frontiers in Research Metrics and Analytics

OPINION
published: 28 May 2021
doi: 10.3389/frma.2021.694307

Enhancing Knowledge Graph Extraction and Validation From Scholarly Publications Using Bibliographic Metadata

Houcemeddine Turki[1,2], Mohamed Ali Hadj Taieb[2], Mohamed Ben Aouicha[2], Grischa Fraumann[3], Christian Hauschke[4] and Lambert Heller[4]*

[1]Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia, [2]Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia, [3]Department of Communication, University of Copenhagen, Copenhagen, Denmark, [4]Open Science Lab, TIB-Leibniz Information Centre for Science and Technology, Hannover, Germany

Keywords: data mining and knowledge discovery, information retrieval and extraction, knowledge graph (ontologies), bibliometric-enhanced information retrieval, bibliographic metadata

OPEN ACCESS

Edited by:
Philipp Mayr, GESIS Leibniz Institute for the Social Sciences, Germany

Reviewed by:
Suzan Verberne, Leiden University, Netherlands
Ingo Frommholz, University of Wolverhampton, United Kingdom

****Correspondence:***
Houcemeddine Turki
turkiabdelwaheb@hotmail.fr

Specialty section:
This article was submitted to Text-mining and Literature-based Discovery, a section of the journal Frontiers in Research Metrics and Analytics

Received: *12 April 2021*
Accepted: *19 May 2021*
Published: *28 May 2021*

Citation:
Turki H, Hadj Taieb MA, Ben Aouicha M, Fraumann G, Hauschke C and Heller L (2021) Enhancing Knowledge Graph Extraction and Validation From Scholarly Publications Using Bibliographic Metadata. Front. Res. Metr. Anal. 6:694307. doi: 10.3389/frma.2021.694307

INTRODUCTION

Fully structured semantic resources representing facts in the form of triples (i.e., knowledge graphs) have a major function in driving computer applications, particularly the ones related to biomedicine, to library and information science and to digital humanities (Haslhofer et al., 2018; Sargsyan et al., 2020). They can be easily processed using Application Programming Interfaces (APIs, like REST APIs) and query languages (mainly SPARQL) to assess the reference semantic information and to generate accurate and precise interpretations and predictions, particularly when the analyzed data is multifactorial and ever-changing such as the COVID-19 knowledge (Turki et al., 2021c), information about the laureates of Nobel Prize in Literature (Lebuda and Karwowski, 2016), and the findings of scholarly publications (Fathalla et al., 2017). In particular, the role of open knowledge graphs to facilitate scientific collaboration has been stressed against the backdrop of the COVID-19 pandemic (Anteghini et al., 2020; Colavizza et al., 2021; Turki et al., 2021a). Effectively, the information included in textual or semi-structured resources such as electronic health records, scholarly publications, encyclopedic entries, and citation indexes can be converted into fully structured Research Description Framework (RDF) triples and included in knowledge graphs and then processed in near real-time using computer methods to obtain evolving research outputs that are automatically updated as the knowledge graphs feeding them is regularly curated. These living research outputs include systematic reviews (Wang and Lo, 2021), clinical trials (Servant et al., 2014), scientometric studies (Nielsen et al., 2017), and epidemiological studies (Turki et al., 2021b).

However, the construction of knowledge graphs is a complex effort including the recognition of scholarly publications related to the scope of the semantic resource (Turki, 2018), the retrieval of abbreviations and terms for every concept (Turki et al., 2021a), and the extraction and validation of semantic relations (Turki et al., 2018a). Many projects depend on advanced neural network-driven machine-learning techniques for applying these tasks as these methods contribute to higher quality (Asada et al., 2021; Fei et al., 2021). However, these techniques are considered as black boxes and cannot be debugged to identify the reasons behind returned false results and consequently to solve these limitations in a transparent way (Turki et al., 2021b). What is more, the quality of these techniques is considered imperfect in some cases, requiring more time to achieve the same results as specific well-defined algorithms (Turki et al., 2021b). Here, Bibliometric-Enhanced Information Retrieval (BIR) has evolved as a novel field that utilizes bibliographic metadata to efficiently drive the extraction and refinement of semantic data from scholarly publications (Cabanac et al., 2018). This field contributed to the development of many intuitive and explainable algorithms for knowledge

engineering. On the one hand, this has been achieved through the restriction of the analysis of full texts to the publications including a particular value of a metadata to reveal the bibliographic settings where assessed algorithms perform well or bad (Safder and Hassan, 2019). On the other hand, this could be done thanks to the analysis of the bibliographic information using taxonomies like MeSH and Wikipedia Category Graph (Hadj Taieb et al., 2020) or using the probabilistic heuristics and constraints inferred from publications using statistical models including TF*IDF (Ramos, 2003) or extracted from knowledge graphs using inference engines, particularly HySpirit (Fuhr and Rölleke, 1998) and F-OWL (Zou et al., 2004).

In this opinion article, we explain how each type of bibliographic metadata can provide useful insights to enhance the automatic enrichment and fact-checking of knowledge graphs from scholarly publications based on the outcomes of research efforts about BIR.

BIBLIOGRAPHIC METADATA FOR ENHANCING INFORMATION RETRIEVAL FROM PUBLICATIONS

In the following subsections, we provide an overview on bibliographic metadata that plays a pivotal role when employing BIR. We selected these types of metadata based on its frequent use in BIR and its wide availability. Title, abstract, controlled keywords, citation analysis, section title as well as further metadata of scholarly publications can be easily retrieved compared to full texts that are sometimes hidden behind paywalls. As such, these types provide the opportunity for better precision and recall for information retrieval.

Title and Abstract

To start, title and abstract are two metadata elements that can enrich knowledge graphs. Even though titles and abstracts of scholarly publications are written in natural languages and are not semi-structured as the other types of bibliographic information, they give insights about the purpose and outcomes of research outputs in a concise way. That is why they can efficiently represent the topics and the format of research papers (Stotesbury, 2003; Letchford et al., 2015). Consequently, the time-consuming and demanding natural language processing of full texts is not required when a brief analysis of titles and abstracts can return required information for information retrieval purposes. As such, the availability of open abstracts as research data has been recently emphasized by the Initiative for Open Abstracts (I4OA) (Tay et al., 2020).

The application of feature-based measures of sentences' semantic similarity to compare the titles or abstracts of two scholarly publications can be efficient to identify whether the two papers describe similar topics or not (Hadj Taieb et al., 2019; Hadj Taieb et al., 2020) and this can serve to contextualize the co-citation and citation links between papers as well as to filter term co-occurrences for a more precise knowledge graph construction and refinement (Hadj Taieb et al., 2020). This is particularly true for the domain of STEM (Science, Technology, Engineering, and Mathematics) since it features, compared to arts and humanities as well as social sciences, an agreed-upon vocabulary which mostly addresses directly its subjects. Semantic similarity measures are driven by external knowledge resources like knowledge graphs and ontologies and can consequently compare brief texts with high accuracy and speed (Hadj Taieb et al., 2015) and full transparency (Turki et al., 2021b) by contrast to other advanced techniques applied to full texts, particularly deep learning, semantic embeddings (Sargsyan et al., 2020), TF-IDF[1] (White, 2018), and Latent Dirichlet Allocation (Jeong et al., 2014).

The consideration of the format of the titles and abstracts when applying information retrieval techniques can be an important factor for advancing the state of the art of the knowledge engineering field. The letter case of words in titles and abstracts can be useful for many information retrieval applications. For example, it can help identify scholarly abbreviations that are generally written in uppercase letters (Zhou et al., 2006) or to extract structured abstracts including uppercase section titles (Ripple et al., 2011). Such algorithms should be considered with care, particularly when the title or the abstract is fully written in uppercase letters such as in Telford et al. (1985). In this situation, case-sensitive algorithms should not be applied to uppercase titles and abstracts as this can alter the efficiency of the methods. Such systems cannot even be applied to the full text of a scholarly publication when both the title and the abstract are in uppercase letters. Consequently, such decisions can only be made by human reading of the title, abstract, and full text, if necessary.

The restriction of several natural language processing algorithms to the titles and abstracts of scholarly publications can be associated with higher accuracy rates for methods. The usage of several patterns in titles and abstracts, particularly parentheses, can be less complicated in titles and abstracts than in full texts and this explains in part the higher accuracy of parenthetic abbreviation extraction from titles (Zhou et al., 2006). For instance, there are more situations where parentheses are used in full texts for explaining facts, mentioning in-text references and defining *p*-values for evaluating assumptions when parentheses are mainly used in titles for stating abbreviations and declaring chemical formulas (Zhou et al., 2006). This phenomenon should raise concerns about the application of information retrieval methods only tested on titles and abstracts to full texts on the one hand and detailed guidelines for deciding when the used methods should be restricted to titles and abstracts to obtain a better precision and recall for information retrieval.

Controlled Keywords, Citation Analysis, and Section Title

As a rule, richer metadata on publications are available than presented in the previous section. This includes contextual information such as content classifications, relationships to other documents as well as structure of content within an article.

Regarding classification of the publication as a whole, controlled keywords are featured as terms from a reference

[1]TF-IDF: term frequency–inverse document frequency.

terminology that are used to label scholarly publications in several bibliographic databases. Examples of these keywords are KeywordPlus attributed to Web of Science records (Zhang et al., 2016) and MeSH Keywords assigned to PubMed publications (Turki et al., 2018a). The advantage of using these keywords is that they allow the use of a unique term and not of synonyms to assign each concept to publications and allow to prevent the redundancy of many variants of the same term across the maintained citation index allowing a more precise data mining and knowledge engineering of the bibliographic database (Henry and McInnes 2017; Turki et al., 2018a). That is why the co-occurrence analysis of controlled keywords, particularly MeSH Keywords, is nowadays used in semantic relation extraction and validation and provides a high accuracy rate for such an action (Henry and McInnes, 2017). Concerning MeSH Keywords, the recognition of a semantic relation can be achieved through the identification of the compatibility of the qualifiers of two significantly co-occurring keywords (e.g., Sofosbuvir/*therapeutic use* and Hepatitis C/*drug therapy*), of the complementarity of the qualifier of a keyword with the class of the heading of another largely co-occurring keyword [e.g., Sofosbuvir/*therapeutic use* and Hepatitis C (*disease*)], or of the association of the classes of the headings of two mainly linked keywords [e.g., Sofosbuvir [*drug*] and Hepatitis C (*disease*)] (Turki et al., 2018a).

Despite the fact that citations also have some shortcomings, they are currently recognized as major information in Scientometrics as they can provide important details about the impact of scholarly publications as well as the evolution of scientific outcomes (Zhai et al., 2018). That is why they can be useful for refining and enriching the outputs of information retrieval from research papers. As the majority of research publications is most likely to be cited by and co-cited with related papers dealing with the same topic, initially considered papers for constructing and validating a knowledge graph can be odd ones and should not be processed if they do not belong to the citation or co-citation network of the topic of the semantic resource (Turki, 2018). By contrast, the papers that have the best centrality in the citation or co-citation network of the field of the knowledge graph should be considered as reference resources that should drive the beginning and reasoning of the information retrieval algorithms as these publications are the main papers in the field upon which all other papers have been developed (Diallo et al., 2016).

Controlled keywords and citations can be combined together to provide an added value to knowledge graph creation and validation from scholarly publications. The sentence including an inline citation to a work can be a key for enriching information about the citing paper as well as the cited one (Aljaber et al., 2011). The controlled keywords and the title analysis of the cited paper can be used to enrich the semantics of the citing sentences and recognize a hidden scientific relation or entity that has been discussed without being clearly stated (Aljaber et al., 2011; Hadj Taieb et al., 2020). The analysis of the inline citation using automatic named entity annotation and scientific relation embedding can reveal controlled vocabularies and relations that are not originally used to describe the cited paper in bibliographic databases (Aljaber et al., 2011).

In another context, several types of semantic relations are available in particular sections of a scholarly publication (Turki et al., 2018a; Alexander and de Vries, 2021). For example, information about research funding for a given paper can only be found in specific parts (Alexander and de Vries, 2021). Alexander and de Vries (2021) note that the choice of an algorithm is important at the beginning of a research project. They use their algorithm to extract funding information from scholarly publications. The advantage of this text is that it is typically included under the section "Acknowledgments" or "Funding Information" in scholarly publications and adheres to certain writing standards, for instance, "This research is funded by'" followed by the name of the research funder, in some cases the funding program, and finally the grant number.

Similarly, the section titles in narrative literature reviews provide an outlook about the information included in each part where a section entitled "Symptoms" in a review about Hepatitis C includes semantic relations about the symptoms of the described medical condition (Turki et al., 2018a). Subsequently, considering section titles during semantic relation extraction and validation can not only reduce the complexity of the recognition of the relation types but also minimize the time allocated for such a task by restricting the process to the sections that are expected to include the required relations.

Other Metadata

What other metadata elements of scholarly publications can be considered when it comes to building knowledge graphs? Scholarly publications have different levels of evidence according to their settings, the age, the type, the status, the research area, and the source title of a given output. All these factors may influence the significance of its findings to the research community (Burns et al., 2011). The age of a research paper can typically determine whether the information included in it is outdated or not as terms and abbreviations might change over years due to nomenclature updates (AlRyalat et al., 2018) and as several findings can be disproven after a time period thanks to advances in experimentation techniques and scientific reasoning (Arbesman, 2013). Although science is an ever-evolving enterprise, it is also based on certain classical literature, for example, the importance of the founders of academic disciplines, such as sociology. Consequently, scientists have to be mainly based on new publications to create better and updated knowledge graphs about their topics of interest and even to predict the evolution of the constructed knowledge graph in the next years (Choudhury et al., 2020). The type of a given publication can affect the amount and quality of information it includes. When letters only present a limited number of facts in a few pages (Turki et al., 2018b), reviews provide a detailed overview of the concepts and findings related to a given topic from the synthesis of many papers and are consequently more adequate as resources for scholarly information retrieval (Burns et al., 2011; Turki et al., 2018a).

Acronyms used in scholarly literature can serve as a reliable goal of matching concepts in a research field. Many acronyms seem to be established terminology that is referred to frequently, in an unambiguous way. It has been shown that even for large corpuses of scientific papers from diverse fields automated

disambiguation can be applied unsupervized and at scale (Charbonnier and Wartena 2018; Veyseh et al., 2021).

The status of a research publication can be also important for ensuring the quality of the extraction of scientific knowledge. Although bibliographic databases like *PubMed*[2] and features like *Crossmark*[3] state whether a publication is a preprint or a partially or fully retracted paper, most of the projects for the creation and validation of knowledge graphs do not consider this factor when retrieving facts from research papers (Sargsyan et al., 2020). The matter with considering preprints in information retrieval is that these publications have not undergone peer review (Glasziou et al., 2020) and their outputs can be dynamically changed over months (Oikonomidi et al., 2020). That is why using them to generate structured information about a given topic can harm the quality of the created resource and make it less trustworthy. As for retracted publications, they are papers that have been proven to include false or fabricated claims, and were rejected by the scientific community and eliminated from their journal of publication for this reason. Although retractions are continuously cited for various reasons (e.g., lacking knowledge about the retractions), applying information retrieval techniques on them can let the users of the returned semantic data reuse scientifically doubtful findings and probably make wrong interpretations (Sotudeh et al., 2020; Soltani and Patini, 2020).

The restriction of the set of considered publications according to their research areas as revealed in citation indexes or through the analysis of author keywords allows refining the creation and validation of knowledge graphs by eliminating outputs outside the scope of the developed resource (Salatino et al., 2020). This prevents the overlapping of concepts from different fields when they are represented by the same polysemous term and consequently eliminates noise from the generated database. To be precise, a polysemous term has various meanings. The consideration of the research venue of publications can be also efficient in this context. Further than the ability to analyze source titles using semantic similarity measures, among other techniques, to verify the topics of interest of journals and conferences (Hadj Taieb et al., 2015), metrics about the venues such as the journal impact factor and the number of citations can be used to filter the considered sources and only consider the most prestigious and reliable ones (Pal, 2021). While the journal impact factor has shortcomings, it can be a useful indicator in this context.

CONCLUSION

In this opinion article, we showed the kinds of semantic information that can be revealed from each type of bibliographic metadata and that can be later used to strengthen information retrieval for knowledge graph construction and validation from scholarly publications. Given this, we invite the scientific community to collaborative projects considering bibliographic information when extracting domain information from scholarly publications for the creation and validation of trustworthy and precise semantic resources. As a future direction of this work, we suggest investigating how bibliographic metadata can enhance information retrieval algorithms using a series of experiments comparing the accuracy of methods processing full texts of scholarly publications with the one of bibliometric-enhanced information retrieval approaches. We consequently propose to study how bibliometric-enhanced information retrieval can enhance knowledge graph construction and validation as well as other interesting computational tasks such as predicting future scientific breakthroughs and major prize winners, natural language generation and translation of scholarly texts, and the automation of the creation and update of various kinds of research outputs. Researchers may also consider the adaptation of BIR algorithms to support the augmentation of university-level courses and evaluation quizzes with explanatory excerpts from scholarly outputs and to recommend scholarly publications to fight online misinformation. As well, we recommend building a framework for explainable artificial intelligence that returns explanations of the use of machine learning models for a given task based on what is currently available about the matter in research papers. As far as the availability of the data needed for BIR is concerned, it is to be hoped for the future that initiatives such as the I4OA mentioned above will gain momentum and that the applicability and re-usability of bibliographic metadata for BIR will become easier.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

The work of HT, MA, and MBA is supported by the Ministry of Higher Education and Scientific Research in Tunisia (MoHESR) in the framework of Federated Research Project PRFCOV19-D1-P1. The contribution of GF to this article is supported by the Train[2] Wind ITN that has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement number 861291.

ACKNOWLEDGMENTS

Significant parts of this research output have been generated through discussion with the organizers and participants of the International Workshop on Bibliometric-enhanced Information Retrieval at the European Conference on Information Retrieval in 2020 and 2021 (Cabanac et al., 2020).

[2]Further details can be found in the PubMed User Guide at https://pubmed.ncbi.nlm.nih.gov/help/. For PubMed, information about the status of a research publication can be automatically retrieved using Biopython (Chapman and Chang, 2000).

[3]*Crossmark* is a crowdsourcing tool and database developed by CrossRef Organization that allows to annotate the types and statuses of scholarly publications (Meyer, 2011).

REFERENCES

Alexander, D., and de Vries, A. P. (2021). "This Research Is Funded by: Named Entity Recognition of Financial Information in Research Papers.," in Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 43rd European Conference on Information Retrieval (ECIR 2021), Lucca, Italy, April 1, 2021 (Lucca, Italy: . CEUR Workshop Proceedings), 102–110.

Aljaber, B., Martinez, D., Stokes, N., and Bailey, J. (2011). Improving MeSH Classification of Biomedical Articles Using Citation Contexts. *J. Biomed. Inform.* 44 (5), 881–896. doi:10.1016/j.jbi.2011.05.007

AlRyalat, S. A., Rawashdeh, K., El khatib, O., Yasin, A., Alqadi, F., Saleh, N., et al. (2018). The Change from an Eponym to a Representative Name: Wegener to Granulomatosis with Polyangiitis. *Scientometrics* 117 (3), 2077–2089. doi:10.1007/s11192-018-2951-z

Anteghini, M., D'Souza, J., Martins. dos. Santos, V. A. P., and Auer, S. (2020). "Representing Semantified Biological Assays in the Open Research Knowledge Graph," in International Conference on Asian Digital Libraries, Kyoto, Japan, November 30–December 1, 2020 (Cham: . Springer), 89–98. doi:10.1007/978-3-030-64452-9_8

Arbesman, S. (2013). *The Half-Life of Facts: Why Everything We Know Has an Expiration Date* (New York: Penguin). ISBN:9781101595299.

Asada, M., Miwa, M., and Sasaki, Y. (2021). Using Drug Descriptions and Molecular Structures for Drug-Drug Interaction Extraction from Literature. *Bioinformatics*. doi:10.1093/bioinformatics/btaa907

Burns, P. B., Rohrich, R. J., and Chung, K. C. (2011). The Levels of Evidence and Their Role in Evidence-Based Medicine. *Plast. Reconstr. Surg.* 128 (1), 305–310. doi:10.1097/PRS.0b013e318219c171

Cabanac, G., Frommholz, I., and Mayr, P. (2020). "Bibliometric-Enhanced Information Retrieval 10th Anniversary Workshop Edition," in European Conference on Information Retrieval, Lisbon, Portugal, April 14–17, 2020 (Cham: . Springer), 641–647. doi:10.1007/978-3-030-45442-5_85

Cabanac, G., Frommholz, I., and Mayr, P. (2018). Bibliometric-enhanced Information Retrieval: Preface. *Scientometrics* 116 (2), 1225–1227. doi:10.1007/s11192-018-2861-0

Chapman, B., and Chang, J. (2000). Biopython. *SIGBIO Newsl.* 20 (2), 15–19. doi:10.1145/360262.360268

Charbonnier, J., and Wartena, C. (2018). "Using Word Embeddings for Unsupervised Acronym Disambiguation," in Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, August 20–26, 2018 (International Committee on Computational Linguistics), 2610–2619. doi:10.25968/opus-1265

Choudhury, N., Faisal, F., and Khushi, M. (2020). Mining Temporal Evolution of Knowledge Graphs and Genealogical Features for Literature-Based Discovery Prediction. *J. Informetrics* 14 (3), 101057. doi:10.1016/j.joi.2020.101057

Colavizza, G., Costas, R., Traag, V. A., van Eck, N. J., van Leeuwen, T., and Waltman, L. (2021). A Scientometric Overview of CORD-19. *PLoS One* 16 (1), e0244839. doi:10.1371/journal.pone.0244839

Diallo, S. Y., Lynch, C. J., Gore, R., and Padilla, J. J. (2016). Identifying Key Papers within a Journal via Network Centrality Measures. *Scientometrics* 107 (3), 1005–1020. doi:10.1007/s11192-016-1891-8

Fathalla, S., Vahdati, S., Auer, S., and Lange, C. (2017). "Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles," in International Conference on Theory and Practice of Digital Libraries, Thessaloniki, Greece, September 18–21, 2017 (Cham: . Springer), 315–327. doi:10.1007/978-3-319-67008-9_25

Fei, H., Zhang, Y., Ren, Y., and Ji, D. (2021). A Span-Graph Neural Model for Overlapping Entity Relation Extraction in Biomedical Texts. *Bioinformatics*. doi:10.1093/bioinformatics/btaa993

Fuhr, N., and Rölleke, T. (1998). "HySpirit - A Probabilistic Inference Engine for Hypermedia Retrieval in Large Databases," in International Conference on Extending Database Technology, Valencia, Spain, March 23–27, 1998 (Berlin, Heidelberg: . Springer), 24–38. doi:10.1007/BFb0100975

Glasziou, P. P., Sanders, S., and Hoffmann, T. (2020). Waste in Covid-19 Research. *BMJ* 369, m1847. doi:10.1136/bmj.m1847

Hadj Taieb, M. A., Ben Aouicha, M., and Bourouis, Y. (2015). "Fm3s: Features-Based Measure of Sentences Semantic Similarity," in International Conference on Hybrid Artificial Intelligence Systems, Bilbao, Spain, June 22–24, 2015 (Cham: . Springer), 515–529. doi:10.1007/978-3-319-19644-2_43

Hadj Taieb, M. A., Ben Aouicha, M., and Turki, H. (2019). "Paper Co-citation Analysis Using Semantic Similarity Measures," in International Conference on Intelligent Systems Design and Applications, December 3–5, 2019 (Cham: . Springer), 264–277. doi:10.1007/978-3-030-49342-4_26

Hadj Taieb, M. A., Ben Aouicha, M., and Turki, H. (2020). Semantic-driven Bibliometric Techniques for Co-citation Analysis. *His* 16 (2), 111–125. doi:10.3233/HIS-200288

Haslhofer, B., Isaac, A., and Simon, R. (2018). *Knowledge Graphs in the Libraries and Digital Humanities Domain Encyclopedia of Big Data Technologies*. Cham: Springer, 1–8. doi:10.1007/978-3-319-63962-8_291-1

Henry, S., and McInnes, B. T. (2017). Literature Based Discovery: Models, Methods, and Trends. *J. Biomed. Inform.* 74, 20–32. doi:10.1016/j.jbi.2017.08.011

Jeong, Y. K., Song, M., and Ding, Y. (2014). Content-based Author Co-citation Analysis. *J. Informetrics* 8 (1), 197–211. doi:10.1016/j.joi.2013.12.001

Lebuda, I., and Karwowski, M. (2016). Written on the Writer's Face: Facial Width-To-Height Ratio Among Nominees and Laureates of the Nobel Prize in Literature. *Creativity Res. J.* 28 (2), 207–211. doi:10.1080/10400419.2016.1162572

Letchford, A., Moat, H. S., and Preis, T. (2015). The Advantage of Short Paper Titles. *R. Soc. Open Sci.* 2, 150266. doi:10.1098/rsos.150266

Meyer, C. A. (2011). Distinguishing Published Scholarly Content with CrossMark. *Learned Publishing* 24 (2), 87–93. Distinguishing published scholarly content with CrossMark, Learned Publishing. doi:10.1087/20110202

Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2017). "Scholia, Scientometrics and Wikidata," in European Semantic Web Conference, Portorož, Slovenia, May 28–June 1, 2017 (Cham: . Springer), 237–259. doi:10.1007/978-3-319-70407-4_36

Oikonomidi, T., Boutron, I., Boutron, I., Pierre, O., Cabanac, G., and Ravaud, P. (2020). Changes in Evidence for Studies Assessing Interventions for COVID-19 Reported in Preprints: Meta-Research Study. *BMC Med.* 18 (1), 402. doi:10.1186/s12916-020-01880-8

Pal, J. K. (2021). Visualizing the Knowledge Outburst in Global Research on COVID-19. *Scientometrics* 126, 4173–4193. doi:10.1007/s11192-021-03912-3

Ramos, J. (2003). "Using Tf-Idf to Determine Word Relevance in Document Queries," in Proceedings of the first instructional conference on machine learning, Washington, DC, August 21–24, 2003 (AAAI), 242, 29–48.

Ripple, A. M., Mork, J. G., Knecht, L. S., and Humphreys, B. L. (2011). A Retrospective Cohort Study of Structured Abstracts in MEDLINE, 1992-2006. *J. Med. Libr. Assoc.* 99 (2), 160–163. doi:10.3163/1536-5050.99.2.009

Safder, I., and Hassan, S.-U. (2019). Bibliometric-enhanced Information Retrieval: a Novel Deep Feature Engineering Approach for Algorithm Searching from Full-Text Publications. *Scientometrics* 119 (1), 257–277. doi:10.1007/s11192-019-03025-y

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., and Motta, E. (2020). The Computer Science Ontology: A Comprehensive Automatically-Generated Taxonomy of Research Areas. *Data Intelligence* 2 (3), 379–416. doi:10.1162/dint_a_00055

Sargsyan, A., Kodamullil, A. T., Baksi, S., Darms, J., Madan, S., Gebel, S., et al. (2020). The COVID-19 Ontology. *Bioinformatics* 36 (24), 5703–5705. doi:10.1093/bioinformatics/btaa1057

Servant, N., RomÃ©jon, J., Gestraud, P., La Rosa, P., Lucotte, G., Lair, S. v., et al. (2014). Bioinformatics for Precision Medicine in Oncology: Principles and Application to the SHIVA Clinical Trial. *Front. Genet.* 5, 152. doi:10.3389/fgene.2014.00152

Soltani, P., and Patini, R. (2020). Retracted COVID-19 Articles: a Side-Effect of the Hot Race to Publication. *Scientometrics* 125 (1), 819–822. doi:10.1007/s11192-020-03661-9

Sotudeh, H., Barahmand, N., Yousefi, Z., and Yaghtin, M. (2020). How Do Academia and Society React to Erroneous or Deceitful Claims? the Case of Retracted Articles' Recognition. *J. Inf. Sci.*, 016555152094585. doi:10.1177/0165551520945853

Stotesbury, H. (2003). Evaluation in Research Article Abstracts in the Narrative and Hard Sciences. *J. English Acad. Purposes* 2 (4), 327–341. doi:10.1016/S1475-1585(03)00049-3

Tay, A., Kramer, B., and Waltman, L. (2020). Why Openly Available Abstracts Are Important – Overview of the Current State of Affairs. Available at: https://leidenmadtrics.nl/articles/why-openly-available-abstracts-are-important-overview-of-the-current-state-of-affairs (Accessed May 1, 2021). Leiden Madtrics.

Telford, M., Mooi, R., and Ellers, O. (1985). A New Model of Podial deposit Feeding in the Sand Dollar, Mellita Quinqujesperforata (Leske): the Sieve Hypothesis Challenged. *Biol. Bull.* 169 (2), 431–448. doi:10.2307/1541493

Turki, H. (2018). Citation Analysis Is Also Useful to Assess the Eligibility of Biomedical Research Works for Inclusion in Living Systematic Reviews. *J. Clin. Epidemiol.* 97, 124–125. doi:10.1016/j.jclinepi.2017.11.002

Turki, H., Hadj Taieb, M. A., and Ben Aouicha, M. (2021a). Enhancing Filter-Based Parenthetic Abbreviation Extraction Methods. *J. Am. Med. Inform. Assoc.* 28 (3), 668–669. doi:10.1093/jamia/ocaa314

Turki, H., Hadj Taieb, M. A., and Ben Aouicha, M. (2018a). MeSH Qualifiers, Publication Types and Relation Occurrence Frequency Are Also Useful for a Better Sentence-Level Extraction of Biomedical Relations. *J. Biomed. Inform.* 83, 217–218. doi:10.1016/j.jbi.2018.05.011

Turki, H., Hadj Taieb, M. A., and Ben Aouicha, M. (2018b). The Value of Letters to the Editor. *Scientometrics* 117 (2), 1285–1287. doi:10.1007/s11192-018-2906-4

Turki, H., Hadj Taieb, M. A., Shafee, T., Lubiana, T., Jemielniak, D., Ben Aouicha, M., et al. (2021c). Representing COVID-19 Information in Collaborative Knowledge Graphs: the Case of Wikidata. *Semantic Web J.*

Turki, H., Taieb, M. A. H., and Aouicha, M. B. (2021b). Developing Intuitive and Explainable Algorithms through Inspiration from Human Physiology and Computational Biology. *Brief. Bioinform.* doi:10.1093/bib/bbab081

Veyseh, A. P. B., Dernoncourt, F., Nguyen, T. H., Chang, W., and Celi, L. A. (2021). "Acronym Identification and Disambiguation Shared Tasks for Scientific Document Understanding," in Proceedings of the 1st workshop on Scientific Document, February 8–9, 2021 (AAAI). Understanding at AAAI-21.

Wang, L. L., and Lo, K. (2021). Text Mining Approaches for Dealing with the Rapidly Expanding Literature on COVID-19. *Brief. Bioinform.*, 22 (2), 781–799. doi:10.1093/bib/bbaa296

White, H. D. (2018). Bag of Works Retrieval: TF*IDF Weighting of Works Co-cited with a Seed. *Int. J. Digit Libr.* 19 (2), 139–149. doi:10.1007/s00799-017-0217-7

Zhai, Y., Ding, Y., and Wang, F. (2018). Measuring the Diffusion of an Innovation: a Citation Analysis. *J. Assoc. Inf. Sci. Tech.* 69 (3), 368–379. doi:10.1002/asi.23898

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z., and Duan, Z. (2016). Comparing Keywords Plus of WOS and Author Keywords: A Case Study of Patient Adherence Research. *J. Assn Inf. Sci. Tec* 67 (4), 967–972. doi:10.1002/asi.23437

Zhou, W., Torvik, V. I., and Smalheiser, N. R. (2006). ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics* 22 (22), 2813–2818. doi:10.1093/bioinformatics/btl480

Zou, Y., Finin, T., and Chen, H. (2004). "F-owl: An Inference Engine for Semantic Web," in International Workshop on Formal Approaches to Agent-Based Systems, Greenbelt, MD, April 26–27, 2004 (Berlin, Heidelberg: Springer), 238–248. doi:10.1007/978-3-540-30960-4_16

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.