A Thesaurus for the Cybersecurity domain to specialized knowledge management and indexing operations

Claudia Lanza[1][0000−0002−3018−1987] and Erika Pasceri[1][0000−0001−9917−2184]

[1 University of Calabria, Department of Culture, Education and Society, Arcavacata di Rende (CS), Italy
claudia.lanza@unical.it
erika.pasceri@unical.it

Abstract. In this paper the thesaurus targeted to represent the terminological framework on the Cybersecurity domain in Italian language will be presented. The methodological approach is compliant with the ones existing in the literature when creating semantic means to manage and represent specialized fields of study. The thesaurus will be presented under the perspective of using it for indexing operations supporting the detection of specific documentation meant to represent the informative baseline to foster the knowledge about the domain under investigation, in this case the one related to the Cybersecurity sphere. [1]

Keywords: Thesaurus · Cybersecurity · Indexing · Specialized documentation · Information retrieval.

1 Introduction

In a research scenario headed by Neural Networks and Deep learning [39, 18], despite the rapid growth of data-driven and crowdsourcing approaches[36, 6], resources controlled and validated by domain experts continue to have relevance and robustness that is difficult to achieve with other approaches. This phenomenon is even more pronounced in niche and specific domains. Information retrieval tasks over specialized fields of study can be more efficiently executed by the use of domain-oriented terminology in order to gather the documentation aimed to represent the corresponding knowledge [15, 16, 14]. In this scenario, a semantic tool such as a thesaurus represents a solid base from which to start accessing specialized documentation since it collects in an ordered configuration the terminology representing the main concepts of sector-oriented domains. In the context of Cybersecurity the documentation analysed for knowledge management purposes refers to the several areas characterizing this highly heterogeneous field of study. Thesauri are semantic controlled resources through which

[1] Authors all contributed in the realization of this paper, although Claudia Lanza dealt with Section I, Section II and Section III, while Erika Pasceri dealt with Section IV and Section V.

the retrieval of technical documents can be more easily executed and can lead to more accurate results. They provide a knowledge systematization to support information retrieval tasks within database interaction containing a wide number of specialised documents. This results to be meaningful when the access to Cybersecurity information represents both a way to acquire knowledge about the domain in its multiple sub-areas and a valid support tool for decision making processes or defense application strategies. The realization of the thesaurus in the Cybersecurity domain, in Italian language as a starting point towards the English term mapping, has been achieved to provide a terminological baseline representing the main concepts of the sector and to retrieve the documentation starting by a validated system of terms effectively used within the communities of domain experts. The thesaurus enhances the knowledge acquisition on a given domains of study by exploring the conceptual framework and systematizing it in a set of semantic relationships established among the corresponding terms. A set of existing studies on the main functions of thesauri for subject indexing operations will be presented, together with the features commonly characterizing such semantic resources to organize specialized domains. The second part of the paper will be dedicated to the presentation of the thesaurus, tailored in its former version for the Italian framework, with a focus on the terminological coverage it is able to reach with respect to the domain under investigation, as well as on the importance of using it as a list of terms to retrieve accurate documentation in order to enhance knowledge both related to the whole domain and to Cyber Defense strategy tasks implementation.

In specific areas of study there is a need of semantic tools to guide the understanding of the domain which organize the main concepts in structured systems easily browsed by users. In detail, these resources developed from a semantic perspective for specialized domain are commonly included in the so-called Knowledge Organization Systems (KOSs) [34, 20], which represent the variety of resources, e.g., thesauri, taxonomies, ontologies, classification systems, to organize and collect information proper to specialized domains. More specifically, as [21] points out [20], they are used to:

> *Organize materials for the purpose of retrieval and to manage a collection. A KOS serves as a bridge between the user's information need and the material in the collection. With it, the user should be able to identify an object of interest without prior knowledge of its existence. Whether through browsing or direct searching, whether through themes on a Web page or a site search engine, the KOS guides the user through a discovery process.*

In this sense, thesauri, as one of the types of KOSs, accomplish to the specific goal of allowing users to retrieve the most appropriate form of documentation. In detail, they can support the navigation experience of users by providing a set of semantically related terms on a given domain that can be associated to a group of documents. Indeed, the terms selected to be part of thesauri can be used to identify the main subjects of documents which, as a consequence, can be easily retrieved when selecting the terms associated to determined resources.

In the literature plenty of works published according to various domain-related studies, such as those of [21, 7, 23, 4, 11, 10, 1, 32, 33, 37], to cite a few, address the steps to follow to build a thesaurus including main rules to be compliant with to be developed and the investigation of thesauri's use for indexing operations. The Cybersecurity context is characterized by a multidisciplinary granularity of information shared. This implies that the documentation produced for this domain is characterized by a cross-field coverage when analyzing the contents published by the domain experts. To reach a cross-field knowledge acquisition, several semantic tools are properly built to facilitate this process in a non-ambiguous way. Besides the most recent advances arisen in the field of ontology construction, as another type of KOS, for the Cybersecurity domain[2], there are other semantic existing tools to take into account, e.g., Glossary of Intelligence *The language of informative organisms* published by the Informative System for the Republic Security[3] or the CLUSIT reports[4] in the Italian context, while in the English framework is worth mentioning the *Glossary of key information terms* in the NIST 7298r2 in 2013, the vocabulary included in the ISO 27000:2016[5] or the *Cybersecurity Fundamental Glossary* published by ISACA [6].

2 Thesaurus structure and functions

When referring to thesauri, [33] "a structured collection of concepts and terms for the purpose of improving the retrieval of information. A thesaurus should help the searcher to find good search terms, whether they be descriptors from a controlled vocabulary or the manifold terms needed for a comprehensive free-text search — all the various terms that are used in texts to express the search concept". More specifically, the ISO 25964-1: 2011[7] provides a definition of the thesaurus as follows: "Controlled and structured vocabulary in which concepts are represented by terms organized so that relationships between concepts are made explicit and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms." The ISO 25964-1 of 2011, together with the second part published in 2013, offers a broad overview on the methodology to follow in order to create the semantic resource able to represent the terminology belonging to specialized fields of study. The appropriateness with which thesauri represent

[2] See in this regard Syed, Zareen, et al. 2016. "UCO: A unified cybersecurity ontology." Workshops at the thirtieth AAAI conference on artificial intelligence; Obrst, Leo, Penny Chase, and Richard Markeloff. 2012. "Developing an Ontology of the Cyber Security Domain." STIDS; NIST IR 8138. 2016. "Vulnerability Description Ontology (VDO)"; Rastogi et al. 2020. "MALOnt: An Ontology for Malware Threat Intelligence."

[3] https://www.sicurezzanazionale.gov.it/comunicazione/glossario

[4] Clusit, https://clusit.it/

[5] ISO 27000:2016 Information technology — Security techniques — Information security management systems — Overview and vocabulary

[6] ISACA Glossary, https://www.isaca.org/resources/glossary

[7] Information and Documentation - *Thesauri and interoperability with other vocabularies – Part I: Thesauri for information retrieval, p.12.*

the knowledge of specialized sectors is given by the selection of candidate terms in compliance with the specificity they bear with respect to the technicality within the source documentation, as well as with their effective usage by the domain experts within the specific communicative contexts [22]. The inclusion of specific terms in a thesaurus usually undergoes a normalization process. In this sense, the standardization of language and the compliance with the rules in the corresponding standards, indicating which kind of semantic relationship should be established among terms, support the objective of uniforming a lexicon meant to be a reference for a community of users. Typically, the semantic relations shared by the terms to be included in the thesaurus – established in accordance with the distribution of information in the source documentation where the terms come from – are of three types, as deeply described by the ISO 25964-1 and several works in the reference literature [20]:

1) Equivalence relation: stands for the synonymy connection between terms expressing the same concept, marked through the tags Use (USE) and Used For (UF). An example of equivalence relations is included in the Italian thesaurus for Cybersecurity:

Hacker `UF` *White Hat*

2) Hierarchical relation: expresses the connections among concepts belonging to the same category from a subordination point of view, it is marked by the tags Broader Term (BT) and Narrower Term (NT), specifying also two concepts where one of them is part or included in the other. An example of this kind of relationship included in the Italian thesaurus for Cybersecurity is:

Cyber Threat `NT` Hacker

3) Associative relation: established between concepts belonging to the same or to different categories and not hierarchically connected, it is marked by the tag Related Term (RT). An example of this kind of relationship included in the Italian thesaurus for Cybersecurity is:

Hacker `RT` Hacking

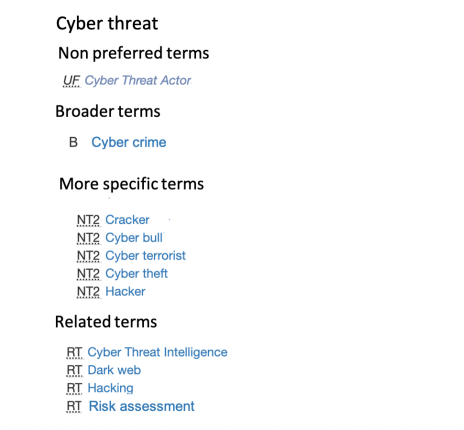
The aforementioned standards provide support also in the steps to follow to include definitions corresponding to terms, the so-called *Scope Notes*, marked by the tag SN, useful to allow the understanding of the terms in a contextualized and unambiguous way. An example of SN included in the Italian thesaurus for Cybersecurity is:

Hacker `SN` Unauthorized user who attempts to or gains access to an information system. SOURCE: CNSSI-4009

The relationships included in the thesaurus have been established from a semantic perspective. The establishment of given semantic connections is strictly related to the analysis of the source corpus documentation from which terms are extracted, and this guarantees a higher specificity level in the representation of a domain under study. The thesaurus taken into account for this paper has been realized for the research purposes within the Cybersecurity Observatory (OCS) platform[8], developed by the Institute of Informatics and Telematics

[8] Cybersecurity Osservatorio, https://www.cybersecurityosservatorio.it/it/Services/thesaurus.jsp

(IIT) at National Council of Research (CNR) in Italy [24]. The OCS platform includes several services to support Cyber Defense operations, among which the thesaurus, guiding the acquisition of specialized knowledge on this domain. The semantic tool has been developed by analysing from a terminological perspective a set of documents, and it includes four main categories decided alongside the quantitative analysis of terms occurrences in the source documents [3] and the validation of domain experts [28]. The thesaurus currently shared on the platform contains **253** terms, distributed, according to the information in the source documentation, in four high-level categories, i.e., Cyber defense, Cyberbullism, Cyber crime, and Cybersecurity. The terms within the main categories are linked together by the use of the aforementioned semantic relationships, thus creating an entangled network of terms in connections with each other through synonymy, hierarchy and association links. The definition of each semantic relationship has been supported by semantic similarity techniques as well as term specificity measures as broadly described in [25]. Figure 1 and Figure 2 show excerpts of the structure of the terms *Cyber threat* and *Hacker* included within the thesaurus connected with other terms within the semantic tool by hierarchical, synonymy and associative correlations.



Cyber threat

Non preferred terms

UF *Cyber Threat Actor*

Broader terms

B Cyber crime

More specific terms

NT2 Cracker
NT2 Cyber bull
NT2 Cyber terrorist
NT2 Cyber theft
NT2 Hacker

Related terms

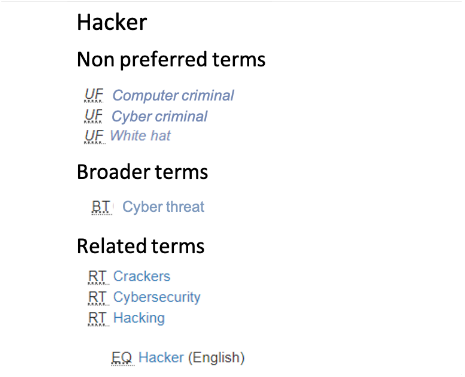
RT Cyber Threat Intelligence
RT Dark web
RT Hacking
RT Risk assessment

Fig. 1. Semantic structure of terms in the thesaurus,*Cyber threat*

A thesaurus results to be a reliable semantic tool able to gather as much as possible a terminological dataset representing the domain under investigation, by organizing it according to a set of connections expressing the distribution of information within the source documentation.

2.1 Source Documentation

The creation of a semantic tool requires the examination of a source corpus of documents reflecting the specialized concepts characterizing the domain to be



Hacker

Non preferred terms

UF Computer criminal
UF Cyber criminal
UF White hat

Broader terms

BT Cyber threat

Related terms

RT Crackers
RT Cybersecurity
RT Hacking

EQ Hacker (English)

Fig. 2. Semantic structure of terms in the thesaurus, *Hacker*

managed and represented. In this regard, a thesaurus, as a means monitoring the terminology proper to specific areas of study marked by a sector-oriented lexicon, is meant to reproduce the informative tissue in a term based structure. To reach this perspective, the Italian thesaurus developed to represent the Cybersecurity domain from a terminological point of view has relied on a representative source corpus made up of authoritative documentation belonging to the sub-areas of legal framework, ICT security field and specialized magazines disseminating technical information on the domain [26]. The ways by which a corpus can be considered representative referring to given domains is a research path broadly investigated in the literature [5, 31, 13, 29]. Authors addressing this subject reflect on the quantitative coverage of the terminology retrieved from source corpora, which, in turn, should be compliant with the current use of language, including in this consideration the evolution language is subjected through time and the competences enhancement. The thesaurus developed for the Italian framework has been based on the information published by authoritative profiles working in this specialized environment as well as on sector-oriented magazines containing a high variety of technical terminology. The corpus has been constituted by **246** legal documents and **314** issues of sector-oriented magazines. As broadly described in [19] [25], the aforementioned documentation, retrieved as specific to the field of study, thus representing the overall concepts related to it, has been processed with Natural Language Processing (NLP) tools [27] as to obtain a list of candidate terms to analyze and include in the thesaural configuration [17, 9]. The quantitative measure of the term bank has significantly varied when using different tools to extract these terminological units. For instance, the thesaurus has been populated by terms alternatively extracted by using the Text2Knowledge (T2K) Italian NLP extractor tool [12] and TermSuite termbank [2], respectively with a consistency of **593 887** for the the first one and **17 083** for the latter. The selection of terms meant to be part of the thesaurus for the Cybersecurity domain has been made under the basis of the Term Frequency/Inverse Document Frequency (TFIDF) statistical measurement [30] which ensured the isolation of the most relevant terminological units referring to their distribution in the source specialized documentation. Documents have been retrieved from the legal framework, specifically they have been constituted by Decree Laws, Parliamentary legislation, Penal/Administrative/Civil Code; Rules (GDPR), CERT guidelines and Government documents, as well as by specific reports on the domain (CLUSIT ones, for example), Guidelines (CERT ones). On this basis, the use of the thesaurus can be considered in a biunivocal dynamic perspective: i) it is the representation of the terminology proper to specialized spheres of study, and this is inferred by the documentation by which this resource is built, characterized by a high sector-oriented way of producing and disseminating informative data; ii) by applying as key access point the preferred terms contained in the thesaurus referring to specialized concepts the specialized documentation can be more efficiently retrieved in databases and platforms.

3 Indexing operations for Cyber Defense tasks

Terms included in the thesauri can be employed to execute a research on platforms gathering specialized documentation on given subjects.A thesaurus is a reliable means through which subject indexing processes can be performed being compliant with the use of a terminology effectively used in the reference context. The semantic relationship configuration a thesaurus provides to systematize the domain-oriented knowledge can act as a guide to select the most appropriate terms to associate to a resource during subject indexing tasks[9]. The synonymy relations, for instance, allows the selection of the preferred terms, mostly used in the indexing operations that involve the use of thesauri, since they ensure the effective use of terms in determined fields of study. For this reason, the choice of a preferred term can enable the retrieval of more accurate results in the users experience when browsing a database within which the documentation of specialized domains is associated to selected terms existing in external resource as thesauri [35, 38]. Using Scope Notes is also useful to understand the scope of a particular set of terms to the domain they belong to. Several fields of study tend to use thesauri as back-end tools to run crawling operations over specific databases, e.g., the European Union Law Access point (Eur-Lex) uses the multilingual multidisciplinary thesaurus EuroVoc[10] created to contain the descriptors to be associated to the legal documentation within the database, as well as PubMed database using the thesaurus MeSH[11] to index its documentation. The Italian Cybersecurity thesaurus can be employed to empower the term bank for existing indexed documents on the platforms referring to this specific domain in order to allow a more appropriate type of research when the acquisition of correct knowledge on technical sectors is required.

[9] See the instructions contained in standard ISO 5963:1985 – Documentation — Methods for examining documents, determining their subjects, and selecting indexing terms

[10] https://eur-lex.europa.eu/browse/eurovoc.html?locale=en

[11] https://www.ncbi.nlm.nih.gov/mesh/

4 Conclusion

The informative system related to the Cybersecurity domain is characterized by a multidisciplinary conceptual framework. The concepts representing this knowledge domain are expressed by using a specialized type of lexicon coming from several sectors, such as that of ICT security or law framework, all of them synergistically guiding the appropriate acquisition of key data on the domain. The specificity designating the documentation on this field of study can be managed and represented by using KOS, which facilitate the association of sector-oriented documentation information with a terminological structures mirroring the connections within the domain-related concepts. This correspondence has proven not only to be advantageous in the systematization of specialized knowledge within a semantic relationship configuration which supports the comprehension of technical information, but also in the subject indexing processes. The thesaurus, particularly the one developed for this field of study on the OCS platform, is meant to be employed as a resource from which to extract the preferred terms to be associated to specialized documentation as its subject abstraction for future Cyber Defense tasks, by acquiring the pertinent informative threshold. The terms in the thesaurus are first retrieved within a set of specialized documents and then these they can be used as key entry points to retrieve more accurate documentation corresponding to several informative needs, in this case those potentially referring to Cyber Defense strategies to be carried out with the adequate domain specific background. The thesaurus configuration helps in targeting in a more accurate way the retrieval of specific information by means of the semantic connections established among terms. Concerning future directions, we plan to investigate emerging approaches promising to exploit current capabilities of traditional neural language models (NLMs) [8] for enriching, updating, and extending such domain-specific resources whose creation requires such a time-consuming effort.

References

1. Aitchison, J., Bawden, D., Gilchrist, A.: Thesaurus Construction and Use: A Practical Manual. Routledge (2003)
2. B., D.: Variations and application-oriented terminology engineering. In Terminology **11**, 181–197 (2005). https://doi.org/10.1075/term.11.1.08dai
3. Bafna, P., Pramod, D., Vaidya, A.: Document clustering: Tf-idf approach. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE (2016)
4. Bessagnet, M.N., Kergosien, E., Gaio, M.: Extraction de termes, reconnaissance et labellisation de relations dans un thésaurus. In: CIDE'12: 12e Colloque International sur le Document Electronique. pp. 275–286. Montréal, Canada (2009)
5. Biber, D.: Representativeness in corpus design. Literary and Linguistic Computing **8**(4), 243–257 (1993)
6. Bonetti, F., Leonardelli, E., Trotta, D., Guarasci, R., Tonelli, S.: Work hard, play hard: Collecting acceptability annotations through a 3d game.

p. 1740 – 1750 (2022), https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144445524partnerID=40md5=9e8e4addd231c367f1897d91c1d974b4

7. Broughton, V., Cavaleri, P.: Costruire thesauri: strumenti per indicizzazione e metadati semantici, vol. 2008. Bibliografica (2008)
8. Buonaiuto, G., Guarasci, R., Minutolo, A., De Pietro, G., Esposito, M.: Quantum transfer learning for acceptability judgements. Quantum Machine Intelligence **6**(1) (2024). https://doi.org/10.1007/s42484-024-00141-8,
9. Cardillo, E., Portaro, A., Taverniti, M., Lanza, C., Guarasci, R.: Towards the automated population of thesauri using bert: A use case on the cybersecurity domain. Lecture Notes on Data Engineering and Communications Technologies **193**, 100 – 109 (2024). https://doi.org/10.1007/978-3-031-53555-0_10,
10. Castellví, M.T.C.: Terminology: Theory, methods and applications. John Benjamins Publishing Co (1999)
11. Darmoni, S.J., Soualmia, L.F., Letord, C., Jaulent, M.C., Griffon, N.: Improving information retrieval using medical subject headings concepts: a test case on rare and chronic diseases. Journal of the Medical Library Association **100**(3), 176–183 (2012)
12. Dell'Orletta, F., Venturi, G., Cimino, A., Montemagni, S.: T2K: a system for automatically extracting and organizing knowledge from texts. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
13. Francom, J., LaCross, A., Ussishkin, A.: How specialized are specialized corpora? behavioral evaluation of corpus representativeness for maltese. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (2010)
14. Guarasci, R., Damiano, E., Minutolo, A., Esposito, M.: Towards a gold standard dataset for open information extraction in italian. p. 447 – 453 (2019). https://doi.org/10.1109/SNAMS.2019.8931822,
15. Guarasci, R., Damiano, E., Minutolo, A., Esposito, M.: When lexicon-grammar meets open information extraction: A computational experiment for italian sentences. vol. 2481 (2019), https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074839710partnerID=40md5=10643863003e88686a4af3749383a489
16. Guarasci, R., Damiano, E., Minutolo, A., Esposito, M., De Pietro, G.: Lexicon-grammar based open information extraction from natural language sentences in italian. Expert Systems with Applications **143** (2020). https://doi.org/10.1016/j.eswa.2019.112954,
17. Guarasci, R., De Pietro, G., Esposito, M.: Quantum natural language processing: Challenges and opportunities. Applied Sciences (Switzerland) **12**(11) (2022). https://doi.org/10.3390/app12115651,
18. Guarasci, R., Silvestri, S., Esposito, M.: Probing cross-lingual transfer of xlm multilanguage model. Lecture Notes on Data Engineering and Communications Technologies **193**, 219 – 228 (2024). https://doi.org/10.1007/978-3-031-53555-0_21,
19. Hazem, A., Daille, B., Claudia, L.: Towards automatic thesaurus construction and enrichment. In: Proceedings of the 6th International Workshop on Computational Terminology. pp. 62–71 (2020)
20. Hjørland, B.: What is knowledge organization (ko)? KO Knowledge Organization **35**(2-3), 86–101 (2008)
21. Hodge, G.M.: Systems of knowledge organization for digital libraries: beyond traditional authority files. No. 91, Digital Library Federation (2000)

22. Joho, H., Sanderson, M.: Document frequency and term specificity. In: Proceedings of the Recherche d'Information Assistée par Ordinateur Conference (RIAO). Sheffield (2007)
23. Lancaster, F.: Thesaurus Construction and Use: A Condensed Course. General Information Programme and Unisist, Unesco (1985)
24. Lanza, C.: An italian thesaurus for cybersecurity and its integration with an ontology structure. AIDAinformazioni **1-2** (2019)
25. Lanza, C.: Semantic control for the cybersecurity domain: investigation on the representativeness of a domain-specific terminology referring to lexical variation. CRC Press (2022)
26. Lanza, C., Daille, B.: Terminology systematization for cybersecurity domain in italian language. In: JEPTALNRECITAL (2019)
27. Marulli, F., Pota, M., Esposito, M., Maisto, A., Guarasci, R.: Tuning syntaxnet for pos tagging italian sentences. Lecture Notes on Data Engineering and Communications Technologies **13**, 314 – 324 (2018). https://doi.org/10.1007/978-3-319-69835-9_30,
28. Nazarenko, A., Zargayouna, H., Hamon, O., Puymbrouck, J.V.: Evaluation des outils terminologiques: enjeux, difficultés et propositions. Traitement Automatique des Langues **50**(1 varia), 257–281 (2009)
29. Pastor, G.C., Seghiri, M.: Size matters: A quantitative approach to corpus representativeness. In: Language, Translation, Reception. To Honor Julio César Santoyo, pp. 1–35. Universidad de León
30. Qaiser, S., Ali, R.: Text mining: use of tf-idf to examine the relevance of words to documents. International Journal of Computer Applications **181**(1), 25–29 (2018)
31. Rapp, R.: Using collections of human language intuitions to measure corpus representativeness. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2117–2128. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (2014)
32. Reich, P., Biever, E.J.: Indexing consistency: The input/output function of thesauri. College & Research Libraries **52**(4), 336–342 (1991)
33. Soergel, D.: Indexing Languages and Thesauri: Construction and Maintenance. Melville Publishing Company, Los Angeles, CA (1974)
34. Soergel, D.: Knowledge organization systems: overview. Dsoergel, Alexandria (2009)
35. Srinivasan, P.: Thesaurus construction. In: Information Retrieval: Data Structures and Algorithms, pp. 161–218 (1992)
36. Trotta, D., Stingo, M., Guarasci, R., Elia, A., Albanese, T.: Multi-word expressions in spoken language: Polisdict. vol. 2253 (2018), https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057730561partnerID=40md5=d1aa2eb532830d71c8eb0dddbacf3a09
37. Tudhope, D., Binding, C., Blocks, D., Cunliffe, D.: Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation **62**(4), 509–533 (2006)
38. Willis, C., Losee, R.M.: A random walk on an ontology: Using thesaurus structure for automatic subject indexing. Journal of the American Society for Information Science and Technology **64**(7), 1330–1344 (2013)
39. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)