Ontology-Based Approach to Validation of Learning Outcomes for Information Security Domain

© Julia Rogushina[1], © Anatoly Gladun[2], © Serhii Pryima[3], © Oksana Strokan[4]

[1] Institute of Software systems of National Academy of Sciences of Ukraine, Kyiv, Ukraine, ladamandraka2010@gmail.com
[2] International Research and Training Center of Information Technologies and Systems of National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, Kyiv, Ukraine, glanat@yahoo.com
[3] Dmytro Motornyi Tavria State Agrotechnological University, Melitopol, Ukraine, pryima.serhii@gmail.com
[4] Dmytro Motornyi Tavria State Agrotechnological University, Melitopol, Ukraine, oksana.strokan@tsatu.edu.ua

Abstract. Validation of non-formal and informal learning results is an effective way to solve a number of socioeconomic problems in various spheres. Recognition of learning outcomes in dynamic domains requires the use of background knowledge acquired from open information resources. Need in such background knowledge causes the interest to ontological representation of learning domain and ontology-based methods of knowledge analysis. Now unstructured natural language texts constitute considerable part of the Web available content. The large data volume necessitates scalable means of analysis, and more efficient and rapid processing can be ensured through the use of task-specific thesauri based on domain ontologies. The approach proposed by the authors to recognize non-formal and informal learning outcomes is exemplified by information security domain. The referencing to this domain is determined by the heterogeneity and dynamics of its information sources, the complex hierarchy of knowledge as well as the growing need for information security specialists in the context of the digitalization of society.

Keywords: Ontology, Information Security, Validation of Learning Outcomes, Non-Formal Learning, In-Formal Learning, Unstructured Data, Big Data.

1 Unstructured data analysis

The rapid development and integration of various domains requires continual updating of relevant professional knowledge and skills for labor mobility. Therefore we need to analyze competence pertinence to new task types at all life activity stages of individuals and organizations. In turn, such knowledge dynamism complicates the learning outcomes determining in formal, non-formal and informal education. Recognition of the non-formal and informal learning outcomes is increasingly seen as a way to improve lifelong education for employment and life quality improving.

European countries emphasize the visualization relevance for learning organizes outside of official education institutions (at home, at work or leisure).

This problem is very important for rapidly changing domains that integrate different research activities. To work effectively in these areas, professionals need lifelong learning, gaining experience and new learning outcomes. These features apply equally to most of the information technologies. On conditions of digitalization it is especially important for information security (IS) domain, since practically all employees need now to have basic knowledge in IS for processing digital data without their loss and disclosure.

Individuals and organizations alike depend on the cyber world. As the world becomes more connected through digital environment, a highly skilled information security workforce is required to secure, protect, and defend global, national, commercial and personal information. The IS strategies and development plans require the know-how improvement of citizens, economic life actors and public administration.

A lot of people who need in modern IS skills and knowledge can receive them only from self-education. Therefore IS competence is not just another professional field of expertise. Rather, it ranges from civic skills all the way to international-level professions and should be included in different educational and institutional levels.

Validation of learning outcomes is a process that recognizes various learning outcomes with the help of such steps as identification, documentation, assessment and certification. Successful validation of learning outcomes is based on use of formalized knowledge of the relevant subject domain: such knowledge is used to match learning results obtained by individual with domain references.

Now the validation problem of non-formal and informal learning outcomes is complicated by the unstructured and unclear nature of many domains, such as IS. It is difficult to clearly define the boundaries of each domain, because many professions are located at the intersection of several industries or have unformalized domain-independent requirements (such as response rate, stress resistance) and require external knowledge from other fields. The problem solution requires the use of modern methods for representation and analysis of knowledge that ensure the information processing at the semantic level.

In this paper we consider IS as an example of complex and dynamic domain where the need to validate the results of non-formal and informal learning is very actual but faces many theoretical and practical obstacles.

2 Formulation of the problem

Digitization of all life spheres of modern society [1] increases the need for relevant information security means and awareness in this area for a wide variety of users and developers of information technology. Therefore, a significant number of people have gained knowledge in the process of self-study, i.e., by non-formal and informal education in this field. Informality of this process and diversity of terminology used for describing of such outcomes complicates to obtain formal definition of learning results for employers for matching with clear criteria and requirements for job responsibilities. This necessitates a reconciliation of domain terminology with knowledge

structures pertinent to current IS standards. But now the big number of standards and examples deals with various IS aspects, new versions and editions appear frequently. Therefor Big Data methods can be applied for tracking of all changes in this sphere.

We suggest to harmonize a hierarchy of IS ontologies to define the terms of this domain and construct thesaurus of IS. Each term of this thesaurus is associated with one or more elements of these IS ontologies (concepts, relations etc.). Use of this thesaurus allows describing various learning outcomes (by natural language (NL) texts or by structured data) obtained by non-formal and informal learning, and provides their comparison with formalized (or semi-formalized) descriptions of jobs and tasks that arise in this field. Such IS thesaurus helps to take into consider the most task-relevant domain knowledge. Time required for processing and construction of thesaurus is much less than the time of matching of domain ontologies with unstructured NL text, which is usually used to describe learning outcomes and vacancies. Thus, the approach proposed in the work should ensure the transition from processing of unstructured large-scale data to analysis of much more compactly represented knowledge obtained from appropriate Big Data and from domain ontologies.

3 Information Security Domain Specificity

The current tendencies regarding the storage, exchange and processing of information are characterized by the intensive implementation of information technologies, the spread of local, corporate and global networks in all life spheres of the state and individuals. This situation creates new opportunities for information exchange and requires secure and safe information at the same time.

The relevance of this problem is determined by the following factors:

- the exponential growth of personal computers, mobile devices and other information exchange means;
- rapid expansion of users with direct access to digital networks and information resources;
- an exponent volume growth of that is accumulated, stored and processed in digital form;
- spread of hardware, software and information technologies that do not satisfy the current safety requirements;
- discrepancy between the rapid development of information processing facilities, the national laws of information security and the development of international standards and legal norms that will provide the necessary level of information protection;
- information war in and around the country caused by external aggression;
- cybercrime, sometimes of national importance, etc.

IS deals with the protection of information systems, both hardware and software, from theft, unauthorized access and disclosure, as well as from intentional or accidental harm. IS has to protects all segments related to the Internet (from the networks themselves to the information transmitted over the network) and stored in databases, to the various applications and devices that control the operation of the equipment through network connections.

Today IS becomes a cornerstone in the activities of every organization or in-

dividual. IS refers to the security of entire organization data against deliberate or accidental actions that cause damage to its owners or users. IS domain is a very dynamic sphere that requires constant monitoring of information coming from both internal and external information resources.

Information resources in IS are generated usually by different tools, sensors and systems expressed using different standards and formats, published by different sources and is often scattered as isolated pieces of information. Furthermore, cyber security data are available in structured, semi-structured and unstructured forms from both, internal sources i.e. within the organization, and external sources i.e. outside the organization. Unifying such scattered information provides better visibility and situational awareness to cyber security analysts.

Now new terms in IS domain appears frequently, and old ones change the meanings making it difficult to agree on different approaches to knowledge representation about IS. In addition, documents (standards, protocols, descriptions) from the IS are characterized by large volume and not structured form. Therefore, it is advisable to use Big Data analytics methods – a new technology that enables the collection, storage, processing and visualization applied for such data volumes. Such analytics takes into account the basic Big Data characteristics [2].

Specialized analytical methods extract useful information from distributed data to obtain the information about the specialists` learning outcomes relevant to the challenges in the IS field. Analytics Big Data can be used to make decisions in such IS tasks [3].

Big Data Analytics applied to IS data can provides insights useful to improve current security practices of network operators and administrators. To use Big Data as an external information source, we first need to filter out multiple datasets by pertinence to task of analytics. Preliminary data cleaning and preparation for analysis can be performed by metadata analysis (and quite often the analysis of the NL part of it, for example, annotations) using the domain knowledge.

It is advisable to treat some information as Big Data if it has one or more characteristics from the "5V": Volume, Variety, Velocity, Value, Veracity [4]. Big Data, also referred to as Data Intensive Technologies, now are becoming a new technology trend in science, industry and business [5] . Big Data can be defined as a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high velocity capture, discovery, and/or analysis.

Now Big Data technology is applied in different human activity domains empowered by significant growth of the computer power, ubiquitous availability of computing and storage resources, increase of digital content production, mobility and security. Each stage of the Big Data lifecycle provides the data transformation or changes in processing of the dataset content. In many cases original data and processed data are linked by referral integrity. This motivates such Big Data features as dynamism and variability that reflect the fact that data are in constant change and may have a definite state, besides commonly defined as data in move, in rest, or being processed. Supporting these data properly will require scalable provenance models and tools incorporating also data integrity and confidentiality.

Problems of big data use for analytic in IS often are caused by the lack of descriptions of metadata datasets. Re-use of data requires as much information as possi-

ble about the origin of the data and about a complete history of the methods used to collect, maintain and analyze. The availability of metadata increases the likelihood that the datasets are reused correctly.

Widespread Big Data used as a source of background IS knowledge need in the scalability of methods that can be used to structuring and validation of learning outcomes in this domain.

4 Unstructured data in the IS domain

Today, the major portion of cyber security data is represented by unstructured NL data. Unstructured data (USD) is information that does not have a predefined model of organization. This causes problems related to storage (traditional databases are not designed for such uncertainty) and analysis. USD contain potentially the greatest value as sources of new knowledge, and the total amount of such data influences on accuracy of the analysis results.

In many cases, USD are represented by textual information in electronic form by sets of natural language words of arbitrary length, combined according to weakly formalized linguistic rules. Such textual information contains the most useful information for further use but USD may also contain dates, numbers, special symbols etc.

Sometimes it is difficult to distinguish between structured and unstructured data. One of the criteria for determining whether data are structured is to create a parser for the data element. USD as a term is not well defined for several reasons [6]:

- data contain the structure without it's formal definition;
- data structure is not useful for processing purposes;
- information about data structure cannot be processed automatically without further clarification.

Data are considered as USD in cases where information about their structure cannot make the analysis of these data more efficient. Therefore, for example, standards for IS (both domestic and international) in terms of data analysis are considered as USD, although they have certain structural elements.

A lot of important cyber security information can be extracted from USD. For example, various cyber security vulnerabilities are typically identified and accessible publicly on the Web but at first not in official documents. It causes a strong need for systems that can automatically analyze unstructured text and extract vulnerability entities and concepts from various non-traditional unstructured data sources such as cyber security blogs, security bulletins and hackers forums. Now there is no widely used automatic mechanism to understand and process such USD with non-formalized terminological base.

Technologies such as Data Mining, Natural Language Processing, and Text Mining provide different methods for structure acquisition for USD. Common text structuring methods usually include manual tagging with metadata or other specific tags for further structuring. The Unstructured Information Management Architecture (UIMA) standard provides a general framework for processing this information to acquire semantics and create structured data.

The earliest Business Intelligence studies focused on unstructured text data,

not numerical data [7]. The development of Big Data in the late 2000s caused increased interest in the use of USD analysis.

Separating unstructured data analysis (UDA) into the separate scientific and technical area dates back to the early 2000s, when Gartner analysts published information about high time and labor involved in processing data because such routine, non-automated content processing took up half the working time.

Now data without formalized structure are the largest share of stored digitized information (more than 80% of all stored data, and their number is increasing by an order of magnitude faster than structured data), so methods and means of their use are evolving rapidly. These methods are intended to transform USD into structured information that can be used in various ways.

Adding a structure to USD is a complex scientific problem that has been paying attention to scientists for a long time [8]. In the most general form, the solution of the problem is related to the construction of a spaced graph corresponding to the USD content and to the comparison of such graphs generated for different USD sets. Another aspect of this problem is related to the finding of relevant knowledge for USD structuring tags and means of representation of such knowledge.

Text Mining can be defined as the process of acquiring knowledge from collections of USD documents using various toolkits for their analyzing [9]. Text Mining can be considered as a separate case of Data Mining. Processing of NL data should provide a transition from USD to structured ones with subsequent analysis. Most often, this process ignores most of the specific features of the NL, which are used only at the previous stage of parsing the texts, and in the following uses the "bag of words" model, in which the word order is not difficult. Similar to Data Mining, Text Mining tools seek to retrieve information relevant to user's activity. In the case of Text Mining such templates that can be applied for user tasks should be found not in formalized database records, but in non-structured text data in the documents of these collections.

An analysis of current research in this area demonstrates that the most effective approaches to the analysis of NL data are based on the use of the background knowledge of the corresponding domains formalized by ontology [10].

5 Ontologies of IS domain

Cyber security information needs to be machine-readable to enable efficient information exchange, retrieval and operation automation deal with validation of learning in this domain. Today, software development in the IS field is characterized by intellectualization, which allows to anticipate the threats that can lead to APT (advanced persistent threat) attacks, phishing and redemption software. Ontological representation of knowledge is widely used now in distributed intelligent application oriented on processing into the Web open environment to support knowledge sharing and reutilization. Ontologies provide an explicit specification of a conceptualization. Therefore creation of IS domain ontologies covered a set of concepts and categories of this domain, their various properties and relationships between them is of great interest to developers and users of modern IS means.

Ontologies in IS have now evolved into an active provider of data element

links that can use machine learning and artificial intelligence algorithms to adapt to changes in the environment [11]. Development of ontological model for entire IS domain requires integration of existing ontologies and their refinement.

The Unified Cyber Security Ontology (UCO) [12] is designed to support the knowledge integration into IS systems and should unify the most commonly used information security standards. This ontology integrates heterogeneous data and knowledge schemas from different IS subsystems and the most commonly used IS standards for data sharing. UCO can serve as the core of knowledge for the IS domain. The UCO ontology has also been mapped to a number of existing cyber security ontologies as well as concepts in the Linked Open Data cloud.

UCO is an ontology developed for integration of the unifying information from heterogeneous sources. This ontology supports reasoning and inferring new information from existing information. UCO also provides capturing specialized knowledge of a cyber security analyst expressed by ontology classes and individuals as well as rules. UCO ontology provides a common understanding of cyber security domain and unifies most commonly used cyber security standards.

The most important top-level IS concepts represented by UCO classes:

- Means: various methods of attack executing
- Consequences: possible outcomes of attacks.
- Attack: cyber threat attacks.
- Attacker: identifications and characterization of the adversary.
- AttackPattern: descriptions of common methods for exploiting software that provide the attackers perspective and guidance on ways to mitigate their effect.
- Exploit: descriptions of individual exploits.
- Exploit Target: weaknesses and vulnerabilities of software, systems, networks or configurations that are targeted for exploitation by the TTP (cyber threat adversary Tactic, Technique or Procedure).

UCO ontology serves as the core for cyber security Linked Open Data (LOD) cloud and integrates knowledge from some international sources but general ontological model of IS has take into account various national IS standards and laws.

A detailed description of knowledge about the IS domain requires the development of an ontology hierarchy, ranging from the top level to the bottom one. Top-level fundamental ontology includes basic IS domain concepts, and low-level ontologies describe the specific problem aspects with more exact detailing of features. An interesting example of multilevel ontological system of cyber security is proposed in [13]. This ontological framework CRATELO consists of three layers:

- top-level ontology designed on the SPRAY basis (simplified version of DOLCE top ontology),
- Security Core Ontology (SECCO) is a set of security concepts based on DOLCE-SPRAY primitives,
- security-related middle-level ontologies of secure operations (OSCO).

In [14] authors propose reference ontology for cyber security operational information. IS in this approach is considered as the preservation of information confidentiality, integrity and availability. This work takes into account a lot of ontologies of cyber security domain that formalize knowledge organizing of IS risk management, vulnerabilities, attackers, security metrics, countermeasures and other relevant concepts [15]. Many middle-level ontologies were created in the field for representing the

various individual aspects of IS domain. For example, researchers developed application ontologies to identify and classify network attacks:

- an ontology for distinguishing the state of network security [16];
- an ontology of intrusion detection [17];.
- an ontology for automated classification of network attacks [18].
- an ontology of prediction of potential network attacks [19].
-

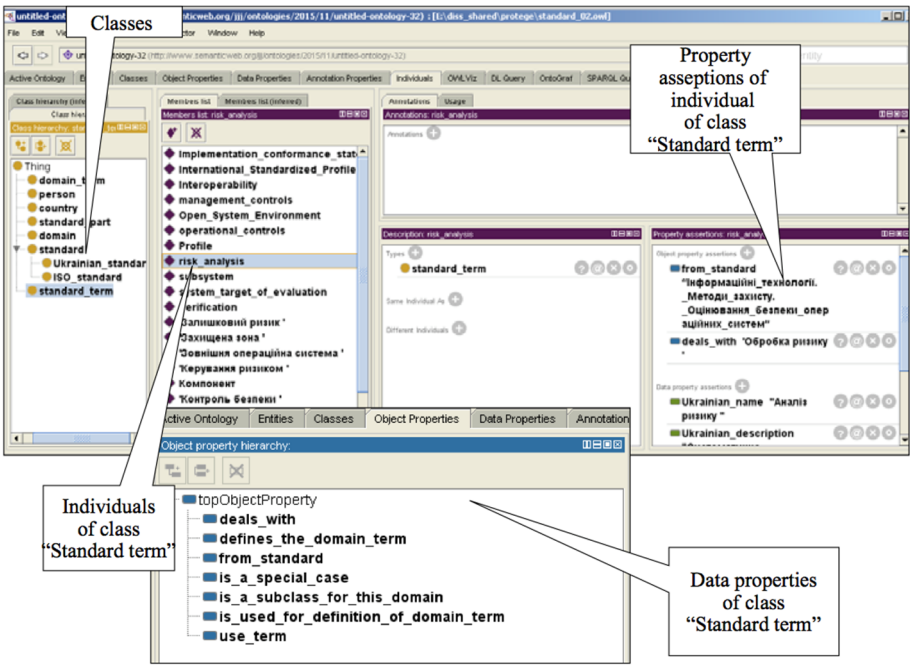

Fig. 1. Information security ontology (fragment)

Other IS ontologies provide and enrich an adaptive vocabulary that can improve behavioral analysis and help stop the spread of threats. As well terms for IS ontologies can be obtained from the open sources, such as the dictionary of cyber security terms [21] and the various standards of this domain. But processing of USD with NL text requires special software and expert participation. Acquisition of knowledge from NL documents that contain semantic markup (such as semantic Wiki resources) is easier. Wiki pages correspond to domain concepts, and links between them explicitly defined in content can be used to build ontological relations. For example, we use the "Security systems" category pages of the portal of the Great Ukrainian Encyclopedia (e-VUE) [22] and Ukrainian Wikipedia as information sources of IS terms (Fig.2).

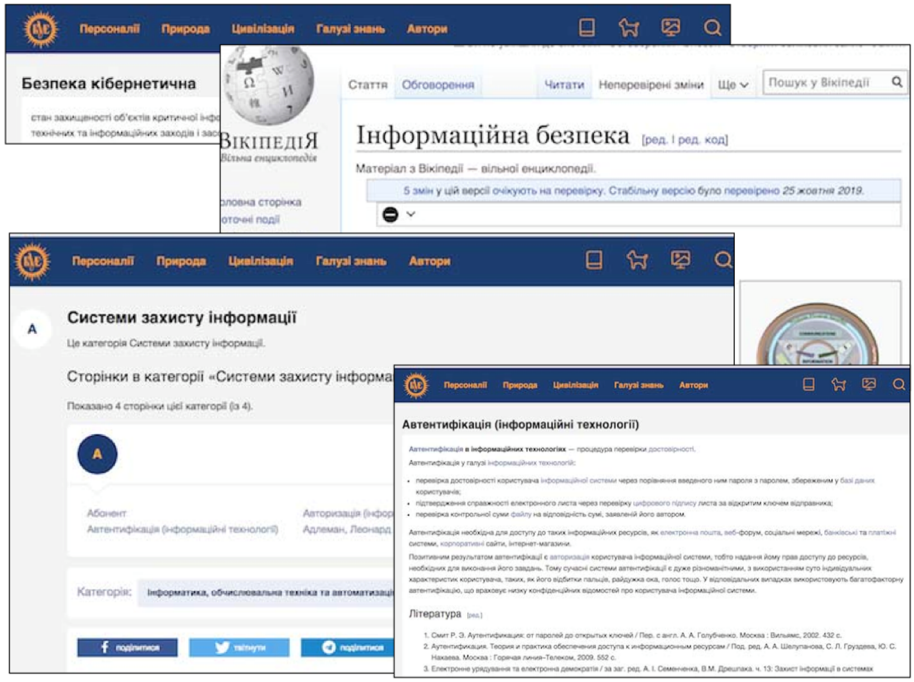

Fig. 2. Wiki pages with semantic markup on e-VUE and Wikipedia.

Wiki markup can be based on existing IS ontologies. It is advisable to create such Wiki pages for each IS standard in whole as well as for their separate subdivisions and definitions. Thus, the semantic markup of NL information resources enables the automated generation of IS ontologies.

6 Use of IS ontological models in validation of learning outcomes

The effective development of IS demands the determination of competence areas and their contents, so as to make it possible to define learning outcomes for each level and type of education [23]. Validation of learning outcomes can be realized by ontology-based matching of results of informal learning with vacancies and task in relevant domain.

Terms and relations of domain ontology provide the semantic matching of NL texts. At the level of ontological models, it is possible to correlate learning outcomes, for example, the competencies of IS professionals who have acquired their knowledge and skills in non-formal and informal learning. These people can describe their learning outcomes in the same terms used by employers to describe the jobs and tasks. The set of competencies is defined by domain specifics.

We proposed to divide each learning result into the unique set of atomic

learning results [24], and to establish the relation and hierarchy of learning outcomes and qualifications through an appropriate ontology. Atomic learning outcomes (atomic competencies) have the following properties:

- $a \in C$, where C is the set of information objects of the class "Learning Outcomes", and C_{atomic} - the set of atomic learning outcomes, $C_{atom} \subseteq C$;

- each learning outcome c is an unique non-empty set of atomic learning results $\forall c \in C \exists a_i \in C_{atomic}, i = \overline{1,n}, k = \bigcup_{i=1} a_i$;

- atomic learning outcomes are not a subset of other learning outcomes $\forall a,b \in C, a \subseteq b \Rightarrow b \notin C_{atomic}$.

In order to relate the learning results to the domain terminology we define for each $a \in C$ the terms $x_i \in X, i = \overline{1,n}, n \geq 1$ and relations $x_{i_a} = r_j(x_{i_b})$ from the domain ontology (in this case, from one of the ontologies included into hierarchy of IS ontological model). Learning outcomes without such ontological correspondences cannot used (if IS ontologies does not contain an appropriate terms for important learning outcomes then we have to expand IS ontology by appropriate terms).

Thus, the matching of individual learning outcomes with task requirements is executed on semantic level by comparison of finite sets of atomic learning outcomes defined in domain terms. Therefore, the time of comparison is linear proportional to the number of atomic learning outcomes, and each atomic learning outcome is a finite unique nonempty set of domain terms.

Use of such atomic learning outcomes requires in generation of domain thesauri that are simpler then ontologies but contain all domain terms deal with some specific task (for example, some job definition).

7 Generation of IS thesauri based on domain ontologies

Thesaurus approach is widely used for analysis of domain term systems. Domain models based on thesauri can be used to generate domain representations assisted by computers with the help of integrated combination of different kinds of techniques from computing, statistics and artificial intelligence [25]. Some methods of thesauri generation use domain ontologies as a knowledge source and integrate thesaurus with current task description.

In [26] researchers propose to use thesaurus approach to formalize IS domain terminology. IS thesaurus displays a wide range of essential properties, attributes and relationships that are inherent in this specific type of security. Another example of IS thesaurus is described in [27]. This thesaurus integrates the concepts of cybersecurity, data security, application security, network security, Web security, and critical information infrastructure security.

Application security is determined by the application software as well as by the software resources and processes involved in their lifecycle. Network security is related to the design, implementation and use of networks within the organization, between organizations, between organizations and users. Web security refers to Web services and related information and communication systems and networks. Security

of critical information infrastructure is characterized by protection against relevant threats, in particular information security.

Researchers [28] distinguish three groups of IS terms:

1. Terms defining the scientific basis of IS. This group includes terms used in many fields of knowledge that are unambiguous, semantically unified, and stylistically neutral. Examples: information, communication, conflict, influence, threat, danger, security, system. The concepts in this group meet the requirements of uniqueness and stability, i.e. they are uniquely applied in one area of knowledge and retain content in any other. This group of terms corresponds with the top-level ontologies.

2. Terms defining the subject basis of IS. This group of terms refers to concepts and their relation to other concepts within information security as a specific area of knowledge. This group of terms corresponds with the middle-level ontologies of IS.

3. Terms defining the nature of IS activities. This group covers terms denoting objects, phenomena, processes, their properties and relationships that are characteristic of this field. This group of terms corresponds with the low-level ontologies of special subspheres of IS.

IS thesaurus oriented on processing NL texts deal with learning outcomes in IS has to contain terms from all categories if they are used into the task description (here task is a NL description of work or problem that employer proposes to specialists with appropriate competencies).

We define task thesaurus as a special case of domain ontology oriented on analyses of NL texts. Thesaurus contains only ontological terms (classes and instances) but does not describe the semantics of relations between them. It can be automatically generated on base of the domain ontology and NL description of the problem [29].

A simple task thesaurus is a task thesaurus based on the terms of a single ontology. A composite task thesaurus is a task thesaurus that is based on the terms of two or more domain ontologies. It should be noted that IS domain is very dynamic, and therefore there is a problem in keeping the terminological base of the current state of research works and their integration with the modern grammar of the Ukrainian language.

8 An algorithm for constructing of task thesaurus

A simple task thesaurus is constructed according to the user-selected domain ontology and task – the description of the current problem (Fig. 3). This algorithm is described in detail in [30].

The task description can be represented via NL text that contains elements related to the ontology elements, or through conditions for domain terms relevant to the task. Thesaurus constructing consists from two main steps:

Step 1. Automated generation of a simple task thesaurus by task description;

Step 2. Expansion of a simple thesaurus by other ontology concepts by the set of conditions that use elements of the O ontology different from instances and classes.

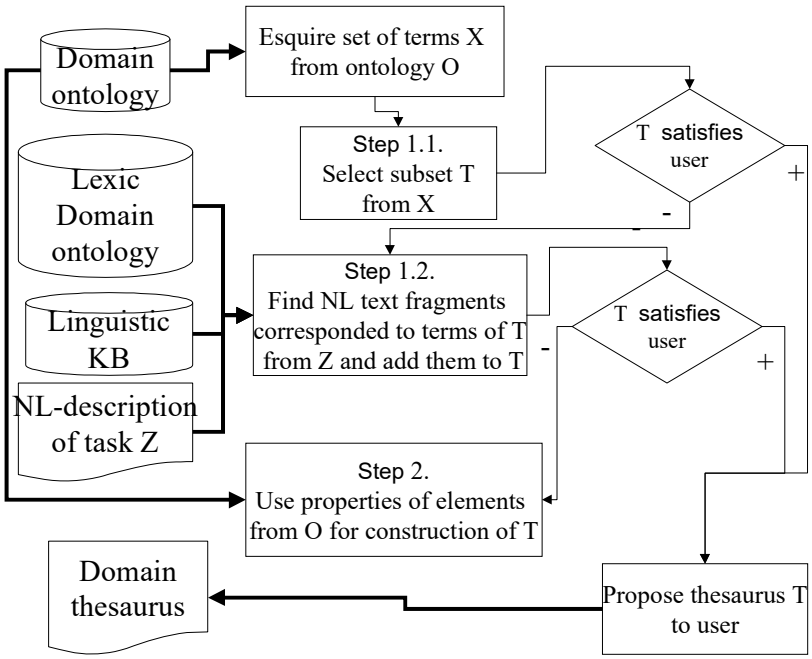

Fig. 3. Main steps for constructing of task thesaurus

We divide step 1 into two substeps. On Step 1.1 user explicitly and manually selects task-pertinent terms from the automatically generated list of classes and instance X. In the simplest cases, the construction of the thesaurus may be completed in this step, but it requires more efforts from the user.

Stage 1.2 uses a variety of methods for processing a natural-language task description (linguistic analysis, statistical processing, semantic markup analysis) that allow detecting NL fragments related to terms from O.

Those ontological terms that correspond to some fragments of task description are added to the simple task thesaurus.

Linguistic knowledge bases (KB) can be used to construct a thesaurus. We can use specific domain-oriented linguistic KBs if a large amount of lexical information has already accumulated. Such information is not universal and depends either from domain and natural language used in task definition.

Therefore we cannot use Text Mining systems – they often are oriented on processing only English texts. We apply direct updating of the domain lexical ontology by users and export linguistic knowledge from relevant vocabularies and knowledge bases, as well as from semantically marked Ukrainian texts.

In many cases, information about other elements of the ontology is appropriate in thesaurus constructing. This information allows taking into account the properties of individual terms and their relations with other terms.

Such actions are applied on step 2 for refining the initially formed thesaurus in accordance with explicitly formulated user conditions. These conditions depend from the specific nature of the task, but are not derived from its description. Such

conditions can be considered as a set of meta-rules to describe the information the user is seeking.

Stage 2 can be represented as a function that transforms the O ontology into simple task thesaurus according to proposed meta-rules deal with the finite set of conditions that the user formulates for classes and instances of domain ontology.

Complex task thesauri are generated from the built earlier task thesauri (simple or complex) with the help of set theory operations such as sum of sets, intersection of sets etc.

9 Use of task thesauri for validation of learning outcomes

Task thesauri can be used in various intelligent tasks. For example, such thesauruses can be very useful to validate learning outcomes that are represented by NL unstructured text. The main requirement of such approach is an ontological model of learning domain. As we describe above, complex and dynamic IS domain has a lot of ontological representations of its various aspects. Therefore we can apply this approach to problem of learning outcomes validation for IS.

Processing of task thesauri for learning outcomes validation consists of such main steps:

Step 1. Task thesaurus is generated by processing of NL job descriptions received from employers (vacancy description, set of requirements for a specific position or qualification etc.). Thesaurus consists of concepts used in this description.

Step 2. The pertinent set of IS ontologies is generated by retrieval in the Web and ontology repositories. Thesaurus concepts are used as search keyword.

Step 3. We retrieve in the NL descriptions of the learning outcomes not all IS domain concepts but only those that are included in the task thesaurus. This greatly speeds up the comparison because IS ontological model in general contains a great number of concepts and their NL representations.

Descriptions are sufficient for to matching procedure if they contain all the necessary concepts of ontology. It is important that thesauri allow us to group synonyms or words represented by different NL, thus matching all semantically similar terms.

Step 4. If any learning outcome completely matches with job description then we try to find descriptions of the learning outcomes that are most similar to the task description.

10 Acknowledgement

In this research work we use scientific results obtained from Information programs of the NASU «Development of dataware, functional support and software of electronic version of encyclopedic editions of Ukraine» project (2019) in Institute of Software Systems of NASU and from the work fulfilled in accordance with the thematic plan of scientific researches of the Dmytro Motornyi Tavria State Agrotechnological University (project of applied research at the expense of the state budget expenditures «The-

oretical substantiation and development of information system of semantic identification, documentation and processing of results of non-formal and informal learning»).

11 Conclusions

IS industry is a dynamic field that is rapidly changing and improving specific methods and tools. Therefore, cybersecurity-related specialists require actual knowledge arrival, both from internal and external sources. This necessity enhances the non-formal and informal learning importance in this field for the effective accomplishment of various tasks.

The complex hierarchical structure of IS domain knowledge necessitate the ontological analysis use to the data processing. IS domain ontological model contains knowledge conceptualization, integrates heterogeneous data structures and knowledge from different IS subsystems, as well as national and international standards. Ontologies provide formalization of the relations between domain elements and support their automated processing.

Unstructured and large-volume information causes the application of Big Data analytic techniques in preliminary processing of data sources used for generation of IS ontological model. Then this model can be used for Big Data metadata matching with various task areas.

The authors propose to use task thesauri based on appropriate ontologies as a tool for matching the informalized results of IS specialists learning outcomes with task descriptions. These thesauri provide a terminological basis for the integration of IS ontologies of various abstraction levels. Thesauri use reduces the comparing computational complexity of heterogeneous knowledge structures.

References

1. Parviainen, P., Tihinen, M., Kääriäinen, J., Teppola, S.: Tackling the digitalization challenge: How to benefit from digitization in practice. In: International Journal of Information Systems and Project Management, 5 (1), pp.63-77.(2017).
2. Grimes S.: Unstructured Data and the 80 Percent Rule, Clarabridge, Bridgepoints, (2008), http://breakthroughanalysis.com/2008/08/01 / unstructured-data-and-the -80-percent-rule /.
3. Savas O., Deng J.: Big Data Analytics in Cybersecurity In: CRC Press, .- 353p. (2018).
4. Zhang, Y., Ren, J., Liu, J., Xu, C., Guo, H., Liu, Y.: A survey on emerging computing paradigms for big data. In: Chinese Journal of Electronics, 26 (1), pp.1-12, (2017).
5. Demchenko, Y., De Laat, C., Membrey, P. : Defining architecture components of the Big Data Ecosystem. In: 2014 International Conference on Collaboration Technologies and Systems (CTS), pp. 104-112, (2014).
6. Unstructured_data. - https://en.wikipedia.org/ wiki / Unstructured_data.
7. Grimes S.: A Brief History of Text Analytics. B Eye Network, (2016), http://www.b-eye-network.com/view/6311.
8. Buneman P., Davidson S., Fernandez M., Suciu D. Adding structure to unstructured data. In: International Conference on Database Theory, pp. 336-350, (1997).

9. Feldman R., Sanger, J.: The text mining handbook: advanced approaches in analyzing unstructured data. In: Cambridge university press, (2007), https://wtlab.um.ac.ir/images/e-library/text_mining/The%20Text%20Mining%20HandBook.pdf.
10. Rogushina J.: Means and methods of unstructured data analysis. In: Problems in Programming, N 1, pp.57-77, (2019), http://pp.isofts.kiev.ua/ojs1/article/view/348/346. (in Ukrainian).
11. Obrst, L., Chase, P., Markeloff, R.: Developing an Ontology of the Cyber Security Domain. In: STIDS, pp. 49-56, (2012).
12. Syed, Z., Padia, A., Finin, T., Mathews, L., Joshi, A.: UCO: A unified cybersecurity ontology. In: The Workshops at the Thirtieth AAAI Conference on Artificial Intelligence. (2016), https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/download/12574/12365.
13. Oltramari, A., Cranor, L. F., Walls, R. J., McDaniel, P. D.: Building an Ontology of Cyber Security. In: STIDS, pp. 54-61, (2014), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.664.3593&rep=rep1&type=pdf.
14. Takahashi, T., & Kadobayashi, Y.: Reference ontology for cybersecurity operational information. In: The Computer Journal, 58(10), pp.2297-2312, (2015), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8205615
15. Wang, J.A. and Guo, M.: OVM: An Ontology for Vulnerability Management. In: Proc. 5th Annual Workshop on Cyber Security and Information Intelligence Research, pp. 1-4, (2009).
16. Bhandari, P. and Guiral, M.S.: Ontology Based Approach for Perception of Network Security State. In: Recent Advances in Engineering and Computational Sciences (RAECS), pp.1-6, (2014).
17. Ye, D., Bai, Q., Zhang, M., Ye, Z.: P2P distributed intrusion detections by using mobile agents. In: Seventh IEEE/ACIS International Conference on Computer and Information Science (ICIS 08), pp.259-265, (2008).
18. van Heerden, R., Leenen, L., Irwin, B.: Automated classification of computer network attacks. In: International Conference on Adaptive Science and Technology (ICAST 2013), pp.157-163, (2013).
19. Salahi, A. Ansarinia, M.: Predicting Network Attacks Using Ontology-Driven Inference, (2011), http://arxiv.org/ftp/arxiv/papers/1304/1304.0913.pdf.
20. Motik, B., Patel-Schneider, P. F., Parsia, B., Bock, C., Fokoue, A., Haase, P., Smith, M.: OWL 2 web ontology language: Structural specification and functional-style syntax. In: W3C recommendation, (2009), www.w3.org/2007/OWL/draft/ED-owl2-syntax-20090914/
21. Gladun A.Y., Puchkov O.O., Subach I.Y., Khala K.O.: English-Ukrainian Dictionary of Information Technology and Cybersecurity Terms. In: Igor Sikorsky KPI, P. 376, (2018).
22. Rogushina J.: Processing of Wiki Resource Semantics on the Base of Ontological Analysis. In: Proc.of OSTIS-2018, pp.159-162, (2018), https://libeldoc.bsuir.by/ bitstream / 123456789/30389/1 / Rogushina_Processing.PDF.
23. Lehto, M.: Cyber security competencies: cyber security education and research in Finnish universities. In: ECCWS2015-Proceedings of the 14th European Conference on Cyber Warfare & Security: ECCWS 2015, pp. 179-88, (2015).
24. Rogushina J., Priyma S.: Use of competence ontological model for matching qualifications. In: Chemistry: Bulgarian Journal of Science Education, Volume 26, Number 2, pp.216- 228, (2017).
25. Lloréns, J., Velasco, M., de Amescua, A., Moreiro, J. A., Martínez, V.: Automatic generation of domain representations using thesaurus structures. In: Journal of the American Society for Information Science and Technology, 55(10), pp.846-858, (2004).

26. Sachuk Yu.E.: Training of cybersecurity and information security professionals: thesaurus and ontology. In: Problems of Engineering and Pedagogical Education, N 59, pp.35-40, (2018).
27. Kalichensky A.: The Concept of Creating the National Information and Communication Infrastructure of Ukraine. In: Regional Forum MSE (2012) https://www.itu.int/ITU-D/tech /events/2012/Spectrum_CIS_Kiev_Sept12/Presentations/Session2/A_Kalichensky _a.pdf.
28. Diorditsa I.V.: The presentation of the terminology of cybersecurity policy in the texts of regulatory acts of Ukraine. In: Scientific Bulletin of the International Humanities University. Jurisprudence Series, N 29 (1), pp. 64-67, (2017).
29. Gladun A., Rogushina J.: Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources. In: International Journal of Intelligent Systems and Applications, №1, P.11-20, (2012), http://www.mecs-press.org/ijisa/ijisa-v4-n1/IJISA-V4-N1-2.pdf.
30. Rogushina Yu.V. Use of Thesauri for Search of Complex Information Objects in the Web on Base of Ontologies. In: Problems in Programming, No. 4, pp.11-27. (2019) (in Ukrainian).