does require good operations security hygiene. That goes beyond the purpose of this guide, you might want to do additional research if anonymity is really crucial.

Browser without revealing your true IP address. Please do note that increasingly companies are blocking access to their site for traffic coming from ToR.[113]

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.[114]
7	You can add multiple countries
8	*ExitNodes {xx}, {yy}, {zz} StrictNodes 1*
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.

---

113 See: https://blog.intelx.io/2020/07/05/why-we-are-going-to-block-tor-ips/

114 http://archive.is/jiYA9

Andrej Petrovski

| 10 | Save the file, restart Tor |
| 11 | Now your exit node is in the country you specified |

Onion Service

The Tor network has its own (pseudo) *TLD - Top Level Domain,* which is *.onion.* Websites with an address ending in .onion are part of what is called the Onion Service or 'hidden services'. Simply said, the servers where the content of these websites is hosted, are only accessible via ToR which makes them unfindable. Well, nearly unfindable as law enforcement have made some successes in taking Onion Service market places selling illegal goods down.[115] Still, breaking that anonymity takes considerable effort so generally criminals exploit hidden services for basically anything (illegal) that money can buy.[116]

The addresses of .onion websites first were exact 16 characters, which was under Onion Service version 2. As per 15 October 2021 version 2 will be disabled[117] and only the new (version 3) .onion addresses with exact 56 characters long will work. Remember, these websites are only accessible when you use the ToR browser or when your system is connected in another manner to the ToR network.

Finding data from ToR Hidden Services

As noted, the hidden services, recognisable by their .onion domain, are not indexed by the large search engines. However, there is a number of services that provide some kind of access to the content of ToR Hidden Services. There are onion sites that attempt to act like a search engine, there are sites on the surface web that contain indexed content from Onion Services and there are catalogues that keep track of (new) websites hosted as an Onion Service. Let's explore some of these.

---

115 See https://en.wikipedia.org/wiki/Silk_Road_(marketplace)

116 See https://www.lawfareblog.com/tor-hidden-services-are-failed-technology-harming-children-dissidents-and-journalists

117 https://blog.torproject.org/v2-deprecation-timeline

Search engines

There is a number of search engines available on the Onion Services. The way these search engines work is quite similar to their regular counterparts however they only index content from the Onion Services. Don't expect the millions of results that you may get when searching for something via Google, usually a few thousand hits are the maximum you receive. The four most used at the moment (October 2021) are:

TOR66	tor66sewebgixwhcqfnp5inzp5x5uohhdy3kvtnyfxc2e5mxiuh34iid.onion
GRAMS	jubfkzevojqh5x2v6nacdlrrmsfps65favaeqhfvc3uo5dxzxg3m4sqd.onion
LIGHT	i26iai6rgwckuubcsvgobp6tuevgmmkadj2qgogiti55xkzm6ud3tyqd.onion
TORCH	srtwuinx6xvgmwtu6xfgjqa35orqxi4obtw6o7gspfubruqnmi54m3qd.onion

Grams

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 compared to the surface web[118] and searching onion site other than for very specific research will generally not give you much relevant results.

---

118 https://www.recordedfuture.com/dark-web-reality/

Wiki and catalogues

Other than real search engines there are also a number of .onion sites which keep lists of useful sources, which looks sometimes a bit like the 'start.me' and wiki pages you'll find on the surface web. Such curated content could be very interesting as someone may already have done the hard work and discarded all useless or untrustworthy sources. A disadvantage is that these pages are often ridden with adverts.

A well-known page in this category is the so-called Hidden Wiki: kfj2am4ee2asdqflt4tuxxwbeuzmh6tv64ojbqscc4u55skrechsxzad.onion which also keeps track of reported scams.

Another source which is more recent and very interesting is the Ransomware Group Sites which provides an overview of, at the moment, over 40 ransomware groups and their pages, and can be accessed via ransomwr3tsydeii4q43vazm7wofla5ujdajquitomtd47cxjtfgwyyd.onion

On these pages different ransomware groups dump data they copied from the servers they attached. If the victim of the ransomware does not pay up, the ransomware group dumps the data online.

Ransomware Group Sites

Group Name	Onion V.	Link
Arvin Club	v3	Open
AtomSilo	v3	Open
AvosLocker	v3	Open
Babuk	v3	Open
BlackByte	v3	Open
Bl4ckt0r	v3	Open
BlackMatter (former Darkside?)	v3	Open

Warning: tread very carefully. Expect the data dumped by ransomware groups to contain malware and access / download that data only in a VM!

Surface web sources

Lastly there is a number of services on the surface web which have crawled and indexed Onion Services content and present that content on the surface web. Examples are https://www.tor2web.org/ https://darksearch.io/ however it should be noted that their coverage is usually limited.

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.

Cleaning

```
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[119], a common tool for cleaning data. After a successful download and unzipping at the desired location we open the application *Openrefine*.

OpenRefine	

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

input file	

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:

Cyrillic characters	

119 http://openrefine.org/

In order for OpenRefine to show the Cyrillic characters in a proper character encoding.[120] In this case we need to choose *UTF-8* in the encoding section

UTF-8 in the encoding section	**Select Encoding** Common Encodings — All Encodings **Encoding** — **Aliases** ISO-8859-1 — 819, ISO8859-1, l1, ISO_8859-1:1987, ISO_8859-1, 8859_1, iso-ir-100, latin1, cp819, ISO8859_1, IBM819, ISO_8859_1, IBM-819, csISOLatin1 US-ASCII — ANSI_X3.4-1968, cp367, csASCII, iso-ir-6, ASCII, iso_646.irv:1983, ANSI_X3.4-1986, ascii7, default, ISO_646.irv:1991, ISO646-US, IBM367, 646, us UTF-16 — UTF_16, unicode, utf16, UnicodeBig UTF-16BE — X-UTF-16BE, UTF_16BE, ISO-10646-UCS-2, UnicodeBigUnmarked UTF-16LE — UnicodeLittleUnmarked, UTF_16LE, X-UTF-16LE UTF-8 — unicode-1-1-utf-8, UTF8 Cancel

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 ";".

Custom ";".	Columns are separated by ○ commas (CSV) ○ tabs (TSV) ◉ custom: ;

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

Create project	Project name: Money laundering — Tags: — Create Project »

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...*

---

120 https://www.w3.org/International/questions/qa-what-is-encoding

**Split into
several
columns**

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

**Specify "-"
as a separator**

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.


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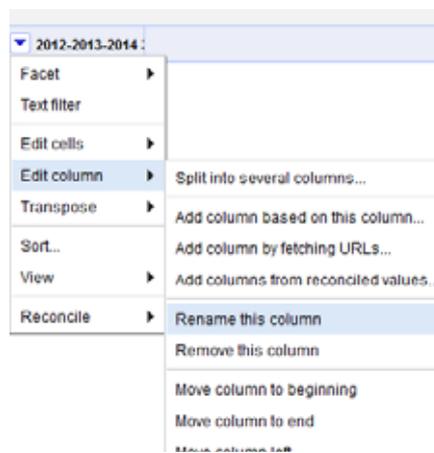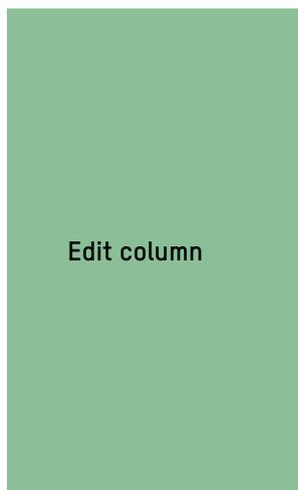
We notice that there are a few duplicate rows in the file. Since this can lead to false conclusions, we need to remove the duplicate rows. This is easier to do in Excel, so we export the our data in Excel format.

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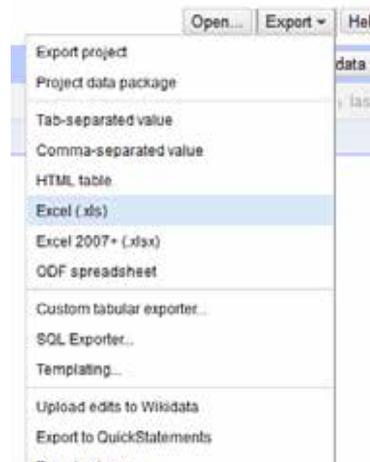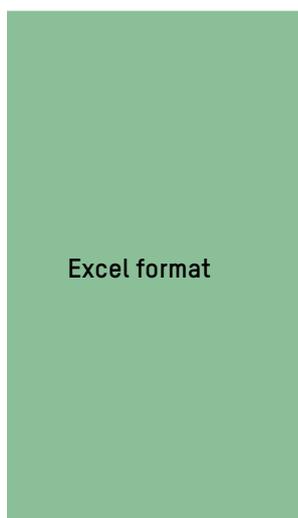
Once the file is open in Excel, navigate to the *Data* tab, to the *Data tools* section, and chose *Remove duplicates.*

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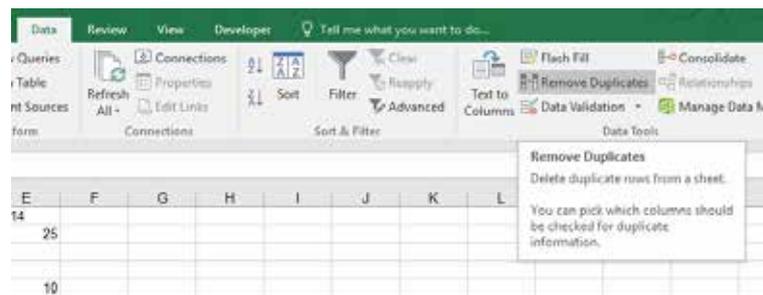
In the remove duplicates we choose which columns should be checked for similar values. Since we only want to remove rows that are identical to other rows, we check all the columns.

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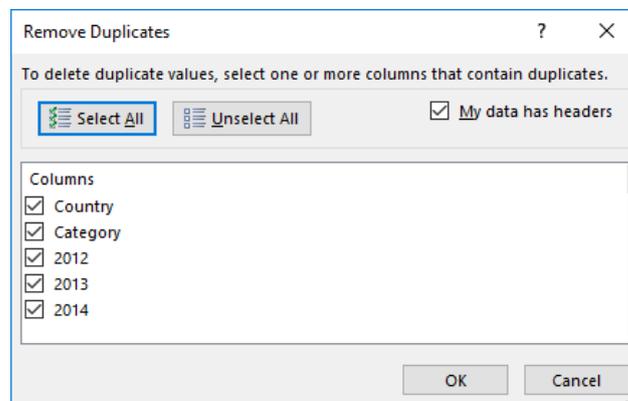
This operation will remove all the duplicates, but we are still left with one empty row (containing data only in the first column) which we don't need, so we delete it the conventional way, by *right-click* and *Delete.*

Clean data

Now we have a more or less clean data.

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...*
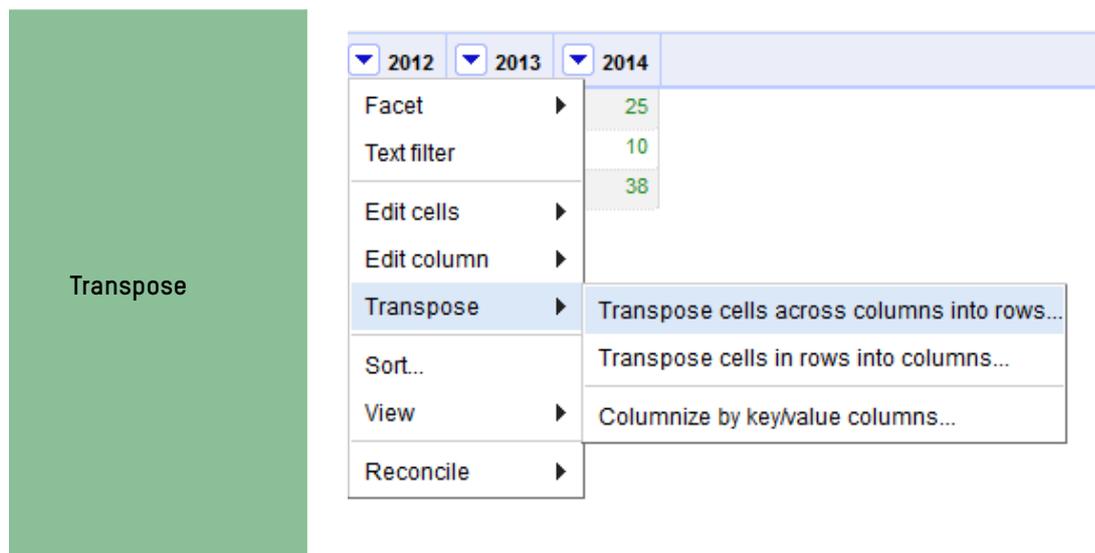
Transpose

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.*

Transpose cells across columns into rows

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

9 rows