Conceptual Model of an Automated System for Processing Information From Open Sources and Detecting Information Deviations

Ralitsa Yotova
University of Library Studies and Infromation Technologies,
National Security Departament
Sofia, Bulgaria
r.yotova@unibit.bg

***Abstract*. This research focuses on the development of a Conceptual Model of an automated system for processing information from open sources and detecting information deviations. The purpose of the model is to provide a framework within which to develop and implement technologies and methods that will enable the system to effectively collect, process, and analyze information from a variety of sources.**

The automated system concept is intended to improve the efficiency and accuracy of intelligence work by using automated methods and algorithms to analyze large volumes of data and information. To achieve this goal, the components and functions of the system will be discussed, as well as the ways in which they interact.

The model proposes an integrated approach combining different technologies and methods to achieve efficient information processing and detection of deviations. The proposed system has the potential to be applied not only in the security sector, but also in various fields such as business, finance, medicine, and others where information from open sources is essential for decision making.

Keywords: open sources, information, model, analysis

I. Introduction

It is definitively clear that in today's world, where information plays an ever-increasing role, intelligence faces significant challenges in the collection, processing and analysis of information gathered from open sources.

Advances in modern open source intelligence, data mining, machine learning, digital forensics, and most importantly, the increasing computing power available for commercial use, are enabling OSINT practitioners to significantly accelerate and even completely automate intelligence collection and analysis.

As the information space expands, the OSINT toolset is constantly changing and improving. This statement is not at all surprising given that open source information has continually increasing volume. To meet this challenge, more and more effective techniques are being developed and introduced which, together with the development of artificial intelligence systems, make productivity inextricably linked to the quality of the technical tools used by analysts.

Undoubtedly, conventional methods of collecting and processing information are becoming increasingly inefficient and unable to respond to the rapidly changing environment. For this reason, the implementation of automated systems to deal with the collection, processing and analysis of information from open sources is becoming more and more necessary.[2-6]

Based on advanced algorithms and operating models, such systems can process and analyse large volumes of data and information faster and more efficiently, assisting the human factor. Moreover, they make detecting information deviations, filtering relevant information and extracting more accurate and up-to-date data an instantaneous process. These types of systems have the potential to change the way security services operate and support their activities many times over. [7-8]

Not only in the field of intelligence, but also for the implementation of various activities that characterise the current social reality, time is of the essence and constant monitoring and timely response in decision-making are vital. [9-12] With the increased levels of network connectivity and constant interaction that underpin the modern information society, monitoring, collecting and analysing data and information from different sources becomes an almost impossible mission without the capabilities of information technology to support these processes. [13-16]

Print ISSN 1691-5402
Online ISSN 2256-070X
https://doi.org/10.17770/etr2024vol4.8230

A major challenge in the processing and analysis of open sources becomes the ability to convert into relevant information large volumes of data, which in most cases are unstructured, unorganized, come from questionable sources, in different forms and from different channels of information. The development and implementation of specific methods and tools to create, validate and improve databases tailored to intelligence needs becomes a high priority.

II. Materials and methods

In this regard, the proposed conceptual model is based on the well-known JDL-model. For this reason, its architecture is rather functional and aims at the operational provision of a segment "necessary operational capabilities" [1] of the central key organizational competences of the state power, directly concerning the structures of the national security system, among whose main activities is the information analysis.

As a methodology of development, the Conceptual Model is constructed of distinct modules according to the specificity of information analysis activities, which partially overlap with the methodology of other methods of structured analysis.

All the components and sub-processes of the structural and functional scheme of the model are appropriately integrated into a common system for generating new information, as well as for simulation modelling, research and training.

The model aims to represent the totality of all components and subsystems and their interaction in the implementation of analytical methods on information from open sources as a basis for the development of an information processing system. Reducing the involvement of the human factor in the processing processes will reduce subjectivity and increase objectivity in the results obtained as a consequence of the automated system.

Based on what has been presented so far, a possible implementation of an algorithm for the operation of the automated system proposed by the Conceptual Model for processing information from detected sources and for detecting information deviations generally includes the following steps:

1. Collecting a training data set

The system collects a training data set that includes information from open sources classified as "normal" or "deviating". This data set is used to train a Bayes classifier.

2. Building a Bayes model

The system builds a Bayes model based on the training data set. This model is used to determine the probabilities of occurrence of various features or attributes in the normal and outlier data.

3. Classification of new data

After successful training of the Bayes model, the system can classify new data from open sources. This is done by calculating the probabilities of occurrence of various features or attributes in the new data and using the Bayes model to determine the classification of this data as "normal" or "deviating".

4. Detection of information deviations

The system uses the classification results of the new data to detect informational deviations. If the new data is classified as "deviating", the system may generate a warning or take other actions to signal the presence of a deviation.

III. Results and discussion

The presented methodology and tools of the research allowed to present the Conceptual Model of an automated system for processing information from open sources and detecting information deviations with the following functional characteristic:

The "entrance" of the information processing system with "dashed circles" in green color represents the movement of information from the information environment to the system. These circles present an array of information representing the interrelationships between events and facts occurring in the objectively existing environment and the occurrence of a problem in the organizational structure, for the purpose of which a solution needs to be found following the application of the Content Analysis method. The graphic shows the process (flow) of information from the objectively existing environment to the information processing and analysis system (Fig. 1).

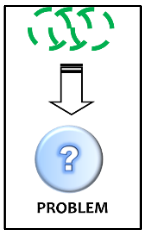


Fig. 1. Entrance

In Module 1 "Areas of Application", the interplay and intertwining of the different areas of human activity that carry out information and analytical work is presented through an iridescent coloured sphere. Thus, the role of the Content Analysis method and the possibilities of its automation in the study of individual processes and events in the fields of national security, politics, media, and medicine are shown. Since Content Analysis is a widely applicable method, the possibility of applying it in other areas of science and human activity is also shown (Fig. 2).

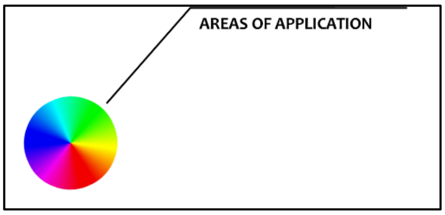


Fig. 2. Areas of Application

Component 2 "Expert" is a core component of the Conceptual Model. Considering the role of the expert in the model, it has basic control functions, which consist not only in controlling and monitoring the operation of the automated system, but also to monitor the flow of information processing in the individual stages. The expert's knowledge and experience play an important role in deciding how to deal with the information received and, in particular, whether to run the automated information system and whether it is suitable for processing the information in question.

A open sources monitoring system is presented in Module 2 by means of a "colour sphere" (Fig. 3). It is a set of technical means (so-called "sensors") by which information is gathered from a self-updating database containing detected information sources. Through them, it provides sensitivity (sense, sensibility) to the movement of the detected sources, and in interaction with them it converts its response into signals, which are coded messages about quantitative or qualitative characteristics of the state of these sources. Its essential characteristic contains a set of specific technical means, representing sensors and sources of information, which provide objective information with respect to the objectively existing environment.

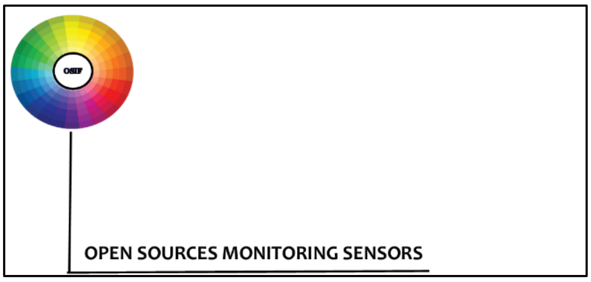


Fig. 3. Open sources monitoring sensors

The main task of the sensor system is to monitor the detected sources of information, thus collecting and summarizing the readings of the different sensors (the types of detected sources) in order to obtain an overall picture of the environment that meets the requirements of completeness, reliability, accuracy, reliability and objectivity.

Module 3 (Fig. 4) presents the integrated databases in the automated information processing system.

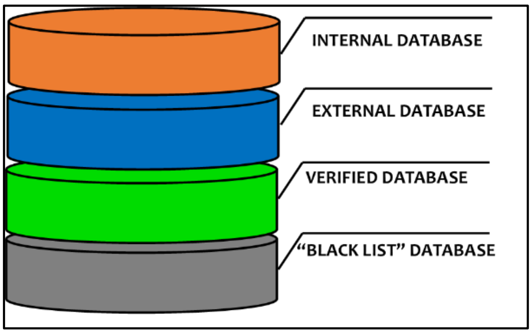


Fig. 4. Database

Databases are essential for the final evaluation and verification of information. In its autonomy and after comparison with the original, basic information, the automated system stores the generated information that does not meet the criteria of information reliability and source reliability in a "black list" database. Thus, the system will subsequently match the collected information with the existing information in the database and not use it.

The guiding task here is the overall monitoring and maintenance of the currency of the sources reviewed and the data used, which should ensure informed decision making at the expert end of the chain.

Subsystem 1 "Defining Analysis Sources" presents the first stage of the application of the Content Analysis method in the automated system, where the selection of the main information sources to be fed into the automated system is performed. This stage of the information analysis is an important part of the overall information analysis process that is the Content Analysis method. The definition of the basic text (data) that the system will process also determines the quality of the results obtained at the end. Both in this stage and in each subsequent stage, the expertise and knowledge of the expert is paramount in controlling the overall system and its effectiveness (Fig. 5).

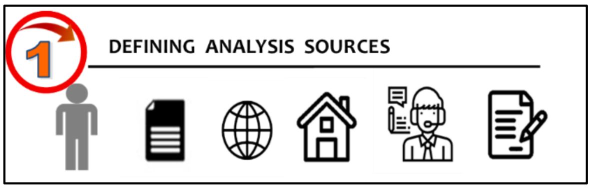


Fig. 5. Defining analysis Sources

Subsystem 2 of the Conceptual Model presents Stage 2, which shows the process of selecting information from the common set of open sources of information. The presented graphic (Fig. 6) shows the process of reviewing and selecting specific information in the system relevant to the problem at hand. In this way, the sources of information that are not suitable for the purpose of the study are eliminated, and only those that meet the content constraints set in the system for the most relevant data are extracted.

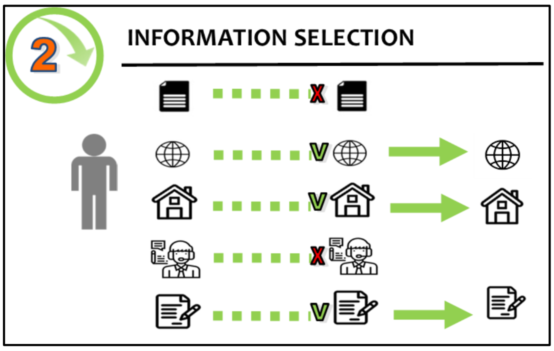


Fig. 6. Information Selection

Subsystem 3 also presents the next stage of automated Content Analysis, which shows the process of identifying the units for analysis. As can be seen in Fig. 7, these units can be words, paragraphs, sentences, symbols or specific topics. By defining predefined criteria for the selection of the units of analysis by the expert, only the information that meets the demand and is suitable for further processing is extracted. Those units that are not large enough to have any semantic value or are too long are dropped from the system, resulting in ambiguities that can cause the system to fail or lock up. An important point at this stage in the search for qualitative entities is that they must be easy to identify and must be contained in a large enough volume of information so that identification can take place.

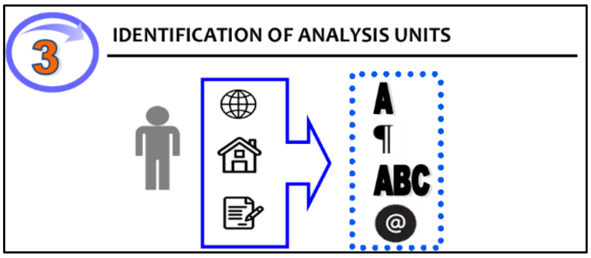


Fig. 7. Identification of Analysis Units

Subsystem 4 of the presented Conceptual Model includes an overview and distribution of the units of analysis. This stage also largely determines the effectiveness of the system as certain units may match in meaning or have a specific character. In this case, the system, processing the accumulated units of analysis from the previous stage, determines the frequency of mention of the selected units, discarding a surplus of them that do not meet the set parameters and criteria. At the end, only the information that has a match in meaning, content or frequency of occurrence in a given text is extracted (Fig. 8).

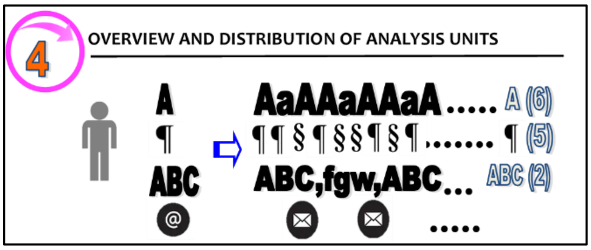
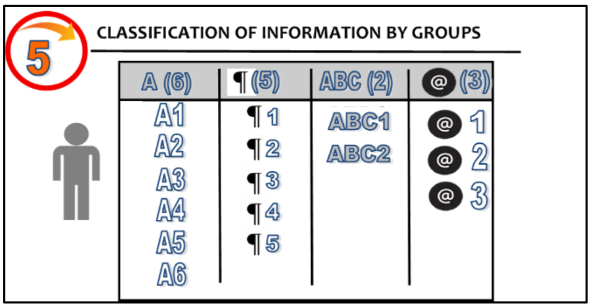


Fig. 8. Overview and Distribution of Analysis Units

The fifth stage, which is Subsystem 5 of the Conceptual Model structure, presents the direct counting of the results obtained from the previous stage in terms of the frequency of mention of units of analysis. Tables, computer programs and statistical calculations are most often used to successfully implement this stage of the analysis. The system tabulates the results obtained for each of the units of analysis (A; ¶; ABC; @) and performs an automatic reclassification by group of the number of these units (Fig. 9).



A (6)	¶ (5)	ABC (2)	@ (3)
A1	¶ 1	ABC1	@ 1
A2	¶ 2	ABC2	@ 2
A3	¶ 3		@ 3
A4	¶ 4		
A5	¶ 5		
A6			

Fig. 9. Classification of Information by Groups

Subsystem 6 (Fig. 10) is also the sixth and final stage of the system operation, but not the last component of the presented Conceptual Model. Upon completion of the information assurance and analytical portions of the system operation, the resulting information product passes through the presented filter for validation and recycling, which stands before the final output point of the overall Conceptual Model.

The purpose of the filter in the system is to scan and detect any information anomalies in the resulting information product. For this reason, the filter has sensors to identify content errors.

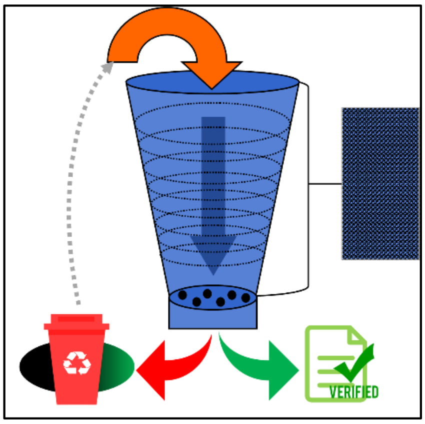


Fig. 10. Information Deviation Detection

and Recycling Filter

The main criterion of the filter operation is based on the principle of authenticity (A = authencity) of the information and reliability (R = reliability) of the source, borrowing for this purpose the functionality and principle of operation of the sensors for monitoring and detection of deviations.

In order to illustrate the filter operation mode, first of all, the information flow (iF) and its possible trajectory change under Average Deviation (Da) and Absolute Deviation (DA) are presented in Fig. 11.

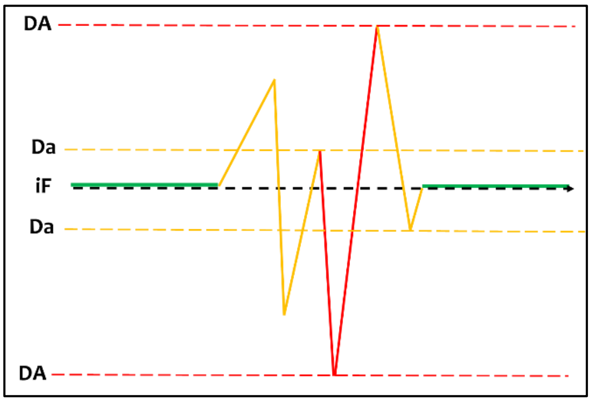


Fig. 11. Graphic Showing Deviations of the Information Flow – example

The information flow will remain on the rights then, when the structural or functional characteristics of the information flowing in it are not influenced by external or internal factors.

Then:

iF = f + s

In the case of mean information drift on the information flow, a partial manipulation has been exerted on the structural **(s)** or functional **(f)** characteristics of the information flow, as a consequence of which it will change its trajectory. Examples of such manipulations are misinformation, propaganda, fake news, deception or online threats.

Then:

Da = A (f - s) - R (f + s)

Or

Da = - A (f - s) + R (f - s)

Da = A (f + s) + R (f - s)

Absolute deviation will be reported by the filter when the information does not meet the criterion in either the credibility or reliability part.

Then:

DA = - A (f - s) - R (f - s)

In the latter case, the information will be transferred to the database for recycling and storage. In this way, the automated system will use the accumulated information resource that does not meet the given criterion, using it to build a database with low credibility and reliability, which the system will then collate and not use. After the recycling mode, information that partially meets the criteria will be transferred for re-verification and final validation.

Information Validation **(iV)** will be executed by the system when the validity of the information and the reliability of the source have been confirmed and no structural or functional changes to the information flow have been identified:

iV = A (f + s) + R (f + s)

Recycling in the filter is the process of storing the information in the databases. In the case where the information does not meet the specified filter performance criterion, it will be stored in the blacklist database. In this way, by learning itself, the system will continuously add to this "list" and will not output the same content again.

Similarly, in the second case, the information that has gone through the non-verification process and comes out in the output as text will be stored in the verified database, which will be an additional guarantee of the quality of the source and will increase the value of the databases.

The seventh stage of the presented Conceptual Model consists in performing an expert interpretation of the results obtained due to the operation of the automated system for processing information from open sources and detecting information deviations.

In this stage, those characteristics of the generated result are identified and evaluated that allow general conclusions to be drawn, such as what is the meaning of the information obtained, whether its content is sufficient to draw conclusions and recommendations, whether the main problem has been solved or whether the causal relationships in it have only been partially inferred. It is no coincidence that the expert as a human factor in the automated system is at the centre of the triangle presented in Fig. 12.

His experience and knowledge determines the overall workflow of the system and, in particular, how the information will be processed so as to create opportunities to perform interpretation on it. This point is also strongly tied to the element of subjectivity in information processing and analysis. This is due to the fact that the experience and knowledge of the expert, based on the value system, world view, cognitive and cultural orientation possessed by him, will break and modify to some extent parts of the flow of information units that pass through all stages of the system. The human factor, other than that of the author or creator of the information in its primary form, will affect the interpretation of the information.

Thus, the result at the end of the proposed Conceptual Model represents the analyst's interpretation, containing his experience and knowledge, combined with the results produced by the automated system. This is also a kind of process of generating new information and adding new knowledge in the domain for which the automated system is used.

As the starting point of the whole system, the distribution/consumption of the final result obtained in the form of a solution to the problem that arises at the input of the presented Conceptual Model is presented. The presented graphic shows the generated solution as new information generated, which comes out of the output and is transmitted to the end user (management unit, organization, institution, area).

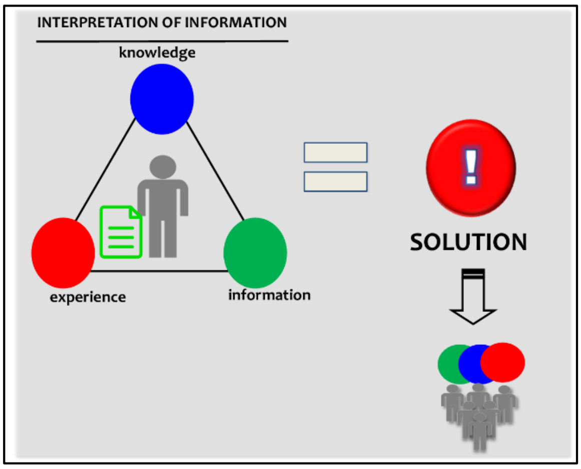


Fig. 12. Interpretation of Information

Conclusions

The developed Conceptual Model (Fig. 13) is suitable for the selection of technical tools and software applications to build a unified information architecture that:

- reduces the impact of incomplete, ambiguous and erroneous data;

- assumes the availability of data at a higher level of abstraction;

- identifies missing information and the need for additional information.

The development of the Conceptual Model based on the JDL-model and borrowing the ideas embedded in human-machine interface methods aims to ensure that the information content of interest will be presented in a form that is suitable for user perception.

The Conceptual Model aims to represent the body of primary knowledge that has acquired a socially relevant status and, at the same time, to represent through specific structurally distinct relationships the knowledge and values that define information-analytic activity and the basic characteristics of information-processing systems.

One of the advantages and contributions of the Conceptual Model is the clear presentation of all stages of the system in a detailed and structured way, which could serve as a scientific basis for the development of an actual system.

The structural and functional characteristics of the Conceptual Model allow to make the connection between the different factors of influence in the implementation of the automated system operation.

The advantages of the proposed Conceptual Model provide:

- All significant knowledge classes that are explicitly described.

- Context-sensitive nature of knowledge extraction algorithms that is observable and controllable.

- Dynamic real-time information and data processing.

- Maintaining a Database Management System containing static declarative knowledge that can be logically divided into context sensitive and context insensitive components.

- Capabilities to introduce correlation algorithms for multilevel, non-standard processing to produce a self-learning algorithm of the analysis procedure.

The use of an automated system to process information from open sources and detect information biases is necessary in intelligence for several reasons:

1. Efficiency: the automated system allows processing large volumes of data and information faster and more efficiently than human operators. This allows information acquisition and analysis to become an extremely fast process, which is essential in operational work.

2. Precision. It can use algorithms and models that are more accurate and reliable to detect potential threats and deviations.

3. Objectivity. The automated system is unbiased and objective as it is based on predefined rules and algorithms.

4. Scope. This allows for wider coverage of the information field and detection of unexpected relationships and patterns.

5. Security. This helps to reduce the risk of false or malicious information spreading and provide greater security to the collected data.

All of these factors make an automated system necessary and valuable to intelligence in processing information from open sources and detecting information deviations.

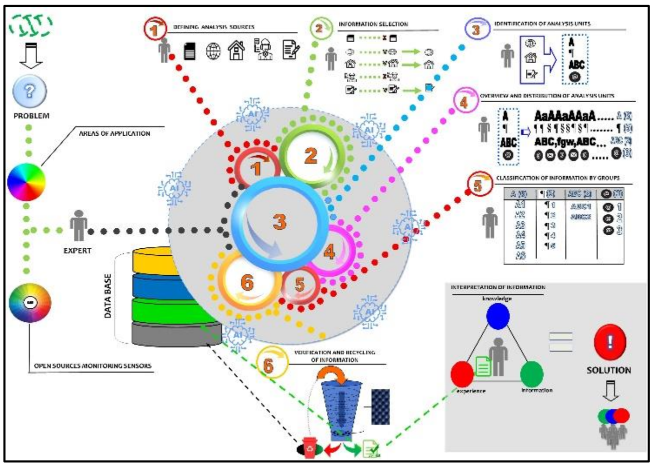

PROBLEM
AREAS OF APPLICATION
EXPERT
OPEN SOURCES MONITORING SENSORS
DATA BASE
1 DEFINING ANALYSIS SOURCES
2 INFORMATION SELECTION
3 IDENTIFICATION OF ANALYSIS UNITS
4 OVERVIEW AND DISTRIBUTION OF ANALYSIS UNITS
5 CLASSIFICATION OF INFORMATION BY GROUPS
6 VERIFICATION AND RECYCLING OF INFORMATION
AI
1
2
3
4
5
6
A
¶
ABC
AaAAaAAaA.....
ABC,fgw,ABC...
INTERPRETATION OF INFORMATION
knowledge
experience
information
SOLUTION

Fig. 13. Conceptual Model of an Automated System for Processing Information from Open Sources and Detecting Information Deviations

ACKNOWLEDGMENTS

This scientific paper was supported by the NSP SD program, which has received funding from the Ministry of Education and Science of the Republic of Bulgaria under the grant agreement no. Д01-74/19.05.2022.

REFERENCES

[1] Semerdzhiev, Ts. Strategicheski informatsionni sistemi. Subekti na avtomatizatsiyata. Sofia: Softreyd, 2007, 273 s. [bulgarian language]

[2] NATO Open Source Intelligence Handbook, November 2001. [Online] Available: https://github.com/lawsecnet/OPSEC/blob/master/NATO%20OSINT%20Handbook%20v1.2%20-%20Jan%202002.pdf. [Accessed: Feb. 20, 2024].

[3] OSINT Handbook by Open Source Center, Romanian Intelligence Service, 2018. [Online]. Available: https://bib.opensourceintelligence.biz/STORAGE/OSINT%20Handbook.pdf. [Accessed: Feb. 20, 2024].

[4] Thompson, J. R., R. Hopf-Weichel, and R. E. Geiselman. The Cognitive Bases of Intelligence Analysis, Arlington, VA: U.S. Army Intelligence and Threat Analysis Center Report No. R83-039C, 1984, pp. 2–9.

[5] Ungureanu, Gabriel-Traian. Open Source Intelligence (OSINT). The Way Ahead. – In: Journal of Defense Resources Management, Vol. 12, Issue 1 (22)/2021, p. 179.

[6] Williams, H., I. Blum. Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise. RAND Corporation, California, 50 p.

[7] Bielska, A., N. Kruz, Y. Baumgartner, V. Benetis. Open Source Intelligence Tools and Resources Handbook. Switzerland: I-Intelligence, 2020. 510 p.

[8] Hirschheim, R., H. Klein, K. Lyytinen et al. Information Systems Development and Data Modeling, November 2011. [Online]. Available: https://www.cambridge.org/core/books/information-systems-development-and-data-modeling/DE7DF31E05AB4F4BF579E7448167B715 [Accessed: Feb. 21, 2024].

[9] Cambridge: Cambridge University Press, 1995. 303 p.

[10] Yordanova, S. Informatsionni voyni. Informatsiyata – orazhieto na savremennia svyat. – V: Sbornik s dokladi ot Godishna universitetska nauchna konferentsia, 30 yuni – 1 yuli 2022, Veliko Tarnovo. Veliko Tarnovo: NVU „Vasil Levski“, 2022, s. 185–192. ISSN 1314-1937. [bulgarian language]

[11] Kazakov, K. Protivodeystvie na zabluda v natsionalnata sigurnost. – V: Obshtestvoto na znanieto i humanizmat na XXI vek: XIX natsionalna nauchna konferentsia s mezhdunarodno uchastie Sofia, 1 noemvri 2021 g. Sofia: Za bukvite – O pismenehy, 2021, s. 438–444. [bulgarian language]

[12] Kazakov, K. Strategichesko upravlenie na informatsionnite uslugi v sigurnostta. Sofia: Softtreyd, 2019. 232 s. [bulgarian language]

[13] Kazakov, K. Falshivite novini kato instrument za manipulirane na obshtestvenoto mnenie. – V: Obshtestvoto na znanieto i humanizmat na XXI vek: XIX natsionalna nauchna konferentsia s mezhdunarodno uchastie Sofia, 1 noemvri 2021 g. Sofia: Za bukvite – O pismenehy, 2021, s. 438–444. [bulgarian language]

[14] Yotova, R. Open Sources af Information. Collection, Classification And Processing. Sofia: Za bukvite – O pismeneh, 2023, 224 p. [bulgarian language]

[15] Zahariev, A. Informatsionnata sigurnost i zashtita na informatsiyata. – V: Sbornik nauchni trudove ot Nauchna konferentsia „Problemi na informatsionnata sigurnost prez XXI vek", Shumen, 2011. Shumen: Natsionalen voenen universitet „Vasil Levski", 2011, s. 245-250. ISBN 978-954-9681-49-9. [bulgarian language]

[16] Angelov, G. Arhitekturen podhod za opisanie na protsesite v komunikatsionno-informatsionni sistemi na organizatsionno-upravlenski strukturi. – V: Obrazovanie, nauchni izsledvania i inovatsii, godina I, knizhka 3, 2023, s. 36-42. ISSN 2815-4630. [bulgarian language]

[17] Boyanov, S. Energiynite resursi kato sredstvo za vliyanie v usloviyata na geopoliticheskо protivopostavyane. Nyakoi aktualni sabitia i proyavlenia, zasyagashti natsionalnata sigurnost. – V: Sigurnost i otbrana. Aktualno sastoyanie, vazmozhnosti i perspektivi. Sofia: Za bukvite – O pismenehy, 2023, s. 380-390. ISBN 978-619-185-593-3. [bulgarian language]