

# Systematically Searching for Identity-Related Information in the Internet with OSINT Tools

Marcus Walkow and Daniela Pöhn<sup>a</sup>

Universität der Bundeswehr München, Neubiberg, Germany

**Keywords:** OSINT, Open Source Intelligence, Taxonomy, Identity, Attack.

**Abstract:** The increase of Internet services has not only created several digital identities but also more information available about the persons behind them. The data can be collected and used for attacks on digital identities as well as on identity management systems, which manage digital identities. In order to identify possible attack vectors and take countermeasures at an early stage, it is important for individuals and organizations to systematically search for and analyze the data. This paper proposes a classification of data and open-source intelligence (OSINT) tools related to identities. This classification helps to systematically search for data. In the next step, the data can be analyzed and countermeasures can be taken. Last but not least, an OSINT framework approach applying this classification for searching and analyzing data is presented and discussed.

## 1 INTRODUCTION

The software company LastPass examined the password behavior of individuals (LastPass, 2021). According to them, 92 percent know that it is risky to use passwords more than once. Nevertheless, 65 percent always or mostly still use the same password or variations. While financial accounts primarily receive stronger passwords (68 percent), work-related accounts and medical records do not (32 resp. 31 percent). For only 8 percent of the participants, a strong password should not be tied to personal information. According to (Zhang et al., 2010), it is possible to predict changes to the password. Consequently, searching for personal information on the Internet may lead to a valid new password. This is even more serious as attacks are increasing, leading to further credentials and personal data being compromised (Verizon, 2022). In organizations, not only one but several digital identities are managed in the identity management system. Typically, users have further accounts, such as web services, where information or credentials can be leaked. Hence, one compromised account in the organization can result in a wider attack.


Open-source intelligence (OSINT) can tackle the problem of the personal factor in passwords and fallback mechanisms. The more knowledge is found about the individual user, the greater the probability

that the authentication factor can be derived from it. Hence, the results of a systematic search can warn the user before an incident happens. In order to systematically search for data, a classification is required. In addition, a modular open-source framework helps to apply this classification. The contribution of the paper is two-fold: 1) a classification of data related to identities and identity management systems and 2) an open-source OSINT framework approach based on the classification. This can be utilized to identify possible problematic information.

The paper is structured as follows: Section 2 provides an overview of the related work. Section 3 introduces and structures OSINT search. The classification is applied by an OSINT framework approach in Section 4. The approach is then discussed based on a real-world example in Section 5. Section 6 concludes the paper and provides an outlook on future work.

## 2 RELATED WORK

Several authors describe OSINT in general. For example, (Pastor-Galindo et al., 2020) provide an overview of OSINT with the basic workflows (collection, analysis, knowledge extraction). Additionally, the authors categorize analysis (lexical, semantic, geospatial, social media) and information (personal, organizational, network). (Martins and Medeiros, 2022) propose a taxonomy for threat intelligence

<sup>a</sup>  <https://orcid.org/0000-0002-6373-3637>

sharing, which is of limited use for our purpose. The Malware Information Sharing Platform (MISP Project, 2022) uses, among others, the categories of blog posts, reports, presentations, news, forums, mailing lists, repositories, and other sources. (Azevedo et al., 2019) further detail the steps of clustering and correlating data, while (Hickey and Arcuri, 2020) explain OSINT including web applications, passwords, and emails. Like many other approaches, the authors focus on generic threat intelligence.

Only a few approaches target OSINT for identities. (Butler et al., 2016) present REAPER, a tool for automated mass credential harvesting. Related to that, (Fang et al., 2019; Peng et al., 2019; Bermudez Vilalva et al., 2018) describe the effects of a password leak. The site Have I been Pwned (Hunt, 2022) applies a similar approach to warn of password leaks. Social media platforms have changed the way people communicate with each other. At the same time, they are an interesting source for further actions. (Kanta et al., 2020; Kanta et al., 2022) propose a concept to generate individual password lists based on data gathered by OSINT search. (V. et al., 2020) go a step further by searching the Internet for information about names, mobile numbers, and email addresses. The authors apply different people's search engines focusing on social media. Similarly, (Sharad Sonawane et al., 2022) performs account matching, extracting user metadata to generate a single report. (Akhgar et al., 2017) uses geographic, statistical, and other public sources, while (Gibson, 2016) speaks of unstructured and structured data as well as the type of procurement and the origin of the data.

Especially on GitHub, different OSINT tool lists can be found with (Cyber Detective, 2022) being the most comprehensive. The author lists, e. g., the categories maps, geolocation, and transport; social media; text/sound/video analysis; image search and identification; cryptocurrencies; messengers; search engines; datasets; passwords; emails; nicknames; phone numbers; contact and leak search. OSINT Framework by (Nordine, 2022) visualizes different OSINT tools by grouping them into 32 categories, e. g., username; email address; images/videos/docs; social networks; instant messaging; people search engines; dating; phone numbers; public records; forums/blogs/IRC; archives; digital currency. Similarly, (Bielska et al., 2020) lists 7,600 tools and services. Two well-known OSINT tools with open-source and commercial variants are SpiderFoot and Maltego. Although both offer modules related to identities, their main target are domains and networks. Maltego Community Edition (CE) only has one person-related machine, searching for email addresses, whereas SpiderFoot Open Source

offers more search options. In addition, full functionality is only available in the paid versions. This shows that further work is required to better protect individual users and organizations with relatively low costs.

### 3 OSINT SEARCH

In order to search for compromised identities and further information, which could lead to that state, relevant data first needs to be explored. In the next step, vulnerabilities can be fixed and data be removed to reduce the number of successful fraud attempts. Hence, the goals are the stages of identity research and cyber reconnaissance. We classify possible sources in a systematic way. First, we detail identity search in Section 3.1. As identity management systems require additional inspections, these are explained in Section 3.2. Next, we describe all-in-one search tools. Last but not least, we show helpers, which aid in the search and visualization (see Section 3.4). These sources and helpers can be applied for an extensive OSINT search, using all the different information.

#### 3.1 Identity Search

The data requested during registration (e. g., usernames, email addresses, phone numbers, and personal information) can be leaked. Other data, such as relationship status and hobbies, can be used for social engineering and are, therefore, particularly interesting for security issues. Even though multi-factor authentication is increasingly applied, it can be circumvented. Therefore, it is important to reduce published data, described next, which can be found in the following sources (Cyber Detective, 2022; MISP Project, 2022; Hickey and Arcuri, 2020):

**Social Media:** Social media intelligence (SOCMINT) is a sub-branch of OSINT and refers to the information collected from social media websites. The data available can be open to the public or private (cannot be accessed without proper permissions). The content comprises posts/comments, replies, multimedia, social interaction, and metadata.

**Search Engines:** Main search engines used by users can be repurposed for OSINT. In addition, meta and specialty search engines are available.

**Public Media:** News from newspapers, radio stations, etc. are published online. News digest and discovery tools try to combine specific news.

**Public Records:** Reliable and legitimate source of

information, e. g., registration of a person or financial data of a company.

**Repositories:** Codes, snippets, documentation, and other information is published at public repositories, such as GitHub.

**Archives:** Website history and capture sites take snapshots of websites that will remain online even if the original page changes or disappears.

**Leak Pages:** Pastebin and alternative Pastebin-type sites contain leaks. These leaks are then checked by specific leak pages.

**Dark and Deep Web:** Another source for leaks is dark and deep web pages, either information in forums or specific web services.

**Further Internet Pages:** This includes forums, blogs, academic resources such as publications, cryptocurrencies, and all other Internet pages.

### 3.1.1 Registration Data

**Email.** Email addresses are often used as a substitute for self-chosen usernames or phone numbers. They immediately offer the advantage of an address for the confirmation link. Different tools search for email addresses or check whether an email address exists (Hickey and Arcuri, 2020; V. et al., 2020; Cyber Detective, 2022). This includes `Snov.io`, `Hunter.io`, `Skrapp.io`, `Prospect.io`, `breachchecker.com`, `spycloud.com`, and `haveibeenold.app`. In addition, `haveibeenpwned.com` searches by email address for leaks.

**Username.** Especially at the beginning of Web 2.0, self-chosen user names were common for logging in (V. et al., 2020; Cyber Detective, 2022). Often, users choose a name, or a variation thereof, with which they have a personal connection. Hence, they often reuse it in the same or in a modified form for other registrations. There are two types of online tools: 1) check whether a profile page exists on various social networks, such as `whatsmyname.app` and 2) create possible usernames based on entered names, for example, `namecombiner.com`.

**Password.** As users tend to reuse passwords, `haveibeenpwned.com` lists leaks based on email address. The leaks though can be found at paste sites, dark and deep web (Hickey and Arcuri, 2020; Butler et al., 2016; Fang et al., 2019; Peng et al., 2019; Bermudez Villalva et al., 2018; Hunt, 2022). In addition, default passwords and password crackers are available online.

**Phone Number.** Eliminating the ownership factor simply by knowing the phone number requires additional technologies. However, there are hardly any online services that link a mobile phone number with a name or email address. A classic method is the telephone book, which mainly publishes landline numbers. The latter can be used, for example, via an SMS for multi-factor authentication if no mobile phone number is available. For practical attacks, landline number cloning is more complicated than mobile number cloning. Some online services provide meta-data, such as the provider for a specified phone number (V. et al., 2020; Cyber Detective, 2022). There is the option of querying cell phone numbers that have been found via Google Dork. The phone number can also be read from social media using suitable crawlers or online services. A possible tool for Instagram, for example, is `istaunch.com`.

**Address.** Personal information such as postal (shipping) addresses can often be found in identity management systems of organizations (Cyber Detective, 2022). This information can be collected online after a leak. In addition, telephone books and public administrations provide further sources. For example, addresses are included in criminal and traffic registers and property searches in the US. `Hitta.se` is a Swedish search engine that offers telephone directories, addresses, and maps. Last but not least, search engines collect information.

### 3.1.2 Further Data

**Texts and Relationships.** Information about social relationships (personal and organizational (Pastor-Galindo et al., 2020)) can provide a plausible background story for social engineering attacks or answers to security questions. Depending on the country, different networks dominate the market (e. g., Russia VKontakte). In consequence, several tools are specialized (Cyber Detective, 2022). As an example, `instahunt.co` looks for usernames in Instagram, while crawlers such as `Osintgram` automate the quest. In contrast, meta-search engines explore different social networks, search engines, archives, and other websites in their forwarded search queries. `yasni.de`, for example, focuses on German-speaking countries, `spokeo.com` addresses the US, and `social-searcher.com` can be used internationally. In order to provide additional background information, further searches, such as about cryptocurrencies (e. g., with `blockcypher.com`) can be applied.

**Image, Video, and Sound.** Users upload several pictures and other material to social media and specific pages (Cyber Detective, 2022). Faces, objects, and logos be recognized in a photo using Google Vision (Google, 2022b) or Microsoft's face recognition API (Microsoft, 2022a). Tools such as `reversesearch.com` can reverse search or analyze the image, e.g., with `Sherloq`. `Huntel.io` and further tools analyze the geolocation if published by exchangeable image file format (EXIF) data. Political information, maps, etc. help to locate the material.

## 3.2 Technical Search

Cyber reconnaissance is a technical investigation aiming to provide attackers with as much information about the target as possible. This includes which (identity) software is being used. On the other hand, publicly available data can be browsed. Hence, the following sources can be utilized (Cyber Detective, 2022; MISP Project, 2022; Hickey and Arcuri, 2020):

**Social Media:** SOCMINT refers to the information (text, multimedia, interaction, metadata) collected from social media websites.

**Search Engines:** Main, meta, and specialty search engines search for selected topics.

**Public Media:** Online news from original sources.

**Public Records:** Reliable and legitimate source of information, e.g., financial data of a company.

**Repositories:** Data published at public repositories, such as GitHub.

**Archives:** Snapshots of sites taken by archives.

**Leak Pages:** Information about leaks.

**Dark/Deep Web:** Information below the surface.

**Further Internet Pages:** All other Internet pages.

**Organisation Website:** Organisations provide information online via their organization websites, such as email addresses, roles, and persons.

**Network:** The organization's network and servers offer information about insecure systems, used operating system and software versions, and internet protocol (IP) addresses.

### 3.2.1 Unsecured Data

Unintentionally leaked information, such as application programming interface (API) keys, credentials, and internal information, can be used during the attack lifecycle. With self-built scrappers, crawlers or uniform resource locator (URL) fuzzers such as (Wilkening et al., 2022), and Google Dorks, publicly accessible folders, files, and data are displayed.

### 3.2.2 Network Data

**Network Scanner.** In order for the identity management systems to be screened using OSINT, the associated hardware must be found on the Internet. `Shodan.io` systemically asks for relevant ports and publishes the results in a queryable format (Daskevics and Nikiforova, 2021). `Censys.io` provides a similar service. Network scanners such as the Network Mapper (NMAP) can be used to find out more about the IT infrastructure.

**Application Testing Software.** Tools such as Burp Suite examine the website of the identity management system. The Burp Suite extension security assertion markup language (SAML) Raider focuses on the federated identity management protocol SAML. The Open Authorization (OAuth) scanner extension detects misconfigurations in the protocol implementations of OpenID Connect and OAuth.

**Security Scanner.** If the identity management system software is known, databases such as `exploit-db.com` display known vulnerabilities and exploits (Cyber Detective, 2022). So-called security scanners such as the Open Vulnerability Assessment Scanner (OpenVAS) thoroughly test the server behind it for possible vulnerabilities. `pentest-tools.com` provides a collection of such security scanners.

## 3.3 All-in-One Search Tools

All-in-one search tools reuse the tools listed above and combine the results across group boundaries (Cyber Detective, 2022). Recon-ng, SpiderFoot, TiDOS, The Harvester, and Maltego are comprehensive representatives. In the case of Maltego and SpiderFoot, the range of tools differs depending on the version. APIs for paid services such as Social Links CE can be integrated into Maltego CE and SpiderFoot HX. In the full version, external services such as `pip1.com` or People Data Labs are purchasable. Although these tools provide all-purpose searches, such as social media, search engines, dark web, and leak pages, their main focus is on organization networks.

## 3.4 Helpers

Due to the huge amount of data that can be found on the Internet, advanced techniques are needed to analyze the data and make a pre-selection.

### 3.4.1 Machine Learning

Machine learning is suitable for this task. Thereby, the collected photos can be evaluated by various social media and provide new insights into identities that were not yet obvious through research. Valuable information is also found in (short) messages and other texts on the Internet, where machine learning algorithms help to extract keywords and analyze the context. Microsoft’s Text Analytics (Microsoft, 2022b), IBM’s Watson API (IBM Developer, 2022), and Google’s Natural Language API (Google, 2022a) provide such analysis services.

### 3.4.2 Natural Language Processing

The aim of NLP (Noubours et al., 2013) is to process natural language and thereby be able to grasp the meaning of texts and language. Just like people, a computer should have eyes and ears to pick up speech and analyze it with the brain, convert it into code or text, and then process it. In NLP, the problem is best addressed through deep learning models, where sufficient learning material is available due to large data collections. For the purpose of the paper, named-entity recognition (NER) (Yang and Lee, 2012; Al-Moslmi et al., 2020), sentiment analyzes (Notz et al., 2019), and text generation (Lee et al., 2022) are of particular interest. A current text generator is a generative pre-trained transformer (GPT)-3 by OpenAI.

## 4 EXAMPLE AND CASE STUDY OF AN OSINT FRAMEWORK

This section describes our open-source OSINT framework to search for identity-related information (see Section 3.1). In addition, the technical search detailed in Section 3.2 can be used if the target is an organization. The framework exerts the workflow described by (Pastor-Galindo et al., 2020): Data collection (see Section 4.2), data analysis (see Section 4.3), and knowledge extraction (see Section 4.4).

### 4.1 Overview

Our OSINT framework has a graphical user interface (GUI) for interaction with the user. Thereby, the user can select different modules for their search. The modules are implemented or attached in the backend, which interacts with the storage (using a pre-defined folder structure) and database. The search results are then displayed in graphs. In order to realize the interactions, the framework Dash was chosen.

Figure 1 provides a brief overview of the GUI. In the top line, new values (e. g., names, identities, email addresses) are added. This can be combined with modules, which come next. In the big frame below, the results are displayed. In the example within the figure, a node with an image was selected and passed to a module. This evolved into new nodes with further information. In the next section, Figure 2 details the overview on an example. Thereby, the workflow can be iterated.

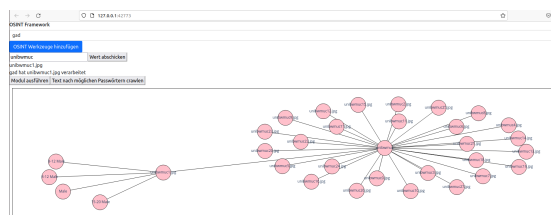


Figure 1: OSINT framework.

### 4.2 Data Collection

For the data collection, we use our own and external tools based on Google Dork, scrappers, and crawlers for various sources including social media sites. For example, if a name is entered, possible email addresses are generated and then checked for existence. Next, those valid email addresses are then used to search for social media accounts with Sherlock (Sherlock Project Team, 2022). Based on the results, different crawlers download texts, images, and videos as well as further data. The raw data is either written directly to the project database or stored in the respective folders. In the next step, images and texts are included to be analyzed with suitable tools. Thereby, the phase data collection especially focuses on usernames, texts, and relationships, as well as media. If addresses and phone numbers are part of the social media profiles, then these are also collected.

### 4.3 Data Analysis

The data analysis again uses external and implemented tools. For example, images are analyzed for geospatial data in EXIF format. Using an API, Google Vision is supposed to recognize texts or faces in images. Images can be further analyzed for location information, such as buildings. The text analysis utilizes an API to Microsoft Text Analysis and the NER extraction tool for the German language. A list of found tokens is returned via the API. In order to receive full words, the words associated with the token are searched in the original text. Results from the analysis are transferred to the database for knowl-

edge extraction. This phase focuses on texts and media, although the selected words are used as input to generate possible passwords. An attacker could apply these passwords for brute-force attacks. In a defensive scenario, the results may help to rise awareness and improve current passwords.

#### 4.4 Knowledge Extraction

For the extraction of knowledge from the images and texts, the APIs of machine learning algorithms provided by Microsoft and Google are applied. Thereby, emotions among others can be discovered. In order to receive age and sex/gender, two convolutional architectures for fast feature embedding (CAFFE) models with an OpenCV library are used. Here, the pictures are transformed into binary large objects (BLOBs) and transferred to the deep neural network (DNN) of the CAFFE model. All liable results serve as input for, e. g., GPT-NeoX to generate text messages, which could be used by attackers. In addition, possible passwords are created by a custom wordlist generator. This shows, that knowledge extraction requires the described helpers in Section 3.4.

## 5 DISCUSSION

We discuss our OSINT framework based on an exemplary search, the intended usage, and a brief comparison with other OSINT tools.

### 5.1 Applying the OSINT Framework

To explain how the OSINT framework works, an exemplary search was conducted on German Chancellor Olaf Scholz. This was limited to gender age detection (GAD) of images from his Instagram page and NER of his tweets. After defining the target person of the search, the new node "olafscholz" as the username for Instagram was entered. This is possible as a self-built Instagram crawler was added via the corresponding menu. The framework was informed that the crawler needs information in string type as input, saves data as result, and inserts the file names as nodes. With this self-built crawler, all images were downloaded.

In Figure 2, 19 found images are displayed. Depending on the results, this overview can get too crowded. In the future, this framework will provide better placement and formatting of the nodes and edges; this may include a reduction of results by grouping.

### 5.2 Data Analysis

For the data analysis, a photo of the results (olafscholz10.jpg) was selected via its representative node and examined with the ML application GAD. Unlike the crawler, GAD requires images as input and returns information. The results of olafscholz10.jpg are inserted as two new nodes on the graph. For Twitter, the tool `vicinitas.io` is used. The tweets are stored as text files in the "Files" folder and a node is added to the graph.

By selecting the node and the German NER ML analysis tool, after pressing the 'crawl text file for possible passwords' button, all tweets are examined for named entities. These appear below the graph and are stored in a text file. In later versions, the results will be sorted by frequency and by sentiment analysis according to emotional significance. The text file can be used by programs such as Hydra to reduce the time for a brute-force attack. The assumption behind this is that users choose their passwords with a personal reference to remember them better. However, it should be noted that further cleaning of the words must take place to remove the # and characters that are typical for Twitter. Furthermore, the German NER also recognizes words as named entities that are none. In the future, information from nodes will be added to the password list, in order to include, for example, Olaf Scholz's wife Britta Ernst as well as other information found about her.

The automated analysis of photos that supports research has been demonstrated with GAD. This capability becomes more helpful when extended with other services such as Google Vision. In the next step, we plan to test the framework on more private individuals as they are typically not aware of the external effects on their posts.

### 5.3 Comparison and Limitations

The basic functioning of collecting and gaining information about identities has been explained. On the technical side, however, there are still some limits and obstacles in comparison to the established tools. The search can only search at locations, where modules with APIs are already written. Further APIs still have to be integrated. In case of APIs are not possible, the search becomes cumbersome. This is also one limitation of existing tools. In comparison, Maltego CE found four unrelated email addresses, whereas SpiderFoot Open Source said the person exists. The latter result did not change when including a Google API.

As programs return different data types, a workaround for Python-based programs was created.

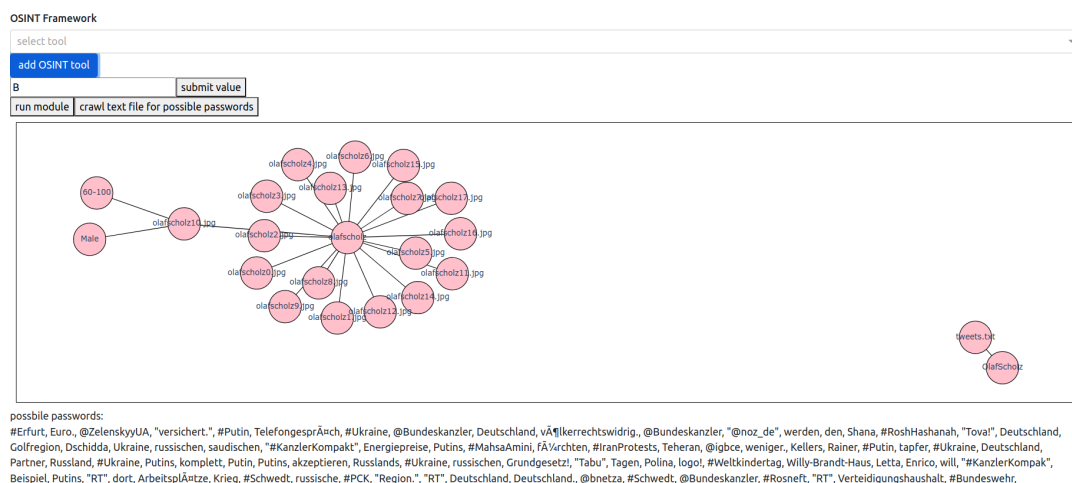


Figure 2: Research on Olaf Scholz.

From the `subprocess.run()` Python method used, each output of the executed program is recorded as a string. For outputs in list format, an interpreter was written that turns the string back into a list. Further interpreters for other Python typical data formats are planned. However, here lies a possible weakness that affects user-friendliness. If developers were to use proprietary data types for the output, users would have to write their interpreter or information extraction process. A first idea would be to create a menu, as envisaged for the integration of APIs, in which users insert a tool output and mark which information is relevant. The framework should then recognize this, save it, and build an interpreter.

## 6 CONCLUSION AND OUTLOOK

OSINT unearths information just waiting to be discovered - either by an individual/organization or an attacker. In order to master the flood of information, classification is necessary for a systematic search. This paper provides a systematic classification for identity and technical data, which is based on a literature review and available tools. In addition, all-in-one search tools and helpers are described. The classification is applied by the open-source OSINT framework approach, which covers the phases of data collection, data analysis, and knowledge extraction. The OSINT framework approach is then discussed based on a targeted search on Olaf Scholz. It shows that open-source tools are possible, though require additional work to produce similar or better results than established tools with a focus on networks.

In order to provide a comprehensive tool for identity protection, further sources will be added in future work. In addition, we plan a user study on the us-

ability and success rate, comparing the results with other open-source and commercial tools. At the same time, countermeasures to hide one's information will be outlined. This OSINT framework will then be extended for organizational purposes.

## REFERENCES

- Akhgar, B., Bayerl, P. S., and Sampson, F. (2017). *Open source intelligence investigation: From strategy to implementation*. Springer.
- Al-Moslmi, T., Gallofré Ocaña, M., L. Opdahl, A., and Veres, C. (2020). Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, 8:32862–32881.
- Azevedo, R., Medeiros, I., and Bessani, A. (2019). PURE: Generating Quality Threat Intelligence by Clustering and Correlating OSINT. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 483–490.
- Bermudez Villalva, D. A., Onalapo, J., Stringhini, G., and Musolesi, M. (2018). Under and over the surface: a comparison of the use of leaked account credentials in the Dark and Surface Web. *Crime Science*, 7(1):17.
- Bielska, A., Kurz, N. R., Baumgartner, Y., and Benetis, V. (2020). Open Source Intelligence Tools and Resources Handbook 2020. [https://i-intelligence.eu/uploads/public-documents/OSINT\\_Handbook\\_2020.pdf](https://i-intelligence.eu/uploads/public-documents/OSINT_Handbook_2020.pdf). Accessed 10-10-2022.
- Butler, B., Wardman, B., and Pratt, N. (2016). REAPER: an automated, scalable solution for mass credential harvesting and OSINT. In *2016 IEEE APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10.
- Cyber Detective (2022). OSINT tools collection. <https://github.com/cipher387/osint.stuff.tool.collection>. Accessed 10-10-2022.

- Daskevics, A. and Nikiforova, A. (2021). ShoBeVODSDT: Shodan and Binary Edge based vulnerable open data sources detection tool or what Internet of Things Search Engines know about you. In *Proceedings of the 2nd IEEE International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 38–45.
- Fang, Y., Guo, Y., Huang, C., and Liu, L. (2019). Analyzing and Identifying Data Breaches in Underground Forums. *IEEE Access*, 7:48770–48777.
- Gibson, H. (2016). Acquisition and Preparation of Data for OSINT Investigations. In Akhgar, B., Bayerl, P. S., and Sampson, F., editors, *Open Source Intelligence Investigation: From Strategy to Implementation*, pages 69–93. Springer International Publishing, Cham.
- Google (2022a). Natural Language API. <https://cloud.google.com/natural-language>. Accessed 10-10-2022.
- Google (2022b). Vision AI. <https://cloud.google.com/vision>. Accessed 10-10-2022.
- Hickey, M. and Arcuri, J. (2020). *Open Source Intelligence Gathering*, pages 55–86. Wiley Data and Cybersecurity.
- Hunt, T. (2022). Have I Been Pwned: Check if your email has been compromised in a data breach. <https://haveibeenpwned.com>. Accessed 10-10-2022.
- IBM Developer (2022). Watson APIs - Resources and Tools. Accessed 10-10-2022.
- Kanta, A., Coisel, I., and Scanlon, M. (2020). Smarter Password Guessing Techniques Leveraging Contextual Information and OSINT. In *2020 International IEEE Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pages 1–2.
- Kanta, A., Coisel, I., and Scanlon, M. (2022). A Novel Dictionary Generation Methodology for Contextual-Based Password Cracking. *IEEE Access*, 10:59178–59188.
- LastPass (2021). The 2021 Password Security Report. <https://www.lastpass.com/de/resources/ebook/psychology-of-passwords-2021>. Accessed 10-10-2022.
- Lee, M., Liang, P., and Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*.
- Martins, C. and Medeiros, I. (2022). Generating Quality Threat Intelligence Leveraging OSINT and a Cyber Threat Unified Taxonomy. *ACM Trans. Priv. Secur.*, 25(3).
- Microsoft (2022a). Face API. <https://azure.microsoft.com/en-us/products/cognitive-services/face/>. Accessed 10-10-2022.
- Microsoft (2022b). Text analytics. <https://azure.microsoft.com/en-us/products/cognitive-services/text-analytics>. Accessed 10-10-2022.
- MISP Project (2022). MISP taxonomies and classification as machine tags. [https://www.misp-project.org/taxonomies.html#\\_osint](https://www.misp-project.org/taxonomies.html#_osint). Accessed 10-10-2022.
- Nordine, J. (2022). OSINT Framework. <https://osintframework.com>. Accessed 10-10-2022.
- Notz, M., Grambau, J., and Hitzges, A. (2019). Evaluation of Sentiment Databases: A Comparison of Sentiment Databases through Social Listening Statements and Azure Machine Learning Studio. In *Proceedings of the 3rd ACM International Conference on E-Business and Internet (ICEBI)*, page 8–12.
- Noubours, S., Pritzkau, A., and Schade, U. (2013). Nlp as an essential ingredient of effective osint frameworks. In *Proceedings of the IEEE Military Communications and Information Systems Conference (MILCIS)*, pages 1–7.
- Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., and Martínez Pérez, G. (2020). The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends. *IEEE Access*, 8:10282–10304.
- Peng, P., Xu, C., Quinn, L., Hu, H., Viswanath, B., and Wang, G. (2019). What Happens After You Leak Your Password: Understanding Credential Sharing on Phishing Sites. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security (Asia CCS)*, page 181–192.
- Sharad Sonawane, H., Deshmukh, S., Joy, V., and Hadsul, D. (2022). Torsion: Web Reconnaissance using Open Source Intelligence. In *Proceedings of the 2nd IEEE International Conference on Intelligent Technologies (CONIT)*, pages 1–4.
- Sherlock Project Team (2022). Sherlock Project. <https://sherlock-project.github.io>. Accessed 10-10-2022.
- V., A. A., A. K., B., R., M., Subbaraj, K., and Kumar Mohan, A. (2020). PeopleXploit : A hybrid tool to collect public data. In *Proceedings of the 4th IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 1–6.
- Verizon (2022). Data Breach Investigations Report 2022. <https://www.verizon.com/business/resources/reports/2022/dbir/2022-data-breach-investigations-report-dbir.pdf>. Accessed 10-10-2022.
- Wilkening, F., Stiemert, L., Schopp, M., Pöhn, D., and Hommel, W. (2022). Investigating Leaked Sensitive Information in Version Control Systems with the Kraulhorizon Framework. In *Sicherheit in vernetzten Systemen: 29. DFN-Konferenz*, pages C1–C21. Books on Demand.
- Yang, H.-C. and Lee, C.-H. (2012). Mining open source text documents for intelligence gathering. In *Proceedings of the International IEEE Symposium on Information Technologies in Medicine and Education (ITiME)*, volume 2, pages 969–973.
- Zhang, Y., Monrose, F., and Reiter, M. K. (2010). The Security of Modern Password Expiration: An Algorithmic Framework and Empirical Analysis. In *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS)*, pages 176–186.