VARSTVOSLOVJE
Journal of Criminal Justice and Security
year 23
no. 4
pp. 425–446

The Use of Tools for Obtaining Data From Publicly Accessible Sources for the Purpose of Competitive Intelligence in Enterprises

Žiga Primc

Purpose:

Article seeks to determine the state of affairs of research in using tools to obtain data from public sources for competitive intelligence activities. The author wishes to offer an overview of what already exists and shed light on areas that merit further research efforts.

Design/Methods/Approach:

A systematic literature review was made using the PRISMA method in the Scopus and IEEE Xplore databases. The review included 23 articles published between 2011 and 2021.

Findings:

Over the past decade, the practice of obtaining data from publicly accessible sources has developed extremely rapidly, and this can be attributed mainly to the rapid advances in technology, the digitalisation the operations of organisations, and to the immeasurable quantities of data that appear every day on the internet. New fields are emerging in obtaining data from publicly accessible sources, and more effective tools are being developed to automate processes. Existing studies highlight the importance of competitive intelligence activities by means of obtaining data from publicly accessible sources for organisations. At the same time, they identify a lack of trained personnel and adequate software systems for optimal business use.

Practical Implications:

The findings of this article offer an insight into research of competitive intelligence activities and highlights the research gaps and formulates starting points for future research.

Originality/Value:

The article is the first systematic literature review in obtaining data from publicly accessible sources for the purpose of competitive intelligence activities to support better decision-making in business environments.

Keywords: Open Source Intelligence, competitive intelligence, systematic literature review, tools, public sources.

UDC: 004.6

Uporaba orodij za pridobivanje podatkov iz javno dostopnih virov v okviru konkurenčne obveščevalne dejavnosti v podjetjih

Namen prispevka:

Namen prispevka je s pomočjo sistematičnega pregleda literature ugotoviti uporabnost orodij za pridobivanje podatkov iz javnih virov v okviru konkurenčne obveščevalne dejavnosti v podjetjih. Na podlagi analize obstoječega stanja želimo izpostaviti vsebine za nadaljnje raziskovanje.

Metode:

Narejen je bil sistematični pregled literature po metodi PRISMA v bazah podatkov Scopus in IEEE Xplore. V končni pregled je bilo vključenih 23 prispevkov, ki so bili objavljeni med letoma 2011 in 2021.

Ugotovitve:

Področje pridobivanja podatkov iz javno dostopnih virov se v zadnjem desetletju razvija izjemno hitro, kar gre pripisat predvsem hitremu tehnološkemu razvoju, digitalizaciji poslovanja organizacij ter neizmerljivim količinam podatkov, ki se vsakodnevno znajdejo na internetu. Odkrivajo se nova področja uporabe pridobivanja podatkov iz javno dostopnih virov, razvijajo se učinkovitejša orodja za avtomatizacijo procesov. Obstoječe študije izpostavljajo pomembnost konkurenčne obveščevalne dejavnosti z uporabo pridobivanja podatkov iz javno dostopnih virov za organizacije, hkrati pa ugotavljajo pomanjkanje usposobljenega kadra in programskih rešitev za optimalno poslovno uporabo.

Uporabnost raziskave:

Ugotovitve prispevka nudijo vpogled v raziskovalno dejavnost na področju konkurenčne obveščevalne dejavnosti ter lahko služijo kot osnova za raziskovanje tega področja. V prispevku so izpostavljene raziskovalne vrzeli in oblikovana izhodišča za prihodnje raziskave.

Izvirnost/pomembnost prispevka:

Gre za prvi sistematični pregled literature na področju uporabe pridobivanja podatkov iz javno dostopnih virov za namen konkurenčne obveščevalne dejavnosti kot podpore za boljše odločanje v poslovnih okoljih.

Ključne besede: pridobivanje podatkov iz javno dostopnih virov, konkurenčna obveščevalna dejavnost, sistematični pregled literature, orodja, javni viri

UDK: 004.6

1 INTRODUCTION

Obtaining data from publicly accessible sources for competitive intelligence activities is a field in which not much has yet been written. Despite this, the field has been gaining attention in recent years. When we speak of obtaining data from publicly accessible sources, we usually encounter the better-known term Open-Source Intelligence or OSINT. This is an expression that dates back more than 70 years, and in comparison, with its beginnings, up to the present day – especially in terms of the method of obtaining information due to technological advances – a lot has changed. Credit for this can be claimed principally by the significant advances in information technology and the digitalisation of everyday life and business. In the review of works dealing with obtaining data from publicly accessible sources, we can quickly see businesses, for the most part, focus on cybersecurity and preventing threats. Open sources of data are excellent for obtaining the majority of information that the state, armed forces or a company might need. Competitive intelligence is the process and forward-looking practices used in producing knowledge about the competitive environment to improve organisational performance (Madureira et al., 2021) and involves systematic collection and analysis of information from multiple sources. It can be described as the action of defining, gathering, analysing, and distributing intelligence about products, customers, competitors, and any aspect of the environment needed to support executives and managers in strategic decision making for an organisation. When we speak of competitive intelligence, we can see that much has been researched and written in terms of preventive measures (cybersecurity, other threats to organisations). For instance, online social networks are used to obtain information about the competition, and they can also be used to identify threats and for taking timely action (Ansari et al., 2013; Al-khateeb & Agarwal, 2020; Yaboah-Ofori & Brimicome, 2017) as well as for detecting potential outbreaks of disease (Bernard et al., 2018). More recent fields of interest include artificial intelligence (AI) in connection with obtaining data from publicly accessible sources (Gonçalves Evangelista et al., 2020), which in the future could also be used in competitive intelligence or for identifying threats to an organisation, based on the identification of specific topics in texts that are freely available (Li et al., 2020).

This article focuses on a review of those sources that relate exclusively to competitive intelligence activities in connection with obtaining data from publicly accessible sources.

2 METHOD

The systematic review of the literature in the field of using tools for obtaining information from publicly accessible sources (open-source intelligence - OSINT) for the purpose of competitive intelligence was conducted in the international bibliographical databases Scopus and IEEE Xplore. The search string used to find articles contained various key words: ((OSINT OR »public sources« OR »corporate intelligence« OR »private *intelligence« OR »competitive intelligence« OR »threat intelligence« OR »cyber intelligence«)). We searched for a combination of selected keywords in titles, summaries and keywords of articles. The review

of the literature was performed on 12 August 2021. Our inclusion criteria covered articles published in 2011 or later and articles written in Slovenian or English. As the field of tools for the use of OSINT in the context of competitive intelligence is developing rapidly technologically, it made sense to include only contributions that are not more than a decade old in the analysis. The use of the specific databases stems from the fact that IEEE Xplore is one of the most important databases linked to information technology, and Scopus is an exceptionally broad and abundant collection of articles. The reason for the absence of the Web of Science database, which would otherwise have been used in our article, is that at the time of searching for material and writing the article that database was experiencing technical difficulties, which prevented us from exporting a list of articles for analysis. We then excluded from the array of articles duplicated articles, those that were not open-source and those that do not relate to obtaining data from publicly accessible sources and competitive intelligence. In reviewing the full text of articles, in the final stage, we excluded irrelevant research and those articles that do not include the basic search terms. The inclusion and exclusion criteria are presented in Table 1.

Table 1: The inclusion and exclusion criteria for systematic literature review

Inclusion criteria	Exclusion criteria
Articles published in scientific journals or conference papers	Books, chapters in books, early accesses, discussions, introductions, reviews.
Articles published between 2011 and 2021	Articles published before 2011
Articles in Slovenian and English	Articles in other languages
Free access to the article	Inaccessibility of the full article
Thematic relevance of the article	Articles that do not relate to the use of public sources, obtaining data from publicly accessible sources and competitive intelligence.

The process of collecting and identifying appropriate articles is presented through a PRISMA diagram in Figure 1.

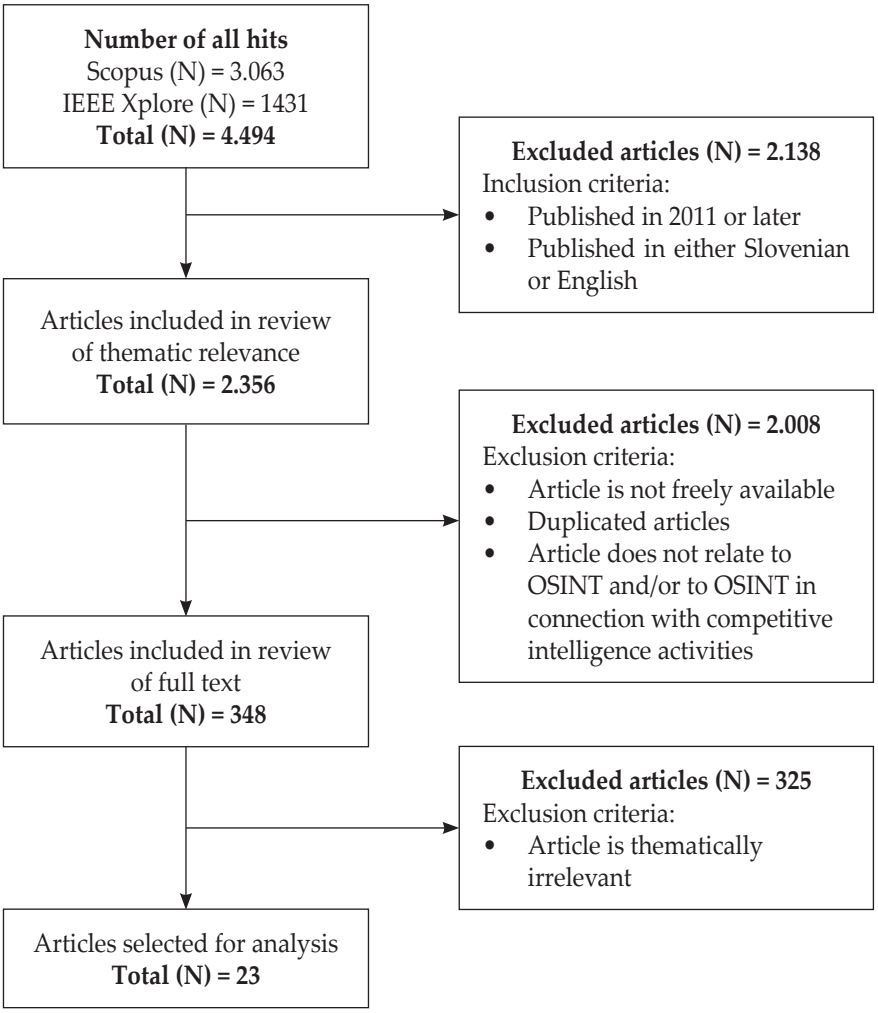

Figure 1: The flow of searching and selecting literature

Based on the criteria presented in Table 1, we can conclude that only a small number of articles are useful for in-depth analysis. Although we initially detected many contributions in both databases, we excluded most of them because they do not analyse the use of the OSINT method in relation to competitive intelligence.

3 RESULTS

The first search for articles with the described search string yielded 4,495 results. Taking into account the inclusion and exclusion criteria, and after excluding duplicates, 23 articles remained for the final analysis. A description of the articles included in the systematic review of the literature is shown in Table 2.

Author and title of article	Type of article and predominant method	Purpose of the article	Key findings of the article
[Anand et al., 2020] PeopleXploit- A hybrid tool to collect public data	review article; presentation of tools	Presents an approach that illustrates the OSINT concept as an important application in profiling individuals.	The PeopleXploit approach makes use of a hybrid tool that aids in the public gathering of available information that is reliable and important for the data searcher.
[Benes, 2013] OSINT, New technologies, education: expanding opportunities and threats. a new paradigm	review article	Analyses how rapidly advanced technology is changing information and consequently the obtaining of data from publicly accessible sources.	The development of new technologies and education in this area are essential for an adequate understanding and use of various OSINT tools. These tools represent the future of obtaining digital intelligence data.
[Best, 2011] Challenges in open source intelligence	review article	Identification of the future challenges regarding tools for obtaining data from publicly accessible sources	Nowadays access itself to gather information using OSINT no longer represents a challenge, it is the analysis of the obtained data that is problematic.
[Buccafurri et al., 2020] Implementing multiple-social-network meta-APIs to support OSINT programming	review article; presentation of model	Points out that despite the availability of tools and platforms for OSINT, gaps exist in terms of software engineering.	Social networks today represent an important source of information and are frequently a basis for OSINT research. In this article the authors present the approach of user-centred social networks programming, whereby pieces of information can be merged and linked, and for OSINT research this is consequently more systematic.

[Černya et al. 2019] Using open data and Google search data for competitive intelligence analysis	review article; presentation of model	Analysis of the competitive environment using OSINT in the health sector.	The analysis of the competitive environment is set out in a reliable time frame, and obtains all the necessary competitive data important for determining the competitiveness of a company.
[Chae et al., 2019] A system approach for evaluating current and emerging army open-source intelligence tools	review article; presentation of model	An assessment of the current situation and examination of emerging OSINT tools and technology.	In the military OSINT is a tool that offers useful information and is used principally for obtaining competitive data and potential threats. At the same time OSINT improves both privacy and protection against threats. Through systemic analysis the authors concluded that certain OSINT tools do not fully meet the requirements of the military, since new tools have been developed on the market that could be more effective.
[Chi & Chen, 2013] Collaborative competitive intelligence: A knowledge base system approach	review article; presentation of model	A definition of the development and presentation of a decision-making platform for boosting the competitive intelligence activities of a company.	The knowledge base system has achieved the aim of targeted competitive intelligence activities and has proven to be successful and reliable. Companies can use these data in formulating strategic objectives.
[Di et al., 2014] Research on enterprise competitive intelligence development and strategies in the big data era	original science article; quantitative research	Presents the development and importance of competitive intelligence activities in the age of digitalisation.	The development of competitive intelligence activities has recently been reaching new heights, since enterprises and organisations are aware of their importance. The digitalisation of data enables the development of various OSINT tools that can comprehensively gather all the necessary information for formulating a competitive strategy based on data about competing enterprises in a completely legal way.

[Dorton et al., 2019] A theoretical model for assessing information validity from multiple observers	original science article; qualitative research and presentation of model	Presentation of a model/approach for assessing the validity of information obtained from HUMINT and OSINT.	In the authors' assessment the approach is appropriate and good, but gaps remain particularly in terms of programming, and these need to be addressed. The focus should be on obtaining information that is not just relevant but also valid.
[Goujon, 2011] Text mining for opinion target detection	review article; presentation of model	Presentation of a model for automated gathering of opinions on a specific topic.	Using text mining it is possible to obtain information on opinions linked to a given topic, for the purpose of competitive intelligence activities. Going forward, a more comprehensive review of sources is needed to confirm the relevance of these opinions for the end user.
[Hribar et al., 2014] OSINT: A "Grey Zone"?	review article	Presentation of elements of intelligence activities and the use of *fuzzy logic* in OSINT.	OSINT is an excellent way of gathering intelligence data that does not violate human rights and is also lawful. OSINT is also a formal tool and means of analytical support in the work of intelligence services –this can also at times be questionable and creates a grey zone in its use.
[Hu & Zhu, 2013] Competitive intelligence acquisition from websites	review article; case study	Analysis of the area of obtaining information from company websites.	Use of competitive intelligence activities is an important method of obtaining information about the competition in the information age. Publicly accessible information on websites offers an excellent insight into information about a company, though this needs to be enhanced by an analysis of data from other publicly accessible databases.

[Pais & Ciobanu, 2014] OSINT for B2B platforms	review article; presentation of software	Determines the possibility of including OSINT via the application Cloud.ro in B2B platforms.	The applicability of OSINT in a B2B platform has been shown to be good and very useful, especially in terms of obtaining information about companies, information on company activities, information on employees and information on business partners.
[Pastor-Galindo et al., 2020] The not yet exploited goldmine of OSINT: opportunities, open challenges and future trend	review article;	Analysis and review of the current state of development of OSINT and a comprehensive overview of the paradigm, with emphasis on technology and services.	OSINT is developing rapidly, and innovative applications are being designed for use in various fields. The authors see the greatest potential for its use as lying in proactive use to detect cyber threats and in the public sphere (especially state authorities), where OSINT represents an ideal tool for obtaining diverse information, including about competitors.
[Rai et al., 2021] Using open source intelligence as a tool for reliable web searching	review article; presentation of model	Analysis of OSINT tools and formulation of a proposed possible new development.	OSINT gathers and obtains data from web servers or sources that are freely accessible. OSINT provides a more rapid means of searching for and filtering data.
[Ranjan & Foropon, 2020] Big data analytics in building the competitive intelligence of organisations	original science article; qualitative research	A study of data in a competitive environment and presentation of the context of the development of competitive intelligence in the future.	Monitoring competitors by means of OSINT has already become an everyday practice, since it enables a determination of the competitiveness of companies. For the most part, companies have a highly centralised, informal process of gathering such data, and development must be oriented towards advanced machine learning based on targeted methods.

[Rasekh, 2015] A new competitive intelligence-based strategy for web page search	review article; presentation of model	Presents a model for searching higher-quality websites, based on competitive intelligence.	The proposed system combines the ICA algorithm and the scheme of classification based on links. The aim of the model is to improve search results, but additional work is needed for the model to be adapted to the web.
[Schaurer & Störger, 2013] The evolution of open source intelligence (OSINT)	review article	Presentation of the development of OSINT.	They define OSINT as a collection, processing, analysis and classification of information obtained from sources in a way that is open and legally accessible. In essence, OSINT emerged as an intelligence discipline in the domain of the state, which in recent times has contributed to the integrated process of managing sources also in the private sector.
[Semerkova et al., 2017] Application of information technologies in competitive intelligence	review article	Presents the concept of competitive intelligence, the intelligence cycle and competitive intelligence tools on the internet.	By means of competitive intelligence, an enterprise can create a competitive advantage, so this is an important strategic orientation for the enterprise. An analysis of the weaknesses and vulnerabilities of competitors is therefore essential, and using OSINT tools can aid in this.
[Sharma et al., 2013] Big data – competitive intelligence	review article; presentation of model	Presents a model for tackling the challenges of organisations adapting to a competitive market by means of obtaining competitive data.	Nowadays obtaining data is no longer as complicated as it once was, due to the availability of various OSINT tools. All the more important data can be obtained from publicly accessible sources, but a systematic review of the data remains the challenge.

[Sondrava et al., 2021] Prevention to sensitive information disclosure via OSINT	review article	A description of OSINT tools and approaches for searching sensitive information online.	The best protection against sensitive information is provided by automated and payable OSINT tools, while at the same time many OSINT tools represent potential 'weapons', since they enable individuals to access information and data which they can use for malicious purposes.
[Wang et al., 2011] Construction and operation of cultivation model for enterprise competitive intelligence competence	review article; presentation of model	Design of a cultivation model for competitive intelligence	This represents an additional value of competitive intelligence, and is regarded as an effective means of raising the competitiveness of a company, if that company has the skills to use it. This plays an important part in shaping business processes and strategies, and competitive must be built both upon internal and external factors.
[Yang & Lee, 2012] Mining open source text documents for intelligence gathering	review article; presentation of model	Formulates an approach for OSINT with the aim of automating the gathering of intelligence data	The gathering and analysis of intelligence data plays an important part in the growth of a company. OSINT is the main approach for gathering and analysing intelligence data, but this is hampered by a lack of automation. The authors have developed a tool that fills this gap and is applicable especially in the area of business intelligence and personal management.

In the use of keywords, we had to set out the search string more broadly since there are too few actual works containing the term OSINT in the title for conducting a systematic review of the literature. For this reason, we spent more time on the actual review of all sources, and consequently obtained better results. In terms of content, we divided the articles into four groups, specifically 1) those that generally address actual OSINT, 2) those that present a specific solution, 3) those that present a model for obtaining and analysing OSINT data, and 4) those that present a model for evaluating the data obtained. We did not include the Web of Science database in our selection of literature because, at the time of gathering articles for review, that website was experiencing technical difficulties that prevented the export of results.

3.1 Review of the field of obtaining data from publicly accessible sources

The best-known definition of Open-Source Intelligence (OSINT) is that this is intelligence data "produced from publicly available information that is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement". The term was defined by both the U.S. Director of National Intelligence (DNI) and U.S. Department of Defense (DoD). Schaurer and Störger (2013) expand the concept somewhat into the "collection, processing, analysis, production, classification, and dissemination of information derived from sources and by means openly available to and legally accessible and employable by the public in response to official national security requirements."

The first signs of efforts to gather data from publicly accessible sources appeared with the establishing of the Foreign Broadcast Monitoring Service (FBMS) in 1941, which was originally intended to monitor foreign (printed and spoken) media. Later on, after the Japanese attack on Pearl Harbour the area of obtaining information from public media gained in importance. After the war, in 1947, this was placed under the administration of the U.S. Central Intelligence Agency (CIA), with the title Foreign Broadcast Intelligence Service (FBIS) (Shaurer & Storger, 2013).

Over decades this field was modified and adapted to changes, and this can be seen most prominently in the last decade, during the period of accelerated digitalisation. While it started out monitoring printed media, radio and later television, today the area of OSINT is focused for the most part on digital media. Pastor-Galindo et al. (2020) state that the quantity of data generated by today's extraordinarily connected world is immeasurable. These data are for the most part accessible to the public and available whenever and wherever. The main sources they mention are the mass media, online social networks, blogs, forums, public administration data, commercial data and publications.

Rai et al. (2021) see obtaining data from publicly accessible sources as a process comprising the following steps: 1) identification of the source, 2) obtaining data, 3) integration and processing of data, 4) analysis of data and 5) delivery of the final result to the relevant recipient. The steps involved in the competitive intelligence cycle were set out in a similar way by Semerkova et al. (2017), who state that

the cycle is made up of defining the objectives, gathering data, analysing data, delivering information to the final recipient and use of the information obtained. Pastor-Galindo et al. (2020) described the process more simply, involving three steps of gathering, analysis and acquiring knowledge. In their article the authors for the most part are concerned with cybersecurity and use of data acquisition from publicly accessible sources for this purpose.

Analysis of the literature has shown inter alia that authors focus on two approaches in using data acquisition from publicly accessible sources: offensive and defensive. The first approach can be more easily placed in the framework of competitive intelligence, i.e., obtaining information on one's competitors on the market, while the second, which is much more prominent, serves to obtain information from public sources through open-source data acquisition for a timely response to cyber and other threats, which ultimately affords organisations a competitive advantage. The timely reaction to an unexpected scenario, for instance a cyber-attack, can in certain cases mean life or death for an organisation.

While obtaining publicly accessible data can have major benefits for an organisation, on the other hand this area is still in a very early stage of development, and the question is, can the development of new technologies catch up in terms of effective use of tools and models for gathering public data. Pastor-Galindo et al. (2020) strikingly delineate the advantages and weaknesses of obtaining data from publicly accessible sources. Among the advantages they list the huge quantity of available information, the massive computer power (when we speak of gathering/searching information), the concept of *big data*[1] and machine learning, the complementarity of data, flexibility of use and the wide selection, while they highlight the weaknesses of complexity and data management, non-structured information, erroneous or false information, reliability of the actual sources of information and ethical or legal reservations. This last factor has been a topic of numerous debates for some time. The grey zone of obtaining data from publicly accessible sources and how this issue can be resolved was written about by Hribar, et al. (2014), who are focused on traditional intelligence activities, although here we can clearly draw parallels with competitive intelligence activities. They state that obtaining data from publicly accessible sources in the majority of cases is accepted, since it involves legal form of information gathering, up until a human right is violated. The personal data of individuals can generate a legal conflict in obtaining data from public sources, so there is a need to respect the legislation governing human rights and privacy (Schaurer & Störger, 2013).

Using tools to obtain data from publicly accessible sources, it is easy to obtain information, and anyone can do it if they decide to acquire these tools, but the harder part is the actual analysis of data, which requires a trained and experienced expert, if you want to achieve optimal results (Hribar et al., 2014). Raw data obtained from public sources can only be assembled into a meaningful whole by someone who requires the legal and technical knowledge, analytical skills and powerful tech support for this (Schaurer & Störger, 2013). The challenge today is no longer how to obtain information, but how to whittle down the

1 *In the opinion of Saša Mojsilović, a court expert for big data at IBM, in 2020 some 40 (Dolenc, 2014) or 44 zettabytes (Microsoft, 2019) of data will be generated in the world (one zettabyte is 44 billion gigabytes). Today the average personal computer has one or two terabytes of memory.*

information obtained to what is relevant. For this reason, there is also a need to further develop the tools that will offer appropriate technical support (Best, 2011). Legitimate use of tools for data acquisition from publicly accessible sources will be possible in the long term through the proper education and training of those who will be involved in it, be it in an organisation or for organisations as a service on demand. Benes (2013) notes that what is needed for an outstanding analyst, alongside excellent theoretical and practical training, is a mindset that espouses lifelong learning. The area of data acquisition from publicly accessible sources is one of the most rapidly developing areas, and the development of personnel here will also require outstanding teachers, schools and training programmes, a general awareness in the public about the opportunities that await in this field and the development of technical support that will deal with the problem of the quantity of information.

Competitive intelligence using big data is an enhanced form of using data acquisition from publicly accessible sources. As determined by Ranjan and Foropon (2020), competitive intelligence using big data involves a complex approach that can offer greater value for practical application. Difficulties arise in analysis, in a way that is similar to obtaining data from publicly accessible sources, except that the quantity of data is significantly greater. In their empirical research they have found that for the most part companies do not see any rational need to implement the big data approach in their competitive intelligence procedure, since they are convinced that they know the market and their competitors well. They also highlight the problems in recruiting appropriate personnel, for as they have found, operatives lack the essential advanced analytical abilities to process, interpret and present results to support decision-making.

3.3 Presentation of systems/tools

In the review of articles over the past ten years we have seen that there have been initiatives and actual development of (software) systems for use in competitive intelligence. Pais and Ciobanu (2014) talk of a purpose-built platform resulted from their development and serves for obtaining data on competitor companies, specifically information on the company, information on activities in their domain, employee data and information on the main business partners. As the authors assert, the system itself can be included in existing platforms.

To obtain data from publicly accessible sources or competitive intelligence, online social networks are a rich source of data, but due to their closed nature, they do not enable free browsing of content in one place. Searching for information can consequently be time-consuming. Buccafurri et al. (2020) propose programming a *meta-API*, which will enable user-centred searching through various social networks. The software itself could enable the developers to create a dedicated tool that could search on various social networks for the location of an individual, their friends or contacts, the use of hashtags and so on. Simultaneous searching by user profiles on different social networks offers us a more complete picture of their activities, since in this way we can merge together information on their leisure

time (e.g. Facebook, Instagram and Twitter) and from the business environment (e.g. LinkedIn).

Checking out individuals is also pursued more broadly in practice, outside the online social networks. Here we can speak of information gathering for competitive intelligence purposes, and background-checking an individual before employment. A very important factor in acquiring publicly accessible information is the already mentioned protection of personal data, i.e. encroachment on privacy. It is always essential to have a legal basis for this kind of searching in such cases. *PeopleXploit* is a tool created for the purpose of increasing the relevance score in profiling, according Anand et al. (2020). This is a hybrid tool that aids in collecting publicly available information that is reliable and relevant to the given input, such as name, surname, e-mail address, telephone number, and so forth, in other words, the user's digital footprints. Once the search is complete, the tool offers an analysis, which differs depending on the purpose of the search.

The majority of the information gathered by means of data acquisition from publicly accessible sources is analysed today manually, since the market does not yet offer sufficient applicable tools that could perform this reliably, or rather such tools have not username yet been perfected. Since analysis is time-consuming, Yang and Lee (2012) and their associates developed an approach for automatic data processing based on text mining techniques. The authors assert that the approach can automatically identify important events, whereby it is able to see from different angles. Actual awareness of such events would offer companies benefits in their decision-making. Text mining techniques can be applicable in the areas of national security and e-learning and also in competitive intelligence.

A similar approach to text mining was taken by Goujon (2011), in connection with obtaining public opinions regarding a specific target. His text mining technique is based on linguistic knowledge and linguistic patterns that automatically identify the target opinions being sought, for the purpose of competitive intelligence. The author tested the tool in specific cases. The results have shown potential in the application of the method, but also deficiencies and indications for further work. Searching opinions and terms can be an extremely complex task due to the dynamics of language and the consequences of dialects and multilingualism. The success of the method therefore depends on a large database of terms through which text is mined and opinions found.

Open-source tools and others that are freely available or available against payment can be found very quickly on the internet. Some tools are extremely powerful, while others are very niche-oriented. In their article, Pastor-Galindo et al. (2020) analyse well the majority of relevant automated data acquisition tools that are currently available and applicable. These tools are divided up by purpose or field of searching:

1. Tools that serve for obtaining information on identity, network information and information on files: FOCA, Maltego, Recon-NG;
2. Tools that serve for obtaining network information and information on files: Metagoofil;
3. Tools that serve for obtaining network information: Shodan, Spiderfoot;

4. Tools that serve for obtaining network information and information on identity: The Harvester, IntelTechniques (limited functioning).

3.3 Models/methods of obtaining data for decision-making purposes

With the aim of improving information searches using data acquisition from publicly accessible sources, numerous authors have presented a method of obtaining data or have set out a model that takes the user through the process.

A simple but effective way of obtaining and analysing data has been presented by Černya et al. (2019) in their attempt to use publicly accessible data in the pharmaceutical sector. They have found that using publicly accessible user data on the use of tablets is possible to obtain a legitimate picture of the state of health trend in the population, which can be used very effectively in terms of competitive advantage. In their experiment the skill level of the analysts was crucial, along with an actual knowledge of where to obtain the data.

Obtaining data from publicly accessible sources can be applied in a broader sense, and we can search for all publicly available information, and be more targeted, for instance to a company's website. Company websites contain some vital information that can be used to gain a competitive advantage, since the data can provide a knowledge of the advantages and weaknesses of competitors (Hu & Zhu, 2013). A review of websites, web applications, and competitors' network can reveal their vulnerabilities and possible sensitive (personal) data, which is made public (Sondrava et al., 2021). In the opinion of Hu and Zhu (2013), the following information is relevant for competitive intelligence: the popularity of the website, age of the domain, appearance in search engine results, number of single device visits (meaning a visit using a certain device, although the same device has visited a website multiple times), the number of visited pages, links to pages, response time and website ranking. The information obtained must be set in the right context and correctly interpreted for an organisation to derive value from it.

Just as we can obtain data on turnover from a website, using tools to obtain data from publicly accessible sources we can rapidly gain information on the vulnerability of a specific website or application. These vulnerabilities appear in the form of leaks of sensitive data, such as account access information, open ports on a network, unprotected web services, outmoded software and an operating system running in the background (Sondrava et al., 2021). Such data are useful for potential attackers or for organisations that search for this information in themselves in the case of a penetration test. Were such data to be used or exploited for the purpose of a competitive advantage, this would constitute a crime, regardless of the fact that the data were freely accessible.

Actual searching on websites can also be time-consuming, since a first search can yield a lot of hits that need to be looked at. Resolving the questionable quality of results and gaining higher quality searches using web browsers is also addressed by Rasekh (2015), who designed a new search system based on competitive intelligence activity with the implementation of a high-quality browser. The proposed system is a combination of an ICA algorithm and a system

of ranking based on links. The described system needs a few more adjustments for use on the internet.

In recent years the concept of big data has of course been a major topic, and researchers are aware that the use of such data can offer value. Di et al. (2014) have identified the importance of competitive intelligence in the increasingly digitalised world and highlight the vast quantities of data generated every day and available on the web for use in conducting in-depth analysis that can signify an extraordinary competitive advantage. Through effective analysis organisations can recognise trends and also dangers, and because of this information they possess, they can make higher quality decisions, or decisions that are better than those of competitors. A similar conclusion was drawn by Sharma et al. (2013), who take the view that older approaches (before the arrival of the big data age) have major deficiencies in obtaining data for the purpose of competitive. Organisations can now obtain information on a dynamic business environment and markets and analyse consumers' responses and opinions in various social networks, forums, and blogs. For the actual analysis of data they present a technical solution for structuring and visualisation of large quantities of unstructured data with a user interface for easier use. They define data analysis steps as follows: posing questions, gathering data, analysis of data, and proposed reactions to the findings.

The added value of competitive intelligence is also recognised by Wang et al. (2011), who take the view that competitive intelligence offer the necessary support for managing companies, along with strategic support. Their article presents a model for acquiring competencies to successfully implementing competitive intelligence, emphasising an organisational culture and other internal and external factors. They have found that competitive intelligence play an important part in the long-term development of a company, including the setting up and operation of various groups and departments within the organisation. A system can be effective if the organisation has set up good training mechanisms to maintain and develop a lasting competitive advantage.

Chi and Chen (2013) researched the area of collaborative competitive intelligence, where they have found that modern organisations, due to their complexity, often require cooperation among various departments, or rather inter-functional cooperation, as they call it. They designed a model for competitive intelligence based on a knowledge base system approach. The system is composed of domain ontology, task ontology and semantic rules. The aim of the system is to deliver to various departments in an organisation news relating to competitive intelligence, depending on the weighting or relevance of the news for the individual department. With a test sample of 512 news entries, the system carried out an average 78% level of recall and 85% accuracy in delivering news. The decision-making model had a correctness level of 99%, and with department weighting 96%. In general they assess the knowledge base system as successful, since it achieves the objective of steering news through the decision-making channels in a company.

3.3 Models for evaluating obtained data and tools

In using OSINT tools and techniques, the seriousness and quality of information can be problematic. It is no surprise that researchers started dealing with this area and designing models for evaluating obtained data and tools. Chae et al. (2019) developed a systemic approach to assessing the appropriateness of current and future tools for obtaining data from publicly accessible sources for military use. The model itself is not oriented towards competitive intelligence, but it is at least partly applicable in assessing the appropriateness of the tool (and consequently also the information obtained with these tools) in the commercial sector. In the qualitative model, which is part of the developed approach, in the assessment they touch upon the actual suitability of the tool in terms of military policy (disclosing data on users of the tool), system decision-making in the final results (quality of information is important), the time needed for training personnel and the limitations of the tool (e.g. necessary internet connection speed, possible false positive results and so on). All the main criteria are in the interest of every slightly larger organisation that has in its ranks a department for competitive intelligence activities.

It is harder to assess the validity of data from multiple sources than to assess tools for using data acquisition from publicly accessible sources. Dorton et al. (2019) developed an approach based on the principles of social sciences, and this enables the mass gathering of opinions and validation of certain information in a quantitative way, in other words a rapid quantitative definition of the validity of information or multiple pieces of information from several observers. Theoretical research is important for conducting qualitative research as well as for analysing information obtained by means of data acquisition from publicly accessible sources and for obtaining data from human sources (Human Intelligence or HUMINT). At this level there are several more limitations to the research. One limitation is that there needs to be empirical research that would confirm or reject the theoretical assertions. Checking assertions is based on axioms which people do not observe consistently, moreover it is assumed that each observer has the same credibility, which is not true in practice. The model also needs to be assessed in a larger group of observers (in the case of theoretical research there were 15 observers). It is also questionable that the model relies on the memory or capacity to recall of individuals, which can cause errors in the validity of the information, since people can remember an event mistakenly.

4 DISCUSSION AND CONCLUSION: WHAT DO WE KNOW AND WHAT NEEDS TO BE FURTHER RESEARCHED?

Using data acquisition from publicly accessible sources is a topical subject that has recently attracted considerable attention, especially in terms of modern technology. For the most part, research related to the use of tools for data acquisition from publicly accessible sources refers to information technology and digital information, which are freely available on the internet. More recent works on this subject no longer talk about classical approaches to obtaining information from people or from print media, although obtaining information

from people is still an important part of competitive intelligence, when we speak of decision-making over important matters. There are a few available tools for obtaining information from public sources, but the majority are problematic since they also capture irrelevant information, and consequently for the general public the acquisition of data from publicly accessible sources can be time-consuming and analytically demanding. For this reason, the majority of companies do not conduct competitive intelligence, or conduct them in a limited scope, since they lack, or else cannot afford, technical and analytical staff that could cover this area

Researchers write about obtaining data from publicly accessible sources and competitive intelligence, but there has still not been any extensive empirical research in this field. In the review of the literature, we had to set the search parameters very broadly, in order to capture all relevant sources, since the industry uses a variety of expressions for similar approaches and fields. Researchers are not always in unison in their use of terminology, which probably lies in the fact that the field of data acquisition from publicly accessible sources is developing very rapidly, and consequently new sub-fields are emerging that attempt to define in greater detail or analyse the activity itself. It is possible to trace in the work of researchers an inclination towards and recognition of the importance of obtaining data from publicly accessible sources also in competitive intelligence, while at the same time they are unanimous that a great number of unanswered questions remain in relation to the immeasurable quantities of data and the necessary technical support for processing and analysing the data. One obvious obstacle is a lack of personnel, who are often inadequately trained to obtain data, or lack analytical abilities to correctly interpret the data obtained and deliver this to management in the form of useful information in support of decision-making processes.

We see numerous possibilities for further work, including greater empirical research in the field of using tools to obtain data from publicly accessible sources and to obtain data from publicly accessible sources for the purpose of competitive intelligence. Research would need to be conducted in organisational environments, involving work on real (decision-making) cases, where the real value of such use could be measured. One possibility is working on the design of an approach model for conducting competitive intelligence using data acquisition from publicly accessible sources, which would serve as a framework in the business environment and whose implementation would offer added value. It would be designed in such a way that in addition to an organisational approach, it would resolve the issue of training for employees and those conducting data acquisition from publicly accessible sources for clients.

The systematic literature review has shown that in the area of obtaining data from publicly accessible sources and competitive intelligence, there is still a lot of room for development, since the distance between where we are and where we would like to be is huge. The path to effective use of obtaining data from publicly accessible sources in organisational environments still faces some obstacles, which can be removed in the coming years, but of course the entire process must have proper IT support and must enable both organisational and technical implementation in various commercial environments.

REFERENCES

Anand, A., Buvanasri, A. K, Meenakshi, R., Karthika, S., & Mohan, A. K. (2020). PeopleXploit- A hybrid tool to collect public data. In *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*. IEEE. https://doi.org/10.1109/ICCCSP49186.2020.9315266

Al-khateeb, S., & Agarwal, N. (2020). Social cyber forensics: Leveraging open source information and social network analysis to advance cyber security informatics. *Journal of Computational and Mathematical Organization Theory, 23*, 412–430. https://doi.org/10.1007/s10588-019-09296-3

Ansari, F., Akhlaq, M., & Rauf, A. (2013). Social networks and web security: Implications on open source intelligence. In *2013 2nd National Conference on Information Assurance (NCIA)*, (pp. 79–82). IEEE. https://doi.org/10.1109/NCIA.2013.6725328

Benes, L. (2013). OSINT, new technologies, education: Expanding opportunities and threats. A new paradigm. *Journal of Strategic Security, 6*(3), 22–37. http://dx.doi.org/10.5038/1944-0472.6.3S.3

Bernard, R., Bowsher, G., Milner, C., Boyle, P., Patel, P., & Sullivan, R. (2018). Intelligence and global health: Assessing the role of open source and social media intelligence analysis in infectious disease outbreaks. *Journal of Public Health, 26*, 509–514. https://doi.org/10.1007/s10389-018-0899-3

Best, C. (2011). Challenges in open source intelligence. In *2011 European Intelligence and Security Informatics Conference*, 58–62. IEEE. https://doi.org/10.1109/EISIC.2011.41

Buccafurri, F., De Angelis, V., & Idone, M. F. (2020). Implementing multiple-social-network meta-APIs to support OSINT programming. In *2020 12th International Conference on Advanced Infocomm Technology (ICAIT)*, (pp. 124–128). IEEE. https://doi.org/10.1109/ICAIT51223.2020.9315457

Černýa, J., Potančoka, M., & Molnára, Z. (2019) Using open data and Google search data for competitive intelligence analysis. *Journal of Intelligence Studies in Business, 9*(2), 72–81. https://doi.org/10.37380/jisib.v9i2.470

Chae, J., Graham, D., Henderson, A., Matthews, M., Orcutt, J., & Song, S. (2019). A system approach for evaluating current and emerging army open-source intelligence tools. In *2019 IEEE International Systems Conference*, (pp. 1–5). IEEE. https://doi.org/10.1109/SYSCON.2019.8836885

Chi, Y. L, & Chen, T. T. (2013). Collaborative competitive intelligence: A knowledge base system approach. In *2013 8th Iberian Conference on Information Systems and Technologies (CISTI)*, (pp. 1–4). IEEE. https://ieeexplore.ieee.org/document/6615735

Di, J., He, B., & Li, W. (2014). Research on enterprise competitive intelligence development and strategies in the big data era. In *2014 IEEE International Conference on Computer and Information Technology*, (pp. 658–663). IEEE. https://doi.org/10.1109/CIT.2014.119

Dolenc, S. (2014). Kaj je Big Data? Kvarkadabra. https://kvarkadabra.net/2014/09/kaj-je-big-data/

Goujon, B. (2011). Text mining for opinion target detection. In *2011 European Intelligence and Security Informatics Conference*, (pp. 322–326). IEEE. https://doi.

org/10.1109/EISIC.2011.45

Dorton, S. L., Frommer, I. D., & Garrison, T. M. (2019). A theoretical model for assessing information validity from multiple observers. In *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, (pp. 52–58). IEEE. https://doi.org/10.1109/COGSIMA.2019.8724242

Gonçalves Evangelista, J. R., 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, 16*(3), 345–369. https://doi.org/10.1080/19361610.2020.1761737

Hribar, G., Podbregar, I., & Ivanuša, T. (2014). OSINT: A »Grey Zone«? *International Journal of Intelligence and Counter Intelligence, 27*(3), 529–549. https://doi.org/10.1080/08850607.2014.900295

Hu, L., & Zhu, M. (2013). Competitive intelligence acquisition from websites. In *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, (pp. 858–862). https://doi.org/10.1109/FSKD.2013.6816314

Li, D., Zhou, X., & Xue, A. (2020). Open source threat intelligence discovery based on topic detection. In *2020 29th International Conference on Computer Communications and Networks*, (pp. 1–4). IEEE. https://doi.org/10.1109/ICCCN49398.2020.9209602

Madureira, L., Popovič, A., & Castelli, M. (2021). Competitive intelligence: A unified view and modular definition. *Technological Forecasting and Social Change, 173*, 121086. https://doi.org/10.1016/j.techfore.2021.121086

Microsoft. (2019). Načrtovanje varne prihodnosti za podjetja [Planning a secure future for your company]. https://www.microsoft.com/sl-si/microsoft-365/business-insights-ideas/resources/planning-a-secure-future-for-your-company

Pais, V. F., & Ciobanu, D. S. (2014). OSINT for B2B platforms. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (p.p. 720–723). IEEE. https://doi.org/10.1109/ASONAM.2014.6921665

Pastor-Galindo, J., Nespoli, P., Gomez-Marmol, F., & Martinez Perez, G. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trend. *IEEE Access, 8*, 10282–10304. https://doi.org/10.1109/ACCESS.2020.2965257

Rai, B. K., Verma, R., & Tiwari, S. (2021). Using open source intelligence as a tool for reliable web searching. *Journal of Computer Science, 2*(402). https://doi.org/10.1007/s42979-021-00777-4

Ranjan, J., & Foropon, C. (2020). Big data analytics in building the competitive intelligence of organisations. *International Journal of Information Management, 56*. https://doi.org/10.1016/j.ijinfomgt.2020.102231

Rasekh, I. (2015). A new competitive intelligence-based strategy for web page search. In *2015 Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, (pp. 120–126). IEEE. https://doi.org/10.1109/ICOSC.2015.7050789

Schaurer, F., & Störger, J. (2013). The evolution of open source intelligence (OSINT). *Journal of U.S. Intelligence Studies, 19*(3), 53–56. https://www.afio.com/publications/Schauer_Storger_Evo_of_OSINT_WINTERSPRING2013.

pdf

Semerkova, L. N., Zaretskiy, A. P., Divnenko, Z. A., Grosheva, E. S., & Vishnevskaya, G. V. (2017). Application of information technologies in competitive intelligence. *In 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM)*, (pp. 804–807). IEEE. https://doi.org/10.1109/SCM.2017.7970730

Sharma, D., Chaudhary, K., Vaidya, P., & Jora, K. (2013). Big data – competitive intelligence. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 684–689). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7100336

Sondrava, S., Sharma, P., & Dholariya, D. (2021). Prevention to sensitive information disclosure via OSINT. *International Journal of Scientific Research in Science, Engineering and Technology, 8*(3), 109–114. https://doi.org/10.32628/IJSRSET218317

Wang, Y., Zhao, X., & Zhang, X. (2011). Construction and operation of cultivation model for enterprise competitive intelligence competence. In *International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, (pp. 2004–2007). IEEE. https://doi.org/10.1109/AIMSEC.2011.6010955

Yaboah-Ofori, A., & Brimicome, A. (2017). Cyber intelligence & OSINT: Developing mitigation techniques against cybercrime threats on social media. *International Journal of Cyber-Security and Digital Forensics, 7*(1), 87–98. http://dx.doi.org/10.17781/P002378

Yang, H. C., & Lee, C. H. (2012). Mining open source text documents for intelligence gathering. In *2012 International Symposium on Information Technologies in Medicine and Education*, (pp. 969–973). https://doi.org/10.1109/ITiME.2012.6291464

About the author:

Žiga Primc holds a bachelor's degree in information security and a master's degree in criminal justice from the Faculty of Criminal Justice and Security, University of Maribor. He is a private detective, licensed as an ethical hacker. E-mail: ziga.primc@student.um.si