AGILE: GIScience Series, 5, 32, 2024. https://doi.org/10.5194/agile-giss-5-32-2024
Proceedings of the 27th AGILE Conference on Geographic Information Science, 4–7 Sept. 2024.
Editors: Alison Heppenstall, Mingshu Wang, Urska Demsar, Rob Lemmens, and Jing Yao.
This contribution underwent peer review based on a full paper submission.

Comparative Evaluation of Keyphrase Extraction Tools for Semantic Analysis of Climate Change Scientific Reports and Ontology Enrichment

Eirini Katsadaki[1], Margarita Kokla[1]

[1] School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Athens, Greece

Correspondence: Eirini Katsadaki (eirinikats@mail.ntua.gr)

Abstract. Keyphrase extraction is a process used for identifying important concepts and entities within unstructured information sources to facilitate ontology enrichment, semantic analysis, and information retrieval. In this paper, three different tools for key phrase extraction are compared to evaluate their accuracy and effectiveness for extracting geospatial and climate change concepts from climate change reports: frequency-inverse document frequency (TF-IDF), Amazon Comprehend, and YAKE. Climate change reports contain vital information for comprehending the complexity of climate change causes, impacts, and interconnections, and include wealth of information on geospatial concepts, locations, and events but the diverse terminology used complicates information extraction and organization. The highest scoring keyphrases are further used to enrich and populate the SWEET ontology with concepts and instances related to climate change and meaningful relations between them to support semantic representation and formalization of knowledge.

Keywords. Keyphrase Extraction, Ontology Enrichment, SWEET ontology, Climate Change, Geospatial Concepts

1 Introduction

In the last few decades, the amount of available unstructured content, such as scientific reports, news articles, travel blogs, and historical archives has increased enormously. These sources contain wealth of information on geospatial concepts, places, events, activities, etc. in a form mainly intended for human use. Keyphrase extraction is a process used for identifying important concepts and entities within these unstructured information sources to facilitate knowledge formalization and organization, semantic analysis, and information retrieval. Keyphrase extraction may also support ontology enrichment by forming bridges between natural language and formalized ontologies, improving semantic representation and integration.

In this paper, three different tools for key phrase extraction from climate change reports are evaluated, followed by the enrichment of the SWEET (Semantic Web for Earth and Environmental Terminology) (https://github.com/ESIPFed/sweet) Ontology with the highest scoring key phrases.

Climate change is an extremely complex issue affecting many systems: atmosphere, land surface, oceans, bodies of water, snow and living organisms on Earth. The multifariousness of this global issue is manifested by the complex terminology and the intersection of different disciplines associated with it, such as environmental science, meteorology, oceanography, economy, and politics. Climate change reports, such as these prepared by the Intergovernmental Panel on Climate Change (IPCC, 2022) and other organizations contain vital information for comprehending the complexity of climate change causes, impacts, and interconnections among different systems, but the diverse terminology used complicated information organization and semantic analysis.

The paper is structured as follows: Section 2 summarizes related work on keyphrase extraction and ontology enrichment. In Section 3 the proposed approach is presented, including the methodology and tools used for keyphrase extraction and ontology enrichment. The main findings and suggestions for future research directions are described in Section 4.

2 Related Work

Keyword and keyphrase extraction is a process that identifies sets of representative words and phrases within a document or corpus of documents that can provide a highly succinct summary of the content and support document analysis, organization, and retrieval based on their content (Siddiqi & Sharan, 2015). Keyphrase extraction techniques is performed using various methods. The most common ones are frequency-based methods which extract the most frequently occurring words or phrases from a text corpus, TF-IDF methods that consider the importance of words as well as the phrases in the textual context, graph-based ranking models that calculate importance based on links between nodes in a graph, machine learning methods, and hybrid ones that combine two or more techniques to improve accuracy (Alami Merrouni et al., 2020). Information retrieval, text classification, sentiment analysis, and topic identification are just a few of the many applications related to keyphrase extraction (Li, 2021). Keyphrase extraction may improve the accuracy of information retrieval systems, reduce the need for manual indexing, and enable efficient organization and management of large volumes of textual data.

Keyphrase extraction is also used for ontology learning and population. Ontology learning refers to the process of constructing a new ontology or enriching an existing ontology with concepts and relations, whereas ontology population refers to the process of adding new instances of concepts to an existing ontology (Kokla, 2021). There are several methods of ontology enrichment from text, including manual curation, automatic extraction, and semi-automatic methods that combine both approaches (Iyer et al., 2019). For the (semi-) automatic extraction, both shallow and deep learning approaches are utilized to enhance and broaden term extraction, relation discovery, and axiom learning (Al-Aswadi et al., 2020).

Calbimonte et al. (2019) proposed a semi-automatic rule-based method for the semantic enrichment of personal data streams based on semantic concepts from standardized specialized vocabularies such as SNOMED-CT. Hasan et al. (2019) implemented a semi-automated mapping approach for enriching existing ontologies associating words or words' meaning with related concepts in WordNet. Reyes-Ortiz (2019) implemented an ontology population and enrichment method using pattern recognition from text to extract criminal events from Spanish text. Tissaoui et al. (2020) used an approach for semi-automatic ontology enrichment from textual corpus based on Latent Dirichlet Allocation (LDA). LDA provided efficient dimension reduction, to capture semantic word-topic and topic-document relations in terms of probability distributions. Mellal et al. (2021) proposed a method that is based on Natural Language Processing (NLP) techniques but augmented by a heuristic algorithm that allows reducing extracted sentences to SVO (Subject, Verb, and Object) and identifying relations with those of the existing ontology as well as the placement of new concepts in it.

The present paper performs a comparative evaluation of three keyphrase extraction approaches / tools regarding their ability to extract geospatial and climate change related keyphrases: Amazon Comprehend, TF-IDF, and Yake. The extracted keyphrases are subsequently used to enrich the SWEET ontology with relevant geospatial and climate change concepts and relations between them.

3 Methodology

The proposed methodology consists of four main steps: (1) pre-processing, (2) extraction of place names, (3) keyphrase extraction, and (4) ontology enrichment (Figure 1).

The keyphrase extraction and ontology enrichment processes used as input the Chapter 16: "Key Risks across Sectors and Regions" of the Sixth Assessment Report of the Intergovernmental Panel on Climate Change Working Group II (O'Neill et al., 2022). This Chapter analyses the state of knowledge about the impacts of climate change, key risks, and the association to adaptation efforts and includes numerous references to places, events, and spatial concepts related to climate change.

The first step involved preprocessing of the input raw text, including the removal of irrelevant information, and tokenization. Tokenization analyses the preprocessed text into individual tokens or units, such as words or subwords, to facilitate the subsequent analysis and extraction of keyphrases from the climate change report.

The second step involved named entity recognition (NER) on the preprocessed text and subsequent visualization of locations on a map using the Python library 'spaCy' (Figure 2).

The third step employed three distinct tools / approaches for keyphrase extraction to compare their accuracy and effectiveness for the extraction process: (a) Amazon Comprehend, (b) TF-IDF, and (c) Yake.

Amazon Comprehend (https://aws.amazon.com/comprehend/) is a web service that uses deep learning-based NLP and topic modelling for topic-based classification, content-based search, and sentiment analysis. Using a combination of statistical techniques, rule-based matching, linguistic heuristics, and deep learning-based models, Amazon Comprehend extracted keyphrases related to geospatial and climate change concepts as shown in Figure 3a.

TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used algorithm for keyphrase extraction that calculates the relevance of a term within a document or corpus (Luhn, 1958). The Python TF-IDF library was used to assign weights to terms based on their frequency within a single document and their inverse frequency across that document. This approach prioritizes terms that appear frequently within the document while being less common overall, highlighting their significance within the context of the document. The resulting scores were used as a guide to rank and select the most significant keyphrases (Figure 3b).

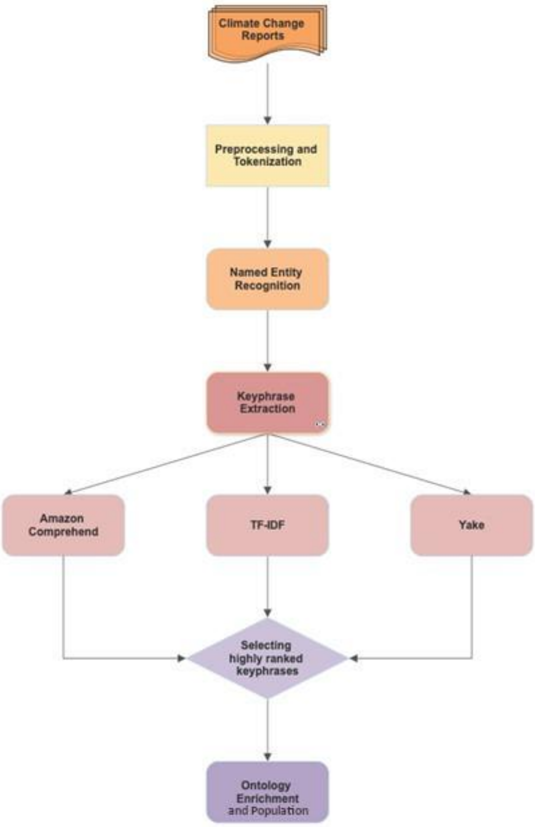
Climate Change Reports
Preprocessing and Tokenization
Named Entity Recognition
Keyphrase Extraction
Amazon Comprehend
TF-IDF
Yake
Selecting highly ranked keyphrases
Ontology Enrichment and Population

Figure 1: Workflow of the proposal approach

YAKE (Yet Another Keyword Extractor) adopts a machine learning-based method consisting of statistical and linguistic features to identify keyphrases (Campos et al., 2020). YAKE uses a sequence labelling algorithm to identify and extract keyphrases based on their statistical properties, such as their frequency and distribution within the text, as well as their linguistic properties, such as part of speech and position in the sentence. The resource is accessible via a Python library and the extracted concepts with the highest scores are shown in Figure 3c.

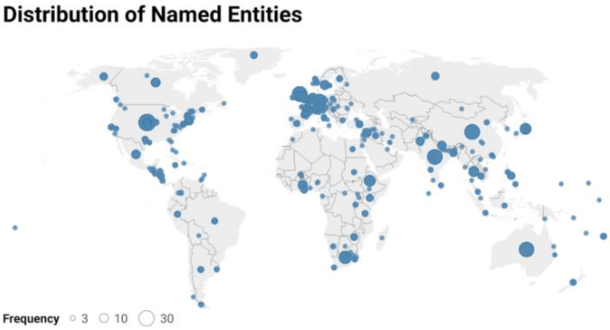

Figure 2: Distribution of locations and frequency of reference.

Table 1. Similarity Score with Locations from NER

Keyphrase Extraction Approach	Total Keywords	Locations Match with NER	Percentage Match with NER
Amazon Comprehend	20875	218	1.04%
TF-IDF	101700	1219	1.20%
Yake	50	2	4.00%

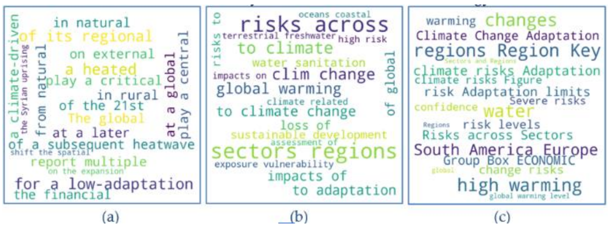

Figure 3: (a) Amazon Comprehend, (b) TF-IDF and (c) Yake's highest Score Concepts

From Figure 3, it can be observed that several keyphrases appear more like n-grams than unique keywords. To properly address this issue, it is essential to consider both the text's fundamental properties and the keyphrase extraction methods used. Because of their reliance on statistical and linguistic patterns, these techniques may include articles, prepositions, and non-useful words in the resulting keyphrases. Furthermore, the complexity and variety of climate change reports, paired with specialized terminology, contribute to the development of longer and more elaborate phrases. To address this issue, improving the preprocessing stages and fine-tuning the tool parameters may improve the quality of extracted keyphrases.

The fourth step leveraged the extracted keyphrases with the highest scores for ontology enrichment to create a more comprehensive and extended representation of the domain concepts used in the input report.

Various ontologies have been developed to formalize the complex concepts and relations related to the environment and climate change such as CCTL (Climate Change TimeLine) (https://github.com/sfpileggi/CCTL-Ontology), ENVO (Environment Ontology) (https://sites.google.com/site/environmentontology/), GEMET (GEneral Multilingual Environmental Thesaurus) (https://www.eionet.europa.eu/gemet/en/themes/), and SWEET (Semantic Web for Earth and Environment Technology Ontology). SWEET has been selected for ontology enrichment in the present research due to the coverage of concepts related to the Earth systems, including climate change. Its modular structure offers flexibility, allowing for targeted enrichment based on specific aspects of climate change addressed in the reports. SWEET also aligns with other ontologies such as ENVO, facilitating interoperability and the integration of knowledge from different sources.

Cosine similarity (Singhal, 2001) was employed to measure the resemblance between keyphrases and the ontology, assessing alignment and compatibility. In general, cosine similarity quantifies the cosine of the angle between two vectors, representing the degree of alignment or resemblance between them. In NLP, cosine similarity is important as it allows the comparison and similarity assessment between textual documents, words, or phrases. Among the approaches considered, Amazon Comprehend consistently yielded the highest cosine similarity score, followed by TF-IDF, and lastly, Yake (Table 2). The extracted keyphrases provided us with important domain-specific terms that were used for semantic analysis and enrichment of the ontology. Table 3 shows indicative keyphrases that are formalized as subclasses of original SWEET ontology concepts.

To enhance the NER process, the extracted locations were cross-referenced with the keyphrases through the use of the 'fuzzywuzzy' library in Python. ‘Fuzzywuzzy’ library employs string matching algorithms to detect similarities between text strings, facilitating the identification of comparable patterns and resemblances in textual data. The purpose of this supplementary stage was to detect any similarities between the locations detected via NER and the keyphrases extracted in the subsequent step (Table 1).

Table 2. Similarity Score with SWEET Ontology

Keyphrase Extraction Approach	**Cosine Similarity Score**
Amazon Comprehend	34.1%
TF-IDF	22.6%
Yake	1.1%

Table 3. Examples of keyphrases added as subclasses of SWEET concepts.

New Concepts	**Relation**	**SWEET Concepts**
Urban Flood Coastal Flood Fluvial Flood River Flood	SubClass Of	Flood
Agricultural Drought Ecological Drought	SubClass Of	Drought
Vector-Borne Disease Water-Borne Disease Food-Borne Disease	SubClass Of	Disease
Marine Biodiversity Terrestrial Biodiversity Alpine Biodiversity	SubClass Of	Biodiversity
Human migration	SubClass Of	Migration
Internal migration International migration Urban migration	SubClass Of	Human migration
Economic impact Societal impact	SubClass Of	Impact

Additionally, locations with associated keyphrases, confirmed through the 'fuzzywuzzy' library, were considered potential new instances related to climate change for populating the SWEET ontology. The decision to populate the ontology through keyword extraction was driven by the richness in extracted instances, places, and events, surpassing the information gathered solely from the NER process in the second step. For example, the extraction process also identified keyphrases related to specific events, such as the 2003 European heatwave, the 2019 Australian bushfires and the 2010 Amazon drought with scores above 80%. These events populated the ontology as instances of climate phenomena, to provide insights into the specific impacts of climate change that have occurred in different regions and timeframes.

Besides subclasses of climate change concepts, in order to capture the multifaceted topic of climate change, it is necessary to address the multitude and diversity of correlations and impacts on the environment, health, society, and economy. These are expressed in various ways in natural language texts and especially through cause-effect and association relations. For example, internal and international migration are linked to extreme events, whereas waterborne diseases are a common cause of morbidity and mortality (Table 4).

Figure 4 shows an excerpt of the enriched ontology. The new concepts are shown with orange outlines and the new relations with orange lines.

5 Conclusions

Our work evaluated three keyphrase extraction approaches/ tools regarding their ability to identify geospatial and climate change related concepts and instances from climate change scientific reports. The most prominent concepts and instances were further used to enrich and populate the SWEET ontology.

Table 4. Cause-effect and other associations between concepts

New Concept	Relation	Sweet Ontology
Climate change risk	Caused-By	Climate change
Internal migration	Linked-To	Extreme event
International migration	Linked-To	Extreme event
Morbidity	Caused-By	Water-borne disease
Mortality	Caused-By	Water-borne disease
Morbidity	Caused-By	Vector-Borne disease
Mortality	Caused-By	Vector-Borne disease
Mortality	Caused-By	Extreme event
Maladaptation	Opposite Of	Adaptation

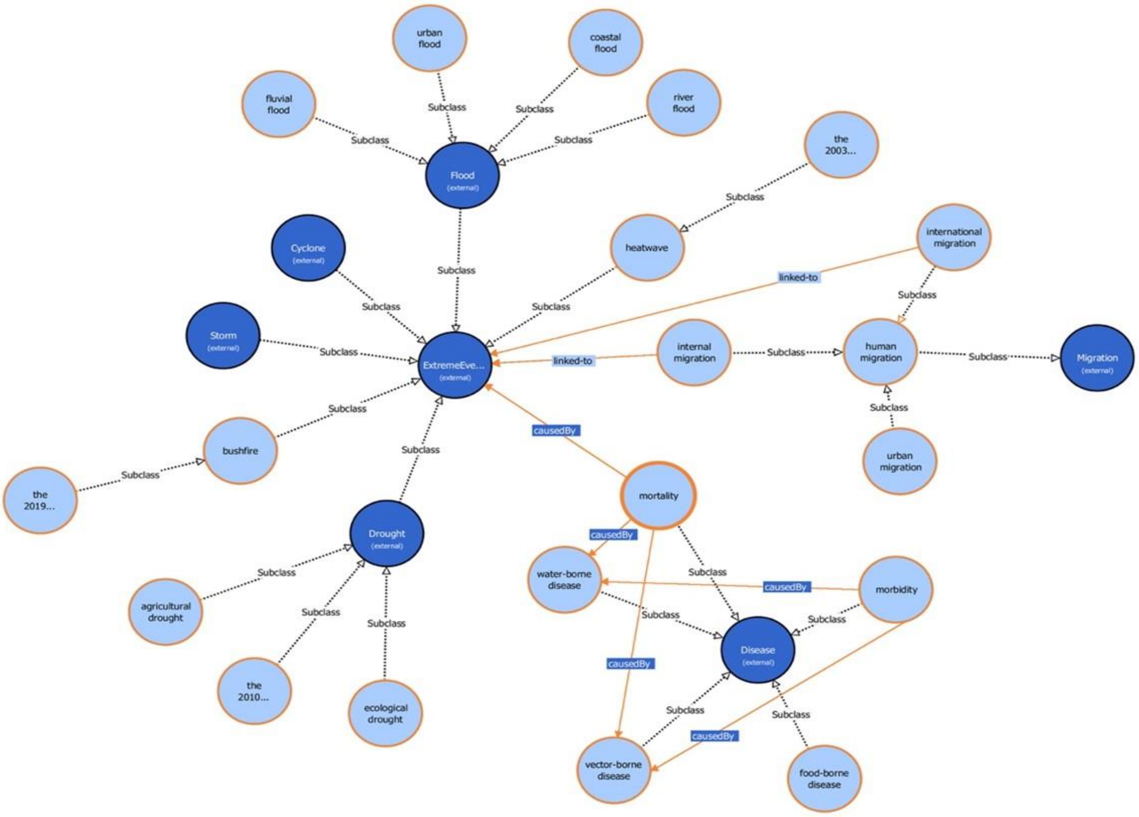

Figure 4. Excerpt of the enriched ontology.

The resulting enriched ontology captures meaningful relations between concepts, uncovering connections between climate change and factors such as urbanization, poverty, human mobility, maladaptation, and other social, economic, and environmental aspects. Moreover, the inclusion of keyphrases related to specific natural disasters, such as droughts, heatwaves, and wildfires has expanded the scope of the ontology and improved its comprehensiveness in capturing complex interactions between climate change and its impacts across regions.

However, there are some limitations to our approach. Climate change is a multi-faceted topic, and relation extraction techniques could be used as an additional process to identify the complicated relations among climate change related concepts, as well as their connections to specific places on Earth.

Future research in this area could also explore the use of other natural language processing techniques combined with deep learning techniques such as neural networks, incorporate additional data sources, and further validate the enriched ontology with domain experts to refine and improve its accuracy and comprehensiveness.

6 Data and Software Availability

We provide the dataset we used for our study, and to ensure reproducibility and transparency to our model and results, the code for running the analysis is available online at https://github.com/EiriniKat94/Comparative-Evaluation-of-Keyphrase-Extraction-Tools-Climate-Change-Reports.git.

References

Alami Merrouni, Z., Frikh, B., & Ouhbi, B.: Automatic keyphrase extraction: a survey and trends. Journal of Intelligent Information Systems, 54, 391-424. https://doi.org/10.1007/s10844-019-00558-9, 2020.

Al-Aswadi, F. N., Chan, H. Y., & Gan, K. H.: Automatic ontology construction from text: a review from shallow to deep learning trend. Artificial Intelligence Review, 53, 3901-3928. https://doi.org/10.1007/s10462-019-09782-9, 2020.

Calbimonte, J., Dubosson, F., Kebets, I., Legris, P., & Schumacher, M.I.: Semi-automatic Semantic Enrichment of Personal Data Streams. International Conference on Semantic Systems, 2019.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A.: YAKE! Keyword extraction from single documents using multiple local features. Information Sciences, 509, 257-289, https://doi.org/10.1016/j.ins.2019.09.013, 2020.

Hasan, M.J., Badhan, A.I., Ahmed, N.I.: Enriching Existing Ontology Using Semi-automated Method. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Advances in Information and Communication Networks. FICC 2018. Advances in Intelligent Systems and Computing, vol 886. Springer, Cham. https://doi.org/10.1007/978-3-030-03402-3_32, 2019.

Intergovernmental Panel on Climate Change (IPCC): Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, https://doi.org/10.1017/9781009325844, 2023.

Iyer, V., Mohan, L., Bhatia, M., & Reddy, Y. R.: A survey on ontology enrichment from text. Proceedings of the 16th International Conference on Natural Language Processing (pp. 95-104), https://aclanthology.org/2019.icon-1.11, 2019.

Kokla, M.: Semantic Information Elicitation. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2021 Edition), John P. Wilson (ed.). https://doi.org/10.22224/gistbok/2021.2.10, 2021.

Li, J.: A comparative study of keyword extraction algorithms for English texts. Journal of Intelligent Systems, 30(1), 808-815, https://doi.org/10.1515/jisys-2021-0040, 2021.

Luhn, H. P.: The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165, https://doi.org/10.1147/rd.22.0159, 1958.

Mellal, N., Guerram, T., & Bouhalassa, F.: An approach for automatic ontology enrichment from texts. Informatica, 45(1), https://doi.org/10.31449/inf.v45i1.2586, 2021.

O'Neill, B., M. van Aalst, Z. Zaiton Ibrahim, L. Berrang Ford, S. Bhadwal, H. Buhaug, D. Diaz, K. Frieler, M. Garschagen, A. Magnan, G. Midgley, A. Mirzabaev, A. Thomas, and R. Warren, 2022: Key Risks Across Sectors and Regions. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press, pp. 2411–2538, https://doi.org/10.1017/9781009325844.025, 2023.

Reyes-Ortiz, J. A.: Criminal event ontology population and enrichment using patterns recognition from text. International Journal of Pattern Recognition and Artificial Intelligence, 33(11), 1940014, https://doi.org/10.1142/S0218001419400159, 2019.

Siddiqi, S., & Sharan, A.: Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications, 109(2), https://doi.org/10.5120/19161-0607, 2015.

Singhal, A.: Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35-43, 2001.

Tissaoui, A., Sassi, S., & Chbeir, R. (2020, November). LEOnto: new approach for ontology enrichment using LDA. Proceedings of the 12th International Conference on Management of Digital EcoSystems (pp. 132-139), https://doi.org/10.1145/3415958.3433076, 2020.