Italian Domain-specific Thesaurus as a Means of Semantic Control for Cybersecurity Terminology

Claudia Lanza
Department of Computer Engineering, Modelling, Electronics and Systems Engineering (DIMES)
University of Calabria
Arcavacata di Rende, Italy
e-mail: c.lanza@dimes.unical.it

***Abstract*— This paper presents an ongoing PhD research project aimed at realizing a tool for semantic control - an Italian thesaurus - that could represent a repository of the Cybersecurity field of knowledge and a means starting from which the representativeness of this domain can be enhanced by increasing the terminological coverage threshold. The paper starts with a description of the methodology followed by the creation of an authoritative corpus. This latter is meant to be the source of the information retrieval for the terms that should be inserted in the Italian controlled vocabulary of Cybersecurity. Afterwards, an overall summary of the semi-automatic terminological extraction will be provided. The paper focuses on the terminological process of mapping the selected terms from the authoritative corpus to the existent standards of Information and Communications Technology Security glossaries and vocabularies by using Python scripts. The paper also focuses on the perspective of how the relationships built in a thesaurus could be migrated to an ontology as a better form of knowledge representation**.

Keywords-Cybersecurity; knowledge representation; information retrieval; ontologies; thesauri.

I. INTRODUCTION

The underlying idea of this PhD research project is to develop a model that is meant to guarantee the terminological coverage of a semantic resource, such as a thesaurus, and its representativeness threshold with reference to semantic variation during time. By building an Italian thesaurus related to the Cybersecurity domain, this project relies on the perspective of providing organizations with a complete knowledge representation of the field of study on Information and Communications Technology (ICT) security. The thesaurus can represent a valid support tool for information access, treatment of data and information retrieval tasks in order to improve the security decision making processes. This research project is included in one of the activities carried out in collaboration with the Informatics and Telematics Institute (IIT) [26] – National Research Council (CNR) institute located in Pisa.

This paper analyses the steps needed to construct the thesaurus related to this particular domain beginning with the selection of the sources, which have been taken into consideration in order to have an authoritative corpus from which the information of the domain can be retrieved. The goal of the research project is, therefore, to provide a solid tool in which the information on Cybersecurity could help in reaching as complete a terminological coverage as possible. To reach this latter perspective, the project aims at enhancing the terminological set of data by taking into account not only legislative documents but also social media infrastructures and, doing so, achieves a heterogeneous information repository.

The main intention of the research project is to create an Italian thesaurus on Cybersecurity, currently not existing, that can help organizations to better frame the information on Cybersecurity and provide a terminological means of support that could be useful to broadly understand the domain from a semantic point of view. Even though there are taxonomies and glossaries on Cybersecurity in Italian language, such as, for example [15], a thesaurus can be considered a service that could give a more detailed overview on this domain thanks to the possibility of creating relationships between the terms that are meant to be representative of this area of study. The semantic tangle that comes out by the creation of a thesaurus is a starting point that can facilitate the process of migrating the knowledge organization within it into an ontology system.

The purpose of this paper is threefold: (1) providing an overall presentation of how to build a semantic tool, such as a thesaurus, as a means of semantic control for a specific domain by describing the steps which characterize the corpus creation and the terminological extraction; (2) presenting a model of mapping the existent standards on Cybersecurity to all the head terms contained in the initial corpus through Python scripts in order to evaluate which candidate terms should be chosen to be part of the thesaurus; (3) opening up the perspective of migrating the terms and their relationships of the Italian thesaurus on Cybersecurity in an ontology system.

The paper is structured into five sections. Section II contains a brief presentation of the state of the art consulted for the creation of the Italian thesaurus on Cybersecurity. The studies taken into consideration in Section II specifically refer to strategies able to establish the terminological coverage threshold of a semantic resource. Section III goes deeper into the methodological approach undertaken towards the realization of the semantic means of control on this particular domain. It describes the phases related to the information retrieval, starting with the authoritative set of documents that make up the source corpus. Section IV describes the way in which the terminological extraction has been executed from these texts by using the Text to Knowledge (T2K software). Section V

 ISBN: 978-1-61208-678-1

outlines the methods employed to select the head-based terms with a particular overview on a mapping system between the glossary derived from the terminological extraction and the major standards vocabularies on the Cybersecurity domain, i.e., this section ends with a description of a Python script used to automatize the process of aligning the terms contained in the standards and the terms obtained by the terminological extraction. Finally, Section VI concludes the paper with the proposal of converting the semantic structure of the thesaurus into an ontology system by migrating not only the terms, but also the relationships.

II. STATE OF THE ART

Many studies on the issue of evaluating the qualitative strength of a thesaurus have been carried out, like [23], which gives practical suggestions on how to create a reliable semantic resource. Other works have focused their attention on the importance of having a group of people with a high expertise in the domain that has to be analysed for the purposes of realizing a semantic tool for information retrieval. For example, [16] deals with the advantage of getting helped by domain experts who can increase the value of a thesaurus especially for what concerns the selection of terms and their relationships. Another study that is worth mentioning for its important description of a way by which corpus representativeness can be measured is [22]. The authors provide statistical formulas to calculate the size threshold of corpora in terms of linguistic coverage of a particular domain of interest. To calculate the ideal situation in which a semantic tool like a thesaurus can continue to be representative with respect to a particular domain, the terminological update within it is an unavoidable aspect to take into account, and [17] addresses this issue in greater detail.

One of the difficulties faced to set up the construction of the Italian controlled vocabulary, that aims at including as much information about Cybersecurity as possible, is that the available Italian sources on this subject seem to be quite few. A list of some Italian resources that have been used to extract information is given in Section III, *Corpus Creation*. The challenge, therefore, is to map the English concepts inside various terminological repositories on the Cybersecurity field to the Italian language and, by doing so, to align the description of these terms with the Italian law systems and ICT shared knowledge. There are many studies which have been carried out to develop reliable terminological sources that could guide the understanding of the Cybersecurity domain and help with the information retrieval on this subject for the creation of the Italian Cybersecurity thesaurus. Among these aforementioned studies, [8] contains one well-defined example of a helpful ontology on this subject displaying various ways of correlating Cybersecurity concepts in a semantic network process. [9] also gives light to some of the most common terms used in Cybersecurity, and [10] collects the terminology referred to the cyber-threats accompanied by the descriptions of all the terms of this repository. The latter has been helpful in placing the relationships inside the Italian thesaurus. Other important repositories that contributed to the population of terms referred to Cybersecurity vulnerabilities and threats or attacks, are the ones collected in the MITRE Corporation systems [13][14].

III. METHODOLOGICAL APPROACH

The first step of the project is based on the retrieval of the terminological authoritative sources that are going to be the documentary corpus of the domain that has to be analysed. Subsequently, a terminological extraction from the selected documents is carried out through specific programs that are meant to execute text-processing tasks. Next, the paper describes the steps according to which a thesaurus is going to be realized starting from a validation of terms by experts belonging to the domain of interest. With the output given by the terminological extraction, that is the controlled terminological list of terms sorted by the Term Frequency–Inverse Document Frequency (TF/IDF) measurement, a process of selection of the head-based terms begins by mapping the terms of the source corpus with the existent standards of ICT Security. Head-based terms are single terms that appear with highest frequency in the texts and that bring together other terms with which they are frequently accompanied, i.e., if “access” appears accompanied with many other lexical units, thus it will be the head-based term that will help in positioning its qualifiers and other terms with which occurs in the thesaurus.

For a better understanding of the system that has driven towards construction of a semantic means of control on Cybersecurity, the following Unified Modelling Language (UML) activity diagram [11] depicts the steps taken for the realization of the Italian thesaurus on this field of knowledge (Fig. 1).

 ISBN: 978-1-61208-678-1

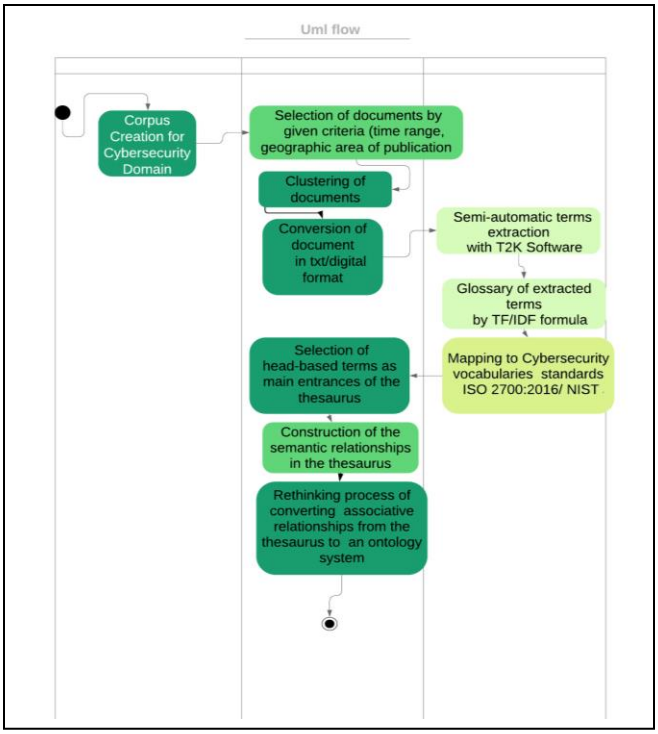

Figure. 1 Process for thesaurus construction.

The first phase of the ongoing project has dealt with the identification, the retrieval and the analysis of the existing Italian authoritative terminological sources referring to the domain of Cybersecurity. The material collected represents the reference context of ICT security and contains, as well, an important dataset both from a quantitative and qualitative point of view.

In terms of the quantitative evaluation of the documents which a corpus is supposed to contain, an established estimate does not exist in the literature. However, it is recommended to have a sizeable set of data that belong to the specific domain of Cybersecurity and consequently have a terminological coverage that could make its representativeness strong enough; a highly reliable dataset should also be present from a qualitative point of view so that the construction of a thesaurus could be as accurate as possible. The sources consulted must be considered authoritative in order for the thesaurus to become a guide that pilots the correct management of the sector-specific language. This process of gathering authoritative sources in order to make the semantic tool wide ranging is based on the principle of the hierarchy of the sources that in law considers three levels: constitutional sources (Constitution, constitutional laws and constitutional revision); legislative sources, also called primary sources (laws, decree law and legislative decree, regional laws) [7]; regulatory laws, also called secondary sources (Government Laws, local authoritative) – books, magazines and specialized articles, field specialized user profiles (experts of Cybersecurity) or social media users. One of the future prospects that will be considered for the development of this proposal is the attention to the social media world, referring to the wisdom of the crowds [2] according to which the implementation of new terms that, over time, become much more common in media jargon, should be considered in detail to understand the necessity of inserting them into the controlled vocabulary.

It goes without saying that this heterogeneity of documents that marks the source corpus from which the information about the domain is going to be studied is a remarkable advantage to reach a higher level of representativeness threshold. To give an example, in the source corpus, various documents deriving from different format and typologies have been taken into consideration, among these:

- Decree Laws;
- Parliamentary legislation;
- Penal/Administrative/Civil Code;
- Rules (GDPR);
- CERT guidelines;
- Government documents;
- Magazines that deal with the domain topics and could give another terminological output to enhance the thesaurus coverage (i.e., “Gnosis”, “Hacker Journal”).
- Glossaries (i.e., “Intelligence Glossary” [15])

Having to do with a domain under development, the hope is to run with the terminological evolution within the corpora so as to test through time the structure of the persistent value of the semantic relationships inside the thesaurus with the emergence of new terms and with the updating of the existing ones.

The constructed corpus will be the starting point from where a thesaurus will be realized and it will be characterized by a flexible set of documents that are going to vary over the years and that will be a very important aspect in order to enhance the terminological coverage level of a given domain of study.

The representativeness of a thesaurus is the key to determine its authoritativeness with respect to certain domains and geographic positioning areas. After completing the construction of the thesaurus structure, reaching what is considered to be the “gold standard” is the first purpose that the research activity aims at. It is unquestionable that this purpose could be reached only after the candidate terms have passed through a validation process by specific field experts. The latter, thanks to their level of authoritativeness in the areas of competence, represents a key step that cannot be avoided to have the approval of terms and of their semantic associations in the systems of knowledge management and organization. The international standards of reference, such as the 25964-1 of 2011 [3] and ISO 2564-2 of 2013 [4] regulations, will be followed: they will provide a standardization of the terms contained in the thesaurus that can guarantee the interoperability between various systems of knowledge management.

IV. TERMINOLOGICAL EXTRACTION

The terminological extraction is carried out once the corpus has been defined by the selection of documents which come from the authoritative legislative sources and

 ISBN: 978-1-61208-678-1

informative channels, such as the official magazines that contain information about the domain taken into account.

Before beginning with a semi-automatic processing of the information contained in the source corpus, the digital native documents, downloaded from the websites of the authoritative sources or Web portals, have been converted into txt format, which is the format required by the textual analysis software.

Native paper documents have been firstly scanned and saved as PDF files and have then undergone an optical character recognition process and finally transformed into txt documents.

Among the pieces of software that have been chosen for the terminological extraction, the Text To Knowledge (T2K) [25] tool has been preferred for the purposes of detecting, in further analysis, head terms that can become part of the controlled vocabulary.

T2K is a software developed by the Institute for Computational Linguistics (ILC – CNR) in Pisa (Italy) by a group of computational linguists and it is a powerful Natural Language Processing (NLP) tool that can provide, through semi-automatic text processing tasks, different forms of wordlists which include candidate terms used to populate the thesaurus. T2K allows to extrapolate the most relevant terms of the corpus on Cloud systems or by virtualization techniques, according to variables, such as accuracy – obtained by the algorithm ULISSE [18] – occurrence, frequency and disambiguation.

The first steps of this project have dealt with the lexical configuration in T2K of the desired semantic chains as output from the documents that make up the imported corpus. What has been obtained by the semi-automatic processing of T2K is a glossary of terms that refer to the domain of study which is meant to be used for the semantic analysis of the candidate terms derived from this set of authoritative digital and paper sources. The statistical measurement that characterized the basis of the terminological representation of the candidate terms in the T2K wordlist is the TF/IDF formula. According to this statistical measurement, the lexical units with the highest frequency are considered less important – such as, adverbs, or articles and prepositions, in brief the stopwords – and the output underlines the words that occur in the documents that in general are less frequent.

Only after having set the semantic configuration of the lexical units of the output given by the NLP text processing, T2K provides different ways to visualize the terms that are meant to be part of the controlled vocabulary. One of these, the so-called Broader Term (BT/NT), that corresponds to the ISO [3][4] tag standard, has been selected in order to help with the decision of inserting preferred candidate terms in the thesaurus. For a better understanding of the results of terms from the authoritative corpus, the following figures show the table given by the extraction (Fig. 2), and the knowledge graph derived from the occurrences of these terms with others inside the corpus (Fig 3). Both of these processes have been developed by using the semi-automatic terminological extraction of T2K NLP Tool. The knowledge graph is a representation of terms that are connected together, and this is the way followed for the selection of the most pertinent head-based terms.

Prototypical form	Lemma of Term	Frequency
Network	network	74
Information	information	69
Fundamental Glossary	fundamental Glossary	36
System	system	47
Access	access	45
Risk	risk	43
Computer	computer	43

Figure. 2 Example of terms extracted with T2K.

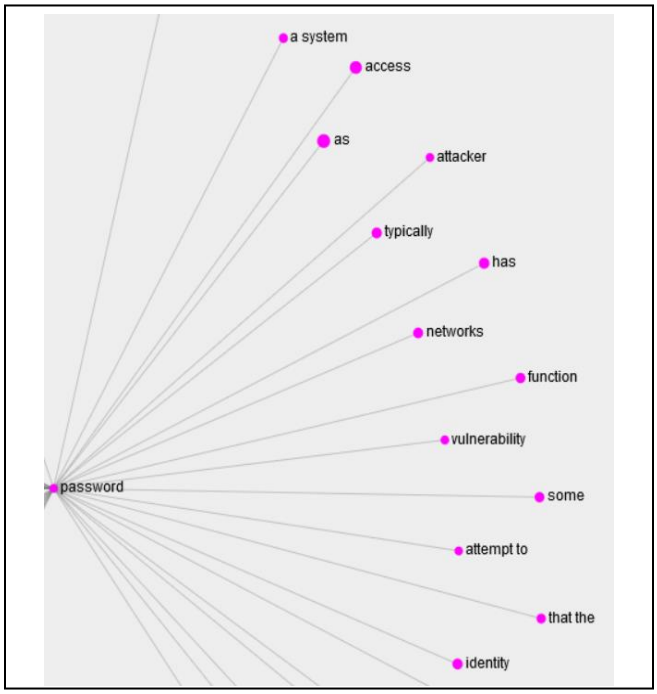

Figure. 3 Knowledge Graph in T2K.

V. HEAD-BASED TERMS

A semantic tool, such as a thesaurus, should represent a reliable source of information that can support the operations of information retrieval and the access to documents related to a specific domain of study. For this reason, the first phase that follows up on the creation of a corpus and the terminological extraction from these authoritative documents is strictly connected to the selection of which candidate terms should be considered as preferred terms that are to be imported in the controlled vocabulary and starting from

which the basic relationships proper of a thesaurus can be inserted.

Typically, in a thesaurus, the classical kinds of relationships that characterize the network of connections between terms are of three types, as [1][3][4] better explain:

1. Equivalent relationship characterizes the synonymy or quasi-synonymy between different terms; in a thesaurus there is one term that is going to be considered as the preferred one to which the other kind of relationships can be developed and it is marked by a tag USE, and its synonym, or the other way by which it can be seen in the domain documents, is marked by the tag UF that stands for *Used For*, e.g., VAPT UF Vulnerability Assessment and Penetration Testing;
2. Hierarchical relationship is marked by two tags: BT that stands for *Broader Term* and represents the more general concept with reference to its more specific one which is, on the other hand, marked as NT, i.e., *Narrower Term*, e.g., Cyber Attacks NT Brute force attacks, or Cyber Threats NT Hacker;
3. Associative relationship defined by the standard tag RT that stands for *Related Term*: it represents a concept that is associated to another one, e.g., Hacker RT Cyber criminality, or Spam RT Virus; as [3] stated, "the associative relationship covers associations between pairs of concepts that are not related hierarchically, but are semantically or conceptually associated to such an extent that the link between them needs to be made explicit in the thesaurus, on the grounds that it may suggest additional or alternative terms for use in indexing or retrieval. The relationship is indicated by the tag 'RT' (related term) and it should be applied reciprocally".

In order to make the Italian thesaurus on Cybersecurity a highly representative source, this paper describes a project phase that has covered the retrieval of terms contained in the NIST Glossary of Key Information Security Terms 7298 [5] and in the ISO-IEC 27000/2016 [6] standard. The purpose was that of checking if those terminological assets were present in the source corpus that has created the basis for the terminological extraction.

Since the source corpus is made up of documents written in Italian language, in order to create a semantic tool for Cybersecurity in Italian, which does not currently exist, the contrastive analysis with the glossaries in the aforementioned standards is useful for three main reasons: (1) because mapping the standard terms can prove, by verifying their presence in the source corpus, if the latter, built for the construction of an Italian thesaurus of Cybersecurity, can be conveyed as a reliable resource; (2) because matching them with the BT/NT structure of the terminological controlled list obtained through text processing software operations is a way of detecting which can be the preferred terms in the thesaurus; (3) starting from the scope notes contained in these standards, the network of the relationships that will characterize the thesaurus can comply as much as possible with the domain language usage.

The steps undertaken to analyse the standard terms list with the terms contained in the source corpus are the following:

1. After having downloaded the NIST 7298 [5] and ISO-IEC 27000/2016 [6], they have been semi-manually translated into Italian language in order to better suit the purposes of the project research of creating an Italian resource for Cybersecurity terminology retrieval; the tool that has been employed to proceed with the translation of the terms present in the ICT standards and their definitions, is TRADOS [12]. This service provides a memory repository that can catch all the translations that have been made on a document which can be used whenever, in another document, there is a term that can be translated exactly the same as in previous texts; this is highly useful in order to achieve coherence in the translation process of many texts;
2. The list of terms contained in the aforementioned standards have been assessed by a group of experts who have played an essential role in the validation of the authoritative source terms to start the cross mapping in the source corpus;
3. A Python script has been realized in order to check if all the words present in the standards were also included in the list obtained by the terminological extraction under the hierarchical form BT/NT. This script takes into account the terminological list from the whole corpus and, through a reading of its lines, the content of the file becomes analysable. Subsequently, an automatic generation of an output file text gives a list of all the occurrences (terms contained in the standards that are present in the controlled list) all at once, facilitating a screening process of the mapping system;
4. Once verified the presence or the absence of determined terms that are inserted in the standards taken into consideration, a process of selection among the head-based terms resulted from the extraction with T2K has been started with the help of the domain experts. Collaboration with the domain experts continues in order to face the challenge of deciding which terms can be considered as the best head terms that can connect the others inside the list and the standards;
5. After having chosen which terms could be considered as the preferred entries in the future Italian thesaurus, the process of building the network of the relationships has begun starting from the term definitions in the standards, and that helped in positioning the terms connected with each other in the thesaurus;
6. A draft prototype of an Italian thesaurus for Cybersecurity has been realized equipped with the Scope Notes derived from the definitions in the standards.

VI. CONCLUSION

This paper aimed at presenting an ongoing work based on a PhD research project that refers to the construction of an Italian Cybersecurity thesaurus, whose terminological coverage is going to be semi-automatically enhanced.

The goal set out in this PhD path is that of migrating all the relationships that will be created in the thesaurus into relationships inside a different structure, i.e., an ontology system.

A large number of studies on the possibility of reengineering a thesaurus into an ontology have proved that

 ISBN: 978-1-61208-678-1

this conversion is possible through the migration of the typologies of relationships in the basic ones of the ontology. In [24], the authors have developed a methodology able to convert the basic relationships proper of the thesaurus to Web Ontology Language (OWL) language. In [19], for the AGROVOC domain, the study converged on a set of relationships the replacement of those used in a thesaurus making them more specific. Indeed, the objective behind the need of converting the thesaurus into an ontology is based mainly on the principle that the latter provides a deeper knowledge representation.

One of the most evident differences that occur between a thesaurus and an ontology is that the associative relationship in the former, RT, can be clarified by customized form of related connections which can be themselves split in different hierarchical and more specific subclasses of relationships. As [21] demonstrates with their experiment of migrating the MeSH thesaurus with OWL language, one of the limits of a thesaurus is that of providing a flattened base of knowledge in terms of RT connections among terms. Although a thesaurus can be a reliable form of domain-specific information retrieval, and can create a dense net of connections which can generate a cross-reference system able to gather all the terminology in a determined field of study, the associative relationship does not suffice to formalize the conceptual links between terms. For this reason, also [20] offers a wide perspective on this issue with its Hasti project, an ontology is able to detect a more specialized kind of relationship customizing, by Universal Resource Identifier (URI) names, the typology of every one of these relationships that have to be different in order to handle a deeper form of knowledge organization of the domain that has to be represented.

The purpose of the conversion from a thesaurus to an ontology is to make information readable through languages that are proper to the ontologies: Resource Description Framework (RDF), Resource Description Framework Schema (RDFs), OWL. The transposition of the concepts represented through an abstract level by terms into an ontology should work to provide a better semantic representation system of the network of relationships. Thanks to OWL, it is possible to have extra semantic properties, such as the disjunction between two sets, so a complex of concepts can be separately analysed, the functional association of a property to a class and consequently the establishment of unique identifiers; the transitive property between classes useful for the creation of a much more articulate and descriptive semantic network. The efficiency of OWL relies on its formalism of language and in the possibility of applying automatic reasoning systems and developing inference on the described knowledge.

Even converting terms inside the controlled vocabulary into a class, entity and property scheme that belongs to a conceptual ontological modelling, the presence of a group of experts who will validate the associative network obtained by the reconstruction of the thesaurus in an ontology will be necessary.

ACKNOWLEDGMENT

This work has received the support of the [26] at the National Research Council in Pisa, especially by the emerging Cybersecurity group.

REFERENCES

[1] W. Broughton, “Building thesauri”, Milan, 2006.

[2] J. Surowiecki, Wisdom of crowds, Anchor Books, 2004.

[3] International Standard ISO 25964-1, “Information and documentation — Thesauri and interoperability with other vocabularies — Part 1: Thesauri for information retrieval”, First edition 2011-08-15.

[4] International Standard ISO 25964-2, “Information and documentation — Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies”, First edition 2013-03-15.

[5] R. Kisserl, NISTIR 7298 Revision 2 “Glossary of Key Information Security Terms” National Institute of Standards and Technology Interagency or Internal Report 7298r2, May 2013.

[6] International Standard ISO/IEC 27000:2016 (E) Information technology – Security techniques – Information security management systems – Overview and vocabulary, Fourth edition 2016-02-05.

[7] V. Crisafulli , “Hierarchy and competence in the constitutional system of law sources” in “Studies in the memory of Guido Zanobini”. Milan: Giuffrè, 1965, vol. III, p. 183.; G. Zagrebelsky, The constitutional system of law sources. Turin: EGES 1984, p. 67;

[8] Z. Syed, A. Padia, T. Finin, L. Mathews and A. Joshi, “UCO: Unified Cybersecurity Ontology, AAAI Workshop on Artificial Intelligence for Cyber Security”, February 2016.

[9] NICCS National Initiative for Cybersecurity Careers and Studies – Glossary, https://niccs.us-cert.gov/about-niccs/glossary [accessed October 2018] .

[10] Sophos “Threatsaurus The A-Z of Computer and data security threats”, https://www.sophos.com/en-us/medialibrary/PDFs/other/sophosthreatsaurusaz.pdf?la=en [accessed September 2018]

[11] Argo Uml, https://argouml.it.uptodown.com/windows [accessed October 2018]

[12] SDL Trados Studio, https://www.sdltrados.com/it/ [accessed October 2018]

[13] “The Common Vulnerabilities and Exposures (CVE) Initiative”, MITRE Corporation, https://cve.mitre.org/ [accessed September 2018]

[14] “The Common Attack Pattern Enumeration and Classification (CAPEC) Initiative”, MITRE Corporation, https://capec.mitre.org/ [accessed September 2018]

[15] Presidenza del Consiglio dei Ministri – Sistema di informazione per la sicurezza della Repubblica, “Il linguaggio degli Organismi Informativi”, Glossario Intelligence, https://www.sicurezzanazionale.gov.it/sisr.nsf/quaderni-di-intelligence/glossario-intelligence.html. [accessed October 2018]

[16] Claire K. Shultz, Wallace L. Schultz, and Richard H. Orr, “Evaluation of Indexing by Group Consensum” (Final Report, Contract No OEC 1-7-070622-3890), Bureau of Research Office of Education, U.S. Department of Health, Education and Welfare, August 30, 1968, 40 pp.

[17] A. Kennedy and S. Szpakowicz, “Evaluation of automatic updates of Roget's Thesaurus” J. Language Modelling, 2014, volume 2, pp.1-49 .

[18] F. Dell’Orletta, G. Venturi, and S. Montemagni, “ULISSE: An Unsupervised Algorithm for Detecting Reliable Dependency Parser”, Conference: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, 2011, pp. 115-124.

[19] D. Sorgel et al., “Reeingeenering Thesauri for New Applications: the AGROVOC Example”, J. Dig. Inf. 4(4), 2004, pp. 1-19.

 ISBN: 978-1-61208-678-1

[20] M. Shamsfard and A. Abdollahzadeh Barforoush, “Learning Ontologies from Natural Language Texts” International Journal of Human-Computer Studies archive Volume 60 Issue 1, January 2004

Article in a conference proceedings:

[21] L.F. Soualmia, C. Golbreich, and S.J. Darmoni, “Representing the MeSH in OWL: Towards a Semi-Automatic Migration” Conference: KR-MED 2004, First International Workshop on Formal Biomedical Knowledge Representation, Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation, Whistler, BC, Canada, 1 June 2004, p. 9.

[22] A. Caruso, A. Folino, F. Parisi and R. Trunfio, “A statistical method for minimum corpus size determination” Proceedings of the 12es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2014), pp. 135-146.

[23] D. Kless and S. Milton, “Towards Quality Measures for Evaluating thesauri” Metadata and semantic research. 4th international conference, MTSR 2010, Alcalá de Henares, Spain, October 20–22, 2010, pp. 312-319.

[24] E. Cardillo, A. Folino, R. Trunfio, and R. Guarasci, “Towards the reuse of standardized thesauri into ontologies”, in Proceeding WOP'14 Proceedings of the 5th International Conference on Ontology and Semantic Web Patterns - Volume 1302, 2014, pp.26-37.

[25] F. Dell’Orletta, G. Venturi, A. Cimino and S. Montemagni, “T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts“. In Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014), 26-31 May, Reykjavik, Iceland, 2014.

[26] Informatics and Telematics Institute – National Council of Research, IIT – CNR, https://www.iit.cnr.it [accessed October 2018]

 ISBN: 978-1-61208-678-1