A primer on Open Source Intelligence (OSINT) leveraging existing tools

Lorena González-Manzano
Associate Professor, PhD in Computer Science and Technology
Computer Security Lab (COSEC)
Universidad Carlos III de Madrid, Leganés (España)
lgmanzan@inf.uc3m.es

Abstract*. The acquisition of intelligence from public open sources, known as Open Source INTelligence (OSINT), is a common practice in many investigations. Data analysis skills are a priority, as well as the knowledge of OSINT tools to relieve the burden of data gathering and analysis. Though data analysis has received a lot of attention, OSINT tools are much less explored. In this regard, and with the intent to be especially useful for newcomers in this discipline, this paper presents a classification of OSINT tools from the professional and academic point of view,. Afterwards, a set of open challenges are highlighted.*

Keywords: open source, intelligence, OSINT, tools.

1. Introduction

"Information is power" -- this is the main motivation that supports the use of Open Source INTelligence (OSINT) techniques. This process requires the ability to establish relationships among collected information from public sources, select the most appropriate ones, combine and analyse them to reach a conclusion. All of this is nourished by experience and creativity [1]. At the light of the large amount of available information OSINT tools support and alleviate the management process.

OSINT was initially used in military environments [2] even before the term was coined. Techniques consisted of collecting information from assorted media different from the web. For instance, in the Second World War, gathering intelligence information simplified the establishment of links between the railway efficiency in France and the price of oranges in Paris, which allowed the successful identification of a railway night raid [1]. Later, in 2001, NATO defined the term OSINT and, in 2005, it was used in printing press, television [3] and continued in the radio, as in times of war [4]. Similarly, with the emergence of social networks, such as Facebook in 2004 or Twitter in 2006, the possibilities offered by OSINT increased, leading to the term SOCial Media INTelligence (SOCMINT) as a sub-branch of OSINT. Social networks have a lot of sensitive information which is useful to identify attackers and victims. As we move forward, the amount of information that the Internet provides facilitates the exchange and increase of knowledge, though without neglecting the complexity of its management.

In an OSINT investigation the process can be divided in five steps, highlighting information gathering, organization and distinction [1]. In the first step all necessary sources are identified to, afterwards, gather information from chosen sources. Third, information is pre-processed and the relevant one is selected for its later analysis. Considering that much information can be collected, synthetization, distinction and analysis are essential processes. Finally, results should be reported and though it may seem simple, this is a challenging task. Reports should be written according to the receiver profile, for instance, removing or including technical aspects as required.

The use of OSINT involves a pair of essential features. On the one hand, analytical skills are a priority to process information, and, on the other hand, the knowledge of tools to simplify information gathering and processing is demanding. The satisfaction of these features requires experience and knowledge on big data analysis in the first case and on OSINT tools in the second one. While manyresearch works and professional tools have been developed towards data analytics [5], OSINT tools are much less explored. Note that, herein, tools refer to any kind of software, in the form of web application, desktop program, mobile app, etc., which could be used in the OSINT process. In this regard, the contributions of this paper are the following:

- Present an overview of different types of existing OSINT tools, distinguishing between those commonly used by the OSINT community, referred as professionals, and those developed by academics to address a research problem. In fact, an ENISA report, published in 2017, pointed out the need to advance in the development of automation in cyberintelligence [6], which highlights the relevance of research in this field.
- Identify open challenges in relation to OSINT tools. The type of information, the assorted networks and so on, make difficult the development of tools to address all needs,

thus several challenges are open to discussion.

As a result, without providing a holistic description of available OSINT tools, this proposal tries to present a general classification to be especially useful for newcomers. Moreover, to simplify the knowledge of where and how such tools are applied, the uses of OSINT are introduced in first place.

This paper is structured as follows: Section 2 presents OSINT uses, also helpful to identify involved parties; Section 3 classifies OSINT tools; Section 4 introduces open challenges; Section 5 describes related work; and Section 6 outlines conclusions.

2. OSINT uses

Either governments or security forces may apply OSINT for different purposes, such as fighting against terrorism, the study of political influence and impact, or the protection of citizens. On the opposite side, where criminals and terrorists are working on, OSINT is also of interest to gather victims' data (related to their real life or their systems) and craft attacks to steal information afterwards, or to use social networks for recruitment purposes. OSINT is also used in companies, being a key part in the candidates' selection process. The search of candidates' information in social networks is a common practice. For instance, companies can know, without asking, if candidates like parties, drink alcohol or have a family, based on the data that is shared with no restrictions by the candidate at stake.

Moreover, OSINT takes part in the first step of emerging and dangerous threats called Advance Persistent Threats (APTs) [7]. They are state-sponsored threats which are complex and hard to be detected. OSINT is applied in the *reconnaissance* phase, the first one applied by APTs, to gather victims' information and later prepare a sophisticated attack.

However, OSINT can be applied for many other uses and there are several research proposals in this regard. [8] shows how cybersecurity risk public information can be used to predict current attack trends. By contrast, [9] applies OSINT tools to do a risk analysis and prevent cyberattacks. Information from, for instance, social networks like LinkedIn is used to set an attack risk, which gets low or high depending on factors such as whether the employee is married or has children. On the other hand, [10] is linked to CARPER European Project, where an OSINT solution has been developed to prevent organized cybercrime, as well as it provides semantic and operational interoperability getting the standardization of the information processing. Tracking terrorist from open sources, [11] proposes the use of data mining and link analysis techniques. A case study about searching information through Twitter of a Somali terrorist Group is described. From Twitter some aliases are collected to check connections in other social networks, thus getting to know some related tweets, followers and even tweets location. The protection of critical infrastructures is also of prime importance and, in this context, [12] shows the potential of using public sources to attack energy sector infrastructures. Information about energy systems and contingency plans is gathered to learn how to avoid energy cuts or power lockouts and finally, to generate attack vectors from known vulnerabilities. However, not just companies or infrastructure managers, but regular users can also be attacked by the use of OSINT-- [13] describes the possibility of being attacked through the reception of customized emails, specially crafted for a given victim who has been carefully investigated through public sources.

3. Tools

Tools are divided between professional and academic. The former set refers to those commonly used in the OSINT process by companies or security experts; while the latter are tools developed by academics to address a research goal.

3.1 Professional

Assorted tools are used in an OSINT process, in this paper they are classified as search engines, social network analysis, users' data retrieval, data extraction from files, web pages information retrieval and multi-purpose, see Fig. 1.

Search engines are tools which, given an input, look for requested items. The input is distinguished in two different classes, namely, text and images. There are assorted search engines which allow introducing text, regardless of the type, as input. In this way, apart from common browsers like Google or Bing which provide general information in the form of links to websites, there are others which allow the identification of network devices to be later exploited, e.g. Shodan[1], the search of places, e.g. Google Maps[2], plane routes, e.g. FlightConnections[3], or even leaked information, e.g. IntelX[4]. On the other hand, images can be also used as input in search engines, either to look for similar images or for websites in which such images appear.

[1] www.shodan.io/ , last access Oct. 2020

[2] www.google.es/maps/preview , last access Oct. 2020

[3] www.flightconnections.com/es , last access Oct. 2020

[4] intelx.io/ , last access Oct. 2020

Google Images[5] or Yandex[6] are common image search engines, being this latter especially appropriate for facial recognition.

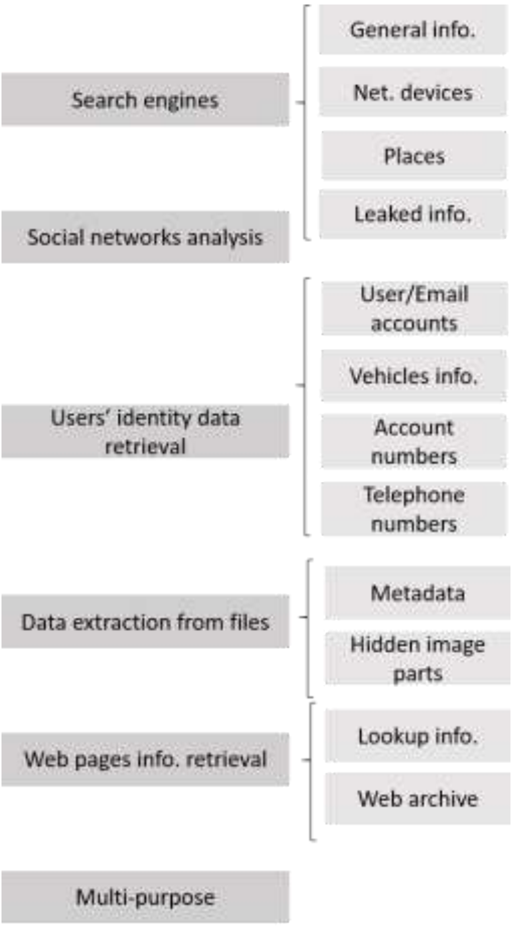

Figure 1. Professional OSINT tools classification

Social network analyses are tools focused on searching data in social networks. Considering the amount of personal data these applications store, their analysis is of vital concern in many situations. Several tools are linked to a social network, for instance, Tinfoleak [7] automates the process of extracting information from Twitter. Though there are other tools like DownAlbum [8], which allow the download of pictures from Facebook, Instagram and more.

Users' identity data retrieval tools look for gathering any type of information directly related to users' identification. There are assorted types but, amount the most common, the identification of usernames and email accounts are an interesting target. Tools like Checkusernames[9] shows if a given username is used in more than 160 social networks, and others like hunter[10] help us to identify company email addresses. Vehicle information can be of interest at some point, for instance, CARFAX[11] is used in Europe to collect vehicles information given a car registration number. Additionally, bank data can be also retrieved, for instance, iban.es[12] verifies a bank account number and provides information on bank details; and Bindb[13] is a bank identification number database which helps in fighting against fraud by giving information about bank details of credit card numbers. Telephone numbers could be also of prime interest in many investigations. Tools like sync.me[14], which can be involved in those referred as "who called me", shows information about a given telephone number.

Data extraction from files tools gather information from files, which despite not being visible, it is part of them. The analysis of metadata regardless of the file type, i.e., using ExifTool[15], can provide remarkable investigation leads. For example, many smartphones' cameras store location coordinates in pictures in the form of metadata. Also considering images, the identification of hidden or missed parts could be also of interest. In the case of high-quality images, any image processing tool can be used to enlarge one part of the image and look for something in detail. Similarly, some pictures, like many posted on Twitter, are not complete, but just a part of them is visualized. In these cases, images can be downloaded, e.g. using browser tools, to study their complete version.

Web pages information retrieval tools can be namely divided between those that provide information about a domain or web page and those that store old internet resources. The first type is commonly referred as Whois service. For instance, in Domaintools[16] either the IP or the domain of a website is introduced to collect information like the website location, the type of server or the hosting history. The second type corresponds to deprecated internet archives, that is web pages and files which were available at some point in time, but that they are not

[5] images.google.com/ , last access Oct. 2020

[6] yandex.com/ , last access Oct. 2020

[7] www.isecauditors.com/herramientas-tinfoleak, last access Oct. 2020

[8] chrome.google.com/webstore/detail/downalbum/cgjnhhjpfcdhbhlcmmjppicjmgfkppok , last access Oct. 2020

[9] checkusernames.com/ , last access Oct. 2020

[10] hunter.io/ , last access Oct. 2020

[11] www.carfax.es/ , last access Oct. 2020

[12] iban.es , last access Oct. 2020

[13] www.bindb.com/ , last access Oct. 2020

[14] sync.me/ , last access Oct. 2020

[15] exiftool.org/ , last access Oct. 2020

[16] www.domaintools.com/ , last access Oct. 2020

available anymore. InternetArchive[17] is the most well-known tool in this regard.

Multi-purpose tools allow gathering data from multiple sources and, in some cases, they also facilitate information analysis. Maltego[18] is an example of this kind of tools. It allows accessing multiple data sources and their visualization in a graphical way using a link graph. Such graphical representation is particularly useful to establish connections between involved parties of an OSINT investigation.

However, all gathered information should be careful checked because, though there are tools and much information can be collected, there are also tools to hijack or deceive these ones. For instance, Burner[19] generates a virtual telephone number and a chosen location can be also set, introducing confusion in the OSINT process. Then, the use of several tools of the same category is recommendable, they may work differently and provide different results, but all of them useful and necessary when working with OSINT.

3.2 Academic

Researchers work towards problems not already addressed or fully solved. The same happens with OSINT tools, there are many, but they do not cover all required needs. Fig. 2 presents the classification of research tools, which is significantly smaller than the professional one. The main difference is that these research tools are generally developed not to just gather a particular type of information but to gather information with a particular purpose in mind, such as the discovery of criminal activities.

Data in the darknet tools refer to those in which the *darknet*, that is the network involving web sites not indexed by common browsers, is used in the OSINT process. P. S. Narayanan et al. presented TorBot [11] to search services in the darkweb. They look for the identification of illegal activities and the visualization of links (e.g. family relationships) between collected data.

Cybersecurity-related data tools are developed to address a cybersecurity need. Due to the lack of tools related to the security development lifecycle (SDL), [14] develops "Threat Miner for SDL" to automate gathering information and delivering product specific threat indicators to inform the SDL while monitoring the disclosure of vulnerabilities along the whole development process. On the other hand, A. Magalhães et al. developed TExtractor [15], a tool to search keywords in videos and audios for helping in attackers monitoring, for instance, analysing content published in forums, social networks, or other channels. By contrast, B. Butler et al. introduce a method, called REAPER [16], to inform about the amount of intelligence information that can be collected looking at criminal activities related to credential dumping, that is, the collection of usernames and passwords. Finally, in [17] the Detection Maturity Level (DML) [18] is modified to identify the maturity of cyberattacks detection. DML, developed to be used by organizations, is improved including OSINT to support attackers' identification.

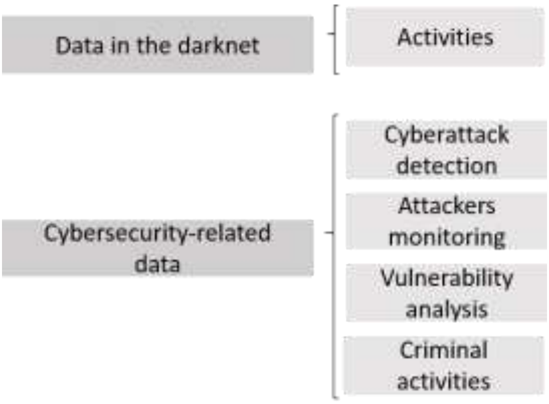

Figure 2. Research OSINT tools classification

4. Open challenges

Despite the great amount of OSINT tools, there are open challenges still to be addressed:

- **Global tools**: in most cases tools are developed to be used in a specific language or country, which is an extremely high limitation. For instance, most common passwords or the format of telephone numbers change between countries and features like these should be considered in the tools development process to increase the scope of their use.
- **Demand of open source tools**: many tools are commercial or just part of the services is available for free. Indeed, most powerful tools are completely commercial, e.g. Maltego[18]. The development of open source tools is especially useful to facilitate customization and, from the security point of view, to allow auditing the tool and ensure that information is securely managed, e.g. just collected the minimum possible.
- **Authenticity and reliability tools**: although a lot of information can be collected from assorted sources, their analysis will not lead

[17] archive.org/ , last access Oct. 2020

[18] www.maltego.com/ , last access Oct. 2020

[19] www.burnerapp.com/virtualnumber/googlevoice , last access Oct. 2020

to conclusive results unless ensuring the authenticity of the information and the reliability of the sources. There are many ways attackers hide themselves, changing information when being chased. This is an intrinsic problem of working with open sources, but more research could be carried towards the development of tools which help in the process of verifying the authenticity of information together with the reliability of sources.

- **Darknet tools**: many illegal activities are hidden in the darknet, as several proposals highlight [19]. Nonetheless, such activities are quite hard to follow and, despite the existence of some OSINT tool to work in this network [20], many more would be valuable to support the search of such activities.
- **Need of automation**: all tools should be as automatic as possible to release the burden of gathering and processing information. The development of multi-purpose tools is appropriate in this regard because they can be pointed out as all-in-one tools. Moreover, to cope with the vast amount of potential information, tools must be designed with scalability in mind. Therefore, cloud-based or other distributed approaches may be considered.
- **Demand of professionals together with tools knowledge**: the huge number of sources, tools and thus gathered information, call for specialist in this discipline. In line with [21], OSINT should be included as a new learning area, detailing the amount of skill to be accomplished. Besides, the training program should include assorted OSINT tools to help specialists along the whole OSINT process. Although some training initiatives are being carried out (e.g., ASSETS+ project for cybersecurity in the defense sector [22]), there is a need to develop suitable academic curricula.

5. Related work

OSINT is a well-known type of intelligence which has been used for a long time, and the research community has also work in this discipline. For instance, [1] studies the evolution of OSINT and [3] reviews the application of OSINT with artificial intelligence.

However, just some works have paid attention to the assorted set of OSINT tools. [23] reviews the use of OSINT as a cybercrime investigation framework. The types of cybercrimes are introduced, the most predominant tools and techniques for OSINT collection and storage are summarized and techniques and methods for a cybercrime investigation are also modeled. Focusing on the web and the sensitive information that can be found, [24] presents a classification of web history tools. Some search engines are used in first place to, later, select web history tools and, finally, present a classification composed of 6 tools. By contrast, [25] introduces a very brief methodology to leverage context information and guess passwords. They point out the need of using OSINT tools, without specifying them, to gather information, which is filtered and translated to meaningful contextual data useful for guessing passwords. [26] presents a framework to help in the assessment of OSINT tools to select the best one for each investigation. It introduces the need of using secure, reliable, and legal tools, but without introducing any of them. Indeed, [27] and [17] are the most similar approaches to the one proposed herein, introducing several tools and being [17] especially remarkable for the number of enumerated tools. [27] presents OSINT tools to search information on the web, verify the authenticity of social network data, look for archives and old web pages, and track the flow of users in communities, called fringe, which are outside of the social media industry mainstream; while [17] presents a model to identify attackers based on OSINT and, apart from that, categorize tools in search engines, social networks, email addresses, usernames, real names, locations, IP addresses and domain names techniques.

Table 1 presents a comparison of works studying OSINT tools. Though some work has already introduced an interesting amount of professional tools, this proposal studies professional and academic tools, as well as it includes some types of tools, like those focused on extracting information from files, which have not been directly mentioned so far. Moreover, just [17] points out several open challenges related to OSINT tools, but this work enhances [17] presenting a more concrete identification of them, including some new ones.

6. Conclusions

The use of OSINT is very widespread according to its low risk and cost. It works with public information which can be easily gathered in comparison with other types of intelligence that require field work. The simplicity of access and the low amount of potential legal problems, as most sources can be unlimitedly shared, also encourage its use. However, there are barriers to overcome. Skills for big data management are specially demanding. Additionally, information can be collected from different sources and with assorted formats, thus being necessary their processing to get a common structure. Indeed, the existence of multiple sources highlight the need of working towards sources reliability and information authenticity, being highly recommendable the knowledge of tools to choose the most appropriate ones in each situation. In this regard, this paper has presented an overview of OSINT tools, especially useful for newcomers in this discipline. Both professional and academic tools have been presented. Moreover, a set of open challenges have been finally outlined.

Table 1. Related work

	Profesional	Tools purpose	Academic
[23]	√	Cybercrime investigations.	x
[23]	√	Web history.	x
[25]	√	-	x
[26]	√	-	x
[27]	√	Search engines, verify users' authenticity in social networks, look for archives and fridge communities.	x
[17]	√	Search engines, social networks, email addresses, username, real, location, ip addresses and domain names techniques.	x
This paper	**√**	**Search engines, social networks, personal data, data extraction from files, web sites information.**	**√**

Acknowledgements

This work was supported by the Spanish Ministry of Science, Innovation and Universities grants TIN2016-79095-C2-2-R (SMOG-DEV),PID2019-111429RBC21(ODIO); by Comunidad de Madrid (CAM) grantP2018/TCS4566 (CYNAMON-CM) funded with European FEDER funds; and CAVTIONS-CM-UC3M funded by UC3M and CAM.

Referencias

[1] Glassman, M., & Kang, M. J. (2012). Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT). Computers in Human Behavior, 28(2), 673-682.

[2] Casanovas, P. (2017). Cyber warfare and organised crime. A regulatory model and meta-model for open source intelligence (OSINT). In Ethics and Policies for Cyber Operations (pp. 139-167). Springer, Cham.

[3] Evangelista, J. R. G., Sassi, R. J., Romero, M., & Napolitano, D. (2020). Systematic Literature Review to Investigate the Application of Open Source Intelligence (OSINT) with Artificial Intelligence. Journal of Applied Security Research, 1-25.

[4] Mercado, S. C. (2009). Sailing the Sea of OSINT in the Information Age. Secret Intell Reader, 78.

[5] Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. Journal of Big data, 2(1), 1-32.

[6] ENISA - Threat landscape report 2017. TR, EU Cybersecurity Agency

[7] Chen, P., Desmet, L., & Huygens, C. (2014, September). A study on advanced persistent threats. In IFIP International Conference on Communications and Multimedia Security (pp. 63-72). Springer, Berlin, Heidelberg.

[8] Ang, C. K., & Datta, A. (2020). Open source intelligence gathering and topic modelling on cyber security incidents. Nanyang Technological University.

[9] Hayes, D. R., & Cappa, F. (2018). Open-source intelligence for risk assessment. Business Horizons, 61(5), 689-697.

[10] Aliprandi, C., Irujo, J. A., Cuadros, M., Maier, S., Melero, F., & Raffaelli, M. (2014, June). CAPER: Collaborative information, acquisition, processing, exploitation and reporting for the prevention of organised crime. In International Conference on Human-Computer Interaction (pp. 147-152). Springer, Cham.

[11] Dawson, M., Lieble, M., & Adeboje, A. (2018). Open source intelligence: Performing data mining and link analysis to track terrorist activities. In Information Technology-New Generations (pp. 159-163). Springer, Cham.

[12] Keliris, A., Konstantinou, C., Sazos, M., & Maniatakos, M. (2019). Open source intelligence for energy sector cyberattacks. In Critical Infrastructure Security and Resilience (pp. 261-281). Springer, Cham.

[13] Uehara, K., Mukaiyama, K., Fujita, M., Nishikawa, H., Yamamoto, T., Kawauchi, K., & Nishigaki, M. (2019, March). Basic study on targeted

e-mail attack method using OSINT. In International Conference on Advanced Information Networking and Applications (pp. 1329-1341). Springer, Cham.

[14] Kannavara, R., Vangore, J., Roberts, W., Lindholm, M., & Shrivastav, P. (2019, February). A Threat Intelligence Tool for the Security Development Lifecycle. In Proceedings of the 12th Innovations on Software Engineering Conference (pp. 1-5).

[15] Magalhães, A., & Magalhães, J. P. (2018, June). TExtractor: An OSINT Tool to Extract and Analyse Audio/Video Content. In International Conference on Innovation, Engineering and Entrepreneurship (pp. 3-9). Springer, Cham.

[16] Butler, B., Wardman, B., & Pratt, N. (2016, June). REAPER: an automated, scalable solution for mass credential harvesting and OSINT. In 2016 APWG symposium on electronic crime research (eCrime) (pp. 1-10). IEEE.

[17] Pastor-Galindo, J., Nespoli, P., Mármol, F. G., & Pérez, G. M. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. IEEE Access, 8, 10282-10304.

[18] Stillions, R. (2014). The DML model.

[19] Broséus, J., Rhumorbarbe, D., Mireault, C., Ouellette, V., Crispino, F., & Décary-Hétu, D. (2016). Studying illicit drug trafficking on Darknet markets: structure and organisation from a Canadian perspective. Forensic science international, 264, 7-14.

[20] Narayanan, P. S., Ani, R., & King, A. T. (2020). TorBot: Open Source Intelligence Tool for Dark Web. In Inventive Communication and Computational Technologies (pp. 187-195). Springer, Singapore.

[21] Gruters, P. C., & Gruters, K. T. (2018). Publicly Available Information: Modernizing Defense Open Source Intelligence. Special Operations Journal, 4(1), 97-102.

[22] ASSETS+ European Project, https://assets-plus.eu/, last access Oct. 2020.

[23] Tabatabaei, F., & Wells, D. (2016). OSINT in the Context of Cyber-Security. In Open source intelligence investigation (pp. 213-231). Springer, Cham.

[24] Evangelista, J. R. G., de Oliveira Gatto, D. D., & Sassi, R. J. (2019, July). Classification of web history tools through web analysis. In International Conference on Human-Computer Interaction (pp. 266-276). Springer, Cham

[25] Kanta, A., Coisel, I., & Scanlon, M. (2020, June). Smarter Password Guessing Techniques Leveraging Contextual Information and OSINT. In 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security) (pp. 1-2). IEEE.

[26] Revell, Q., Smith, T., & Stacey, R. (2016). Tools for OSINT-Based Investigations. In Open Source Intelligence Investigation (pp. 153-165). Springer, Cham.

[27] Hayden, M. E. (2019). Guide to Open Source Intelligence (OSINT).