Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 06 Issue 02
2024

Research Article
https://doi.org/10.56979/602/2024

A Survey on Semantic Representations of Citation Data

Imran Ihsan[1*], and Fawad Salam Khan[1]

[1]Department of Creative Technologies, Air University, Islamabad, 44000, Pakistan.
[*]Corresponding Author: Imran Ihsan. Email: iimranihsan@gmail.com

Received: December 11, 2023 **Accepted:** February 09, 2024 **Published:** March 01, 2024

Abstract: This survey explores the semantic representations of citation data, emphasizing the critical distinction between citations and references in scholarly communication. References denote works listed in a bibliography, while citations refer to individual instances where these works are contextualized within the text. Recent advancements in OWL2 ontologies have facilitated the formal, machine-interpretable encoding of bibliographic and citation data, alongside document components and individual citation characteristics. However, effective automated processing of citation data requires comprehensive machine-readable metadata and ontologies to encapsulate these elements. This study identifies key challenges, including the development of semantic models that accurately represent citation reasons. While existing semantic-based publishing applications provide customized data retrieval, they often neglect the nuances of citation reasons, focusing primarily on basic metadata such as authorship and affiliations. Through a comprehensive examination of various ontologies, this survey assesses their ability to capture citation reasons in a semantically meaningful way and highlights limitations that impede the effective representation of citation data.

Keywords: Citation Reasons; Ontology; Semantics.

1. Introduction

A distinction exists between a citation and a reference [1]. A reference refers to the works listed in the bibliography or reference section of a journal article and may appear one or multiple times within the text. Each instance in the text is a citation, providing context for the referenced work. Advances in OWL2 ontologies [2] have made it possible to formally encode bibliographic and citation data, document components, and the nature of individual citations in a machine-interpretable format. Numerous studies have explored encoding various components of scientific papers, including citations, in a formalized manner.

Automated processing of bibliographic and citations' data also requires machine-interpretable metadata for publications and citations while the ontologies are required to encode these metadata elements [3]. Some of the areas that require answers are:

- Development and adoption of semantic models (ontologies) that permit bibliographic and citations' reasons data in machine-interpretable form, is the core requirement in scholarly authoring and publishing.
- Development of annotation tools to help the authors to enhance the semantic relation of their documents with others, using appropriate semantic assertions for the citations.

Semantic-based publishing applications provide customization of data and the content to reflect the user's needs of retrieval of relevant data with minimal effort. Using the new set of OWL2 ontologies [2], bibliographic and citation data, document components, and the nature of individual citations can be structured. However, existing applications do not follow the basic principle of semantic-based publishing as defined by Peroni and Shotton [3]. Our study reveals that such applications use metadata elements such as Authors and their affiliations, editors and their affiliation, publishing companies, etc., and do not look for the citation reasons. To understand semantic-based citation reasons, we have conducted a survey to examine various available ontologies to find whether they provide options to record the citations reasons in a semantically meaningful away (minimal set of disjoint reasons) and what are their limitations.

2. Semantic Representations

2.1. Bibliographic Reference Ontology

The Bibliographic Reference Ontology (BiRo) [4] is an ontology meant to define bibliographic records, bibliographic references, and their compilation into bibliographic collections and bibliographic lists, respectively as shown in Table 1. Based on FRBR [5] it describes individual bibliographic reference and its relationship to the cited article using two properties; "is referenced by" and "reference" with domain and range as "endeavor" and "bibliographic record" alternatively. It is clear from the meanings that both properties neither define the nature of the relationship between the papers nor citation reasons.

Table 1. Bibliographic Reference Ontology.

No	Properties	No	Properties
1	is referenced by	2	references

2.2. Citation Counting and Context Characterization Ontology

The Citation Counting and Context Characterization Ontology (C4O) [4] is an ontology that permits the number of in-text citations of a cited source to be recorded, together with their textual citation contexts, along with the number of citations a cited entity has received globally on a particular date as shown in Table 2. It keeps track of the number of citations that a paper has received using all possible external sources. The ontology claims to record the "context of citation", however, this is an in-text reference pointer of where the citation has been made. Its "has context" property provides the place where a possible rhetorical motivation for citation exists in a paper but does not exploit the context for possible motivation or reasons for citation.

Table 2. Citation Counting and Context Characterization Ontology.

No	Properties	No	Properties
1	denotes	7	pertains to
2	has context	8	has content
3	has global citation frequency	9	has global count date
4	has global count source	10	has global count value
5	is denoted by	11	has in text citation frequency
6	is relevant to		

2.3. FRBR-aligned Bibliographic Ontology

The FRBR-aligned Bibliographic Ontology (FaBio) [3] is an ontology for describing entities that are published or potentially publishable (e.g., journal articles, conference papers, books), and that contain or are referred to by bibliographic references as shown in Table 3. It mainly records publications such as books, magazines, journals, and their content like algorithms, specifications, vocabulary or technical reports, that are published or in the process of being published, using semantic web descriptions. It is based on FRBR [5] data model to interlink manifestations, items, and expressions and does not deal with the nature of links between them (citations).

Table 3. FRBR-aligned Bibliographic Ontology.

No	Properties	No	Properties
1	has creator	16	has rights
2	has discipline	17	has subject term
3	has embodiment	18	is discipline of
4	has exemplar	19	is embodiment of
5	has format	20	is exemplar of
6	has language	21	is in scheme
7	has license	22	is manifestation of
8	has manifestation	23	is part of
9	has part	24	is portrayal of
10	has place of publication	25	is realization of

11	has portrayal	26	is representation of
12	has primary subject term	27	is scheme of
13	has publisher	28	is stored on
14	has realization	29	stores
15	has representation		

2.4. Document Component Ontology

Document Components Ontology (DoCO) [6] in an ontology that provides a structured vocabulary written of document components, both structural (e.g., block, inline, paragraph, section, chapter) and rhetorical (e.g., introduction, discussion, acknowledgements, reference list, figure, appendix) as shown in Table 4. It decomposes a research paper document into its structural and rhetorical components such as Abstract, Introduction, Results, Conclusion, and Bibliography, etc., and stores these components using RDF - Resource Document Framework. The nature of a citation is the discourse element of a research paper and this ontology does not deal with it.

Table 4. Document Component Ontology.

No	Properties	No	Properties
1	contains	2	is contained by

2.5. Publishing Role Ontology

Publishing Roles Ontology (PRO) [7] is an ontology for the characterization of the roles of agents - people, corporate bodies and computational agents in the publication process as shown in Table 5. These agents can be, e.g. authors, editors, reviewers, publishers or librarians. It also records the time when a role asserts. However, it does not deal with the citation or its nature.

Table 5. Publishing Role Ontology.

No	Properties	No	Properties
1	at time	8	is role in
2	holds role in time	9	relates to
3	is document context for	10	relates to document
4	is organization context for	11	relates to organization
5	is person context for	12	relates to person
6	is related to role in time	13	with role
7	is role held by		

2.6. Publishing Status Ontology

Publishing Status Ontology (PSO) [8] is designed to characterize the publication status of documents at each stage of the publishing process (draft, submitted, under review, etc.) as shown in Table 6. It also records the duration the document took to transit from one status to another and the people involved during that. This ontology also does not deal with citations.

Table 5. Publishing Status Ontology.

No	Properties	No	Properties
1	at time	6	is status in
2	holds status in time	7	results in acquiring
3	is acquired as consequence of	8	results in losing
4	is lost as consequence of	9	with status
5	is status held by		

2.7. SWAN – Discourse Ontology

SWAN 1.0 Discourse Ontology [9] is designed to create an ecosystem that can create, store, access, integrate and exchange semantic context of scientific papers especially in the field of Neuro-medicine and specifically Alzheimer Disease (AD) and is shown in Table 7. The ontology stores a research statement with three possible discourse elements: “citeAsEvidence", “citeLifeScienceEntity" and “citesReagent". These discourse elements relate to each other using a set of relationships that are “discusses", “refutes", “supports" and “alternativeTo". The ontology uses standard biological concepts [9] such as “genes", “proteins", “reagents" etc to assert scientific discourse. Therefore, the ontology works fine in its intended domain but is not helpful in other domains. However, a smaller set of discourse elements provided by the ontology is helpful for the annotators.

Table 7. SWAN – Discourse Ontology.

No	Properties	No	Properties
1	cites As Supportive Evidence	3	refers To
2	research Statement Qualied As		

2.8. Citation Typing Ontology

Citation Typing Ontology (CiTO) [3] is an ontology that enables characterization of nature or type of citations, both factually and rhetorically as shown in Table 8. CiTO asserts and characterizes bibliographic references and citations. Citations have three characteristics "direct and explicit", "indirect", and "implicit". Based on biomedical researchers, the ontology describes citation nature in terms of the "Factual" and "Rhetorical" relationships and subdivides them between "Positive", "Negative" and "Neutral". In total, there are 41 properties and are known as CiTO-Ps. A study [10] has been conducted to cluster these properties that exhibit similar meanings according to the subject's annotation using the Chinese Whispers clustering algorithm [11]. The results show that a certain collection of properties show diffused and overlapped meanings.

3. Experiment and Results

By examining the above ontologies, it becomes clear that the ontology that comes closest to our research goals is CiTO. It defines the nature of citations for intelligent linking and reasoning. However, the characterizations defined by CiTO are very difficult for humans to understand and adopt. Using Ciancarini [10] we have summarized some problems in it after a careful analysis of both experimental data and subjects' feedback. Based on these experiments, some of the limitations in CiTO are:

3.1. Less Used Properties

Several properties defined in CiTO-Ps remain underutilized, particularly those expressing negative citation contexts. For instance, properties that denote adverse relationships, such as "disagreesWith," "disputes," "parodies," "plagiarizes," "refutes," "repliesTo," and "ridicules," appear significantly less frequently than their neutral or positive counterparts [12].

Table 8. Citation Typing Ontology.

No	Properties	No	Properties
1	agrees with	21	disagrees with
2	citation	22	discusses
3	cites	23	disputes
4	cites as authority	24	documents
5	cites as data source	25	extends
6	cites as evidence	26	gives background to
7	cites as metadata document	27	gives support to
8	cites as potential solution	28	likes
9	cites as recommended reading	29	parodies
10	cites as related	30	plagiarizes
11	cites as source document	31	refutes
12	cites for information	32	replies to
13	compiles	33	retracts
14	confirms	34	reviews
15	contains assertion from	35	ridicules
16	corrects	36	speculates on
17	credits	37	supports
18	critiques	38	updates
19	derides	39	uses conclusions from
20	describes	40	uses data from
		41	uses method in

3.2. Most Used Neutral Properties

Certain properties, such as “citesForInformation” and “citesAsRelated,” span across various scholarly domains and are among the most frequently used, likely due to their neutral stance. Ciancarini [10] found that these two properties are commonly applied, even in cases where more specific options, like “citesAsAuthority,” “citesAsDataSource,” and “discusses,” could provide greater precision.

3.3. Lower Inter-Rater Agreement

CiTO provides 41 properties for defining and annotating citation reasons, yet utilizing this full set demands substantial cognitive effort. An experiment [10] comparing T41, which includes all CiTO properties, and T10, a subset limited to 10 properties, demonstrated that the reduced set significantly enhances usability for citation annotation among professors, academic researchers, postdoctoral fellows, and Ph.D. students.

3.4. Non-Taxonomic Organization of CiTO-Ps

CiTO lacks a taxonomic organization; instead, each property is mapped individually based on a mental model. Some CiTO properties exhibit similar conceptual structures, suggesting they could be clustered under broader, parent properties for improved organization.

3.5. Customized Properties

CiTO currently lacks support for customization. When an annotator cannot find a perfectly suitable property, they often choose the closest match to their mental model. The latest CiTO release [3] addresses this limitation by making the ontology structure (i.e., the TBox) static, while allowing users flexibility to precisely express specific characterizations, capturing nuanced details and tones.

3.6. Misinterpretation of Properties

Certain properties within CiTO are often misunderstood or interpreted inconsistently by users, highlighting a clear need for improvements [10].

3.7. Properties Perspective

CiTO properties are designed to align with the annotator's perspective rather than the author’s. For instance, properties like “disagreesWith,” “disputes,” “parodies,” “plagiarizes,” “refutes,” “repliesTo,” and “ridicules” are intended for annotators' use rather than by authors themselves. Allowing authors to define citation reasons directly could result in more semantically accurate and meaningful annotations, as they are best positioned to specify the intent behind each citation.

4. Discussion

With the advent of Knowledge Graphs, research has increasingly focused on storing scientific data in large-scale RDF formats. One notable initiative is the Microsoft Academic Knowledge Graph (MAKG) [13], which encompasses a vast volume of 8 billion triples and is accessible through the Linked Open Data Cloud. However, the effective use of MAKG necessitates the adaptation of various ontologies to accurately encode different components of research articles. For references, MAKG employs the CiTO ontology to model citation information. Nevertheless, due to the coarse granularity of the 41 properties within CiTO, MAKG primarily utilizes a single entity type, cito:citation, while leaving the other properties underutilized. Despite this limitation, MAKG recognizes the importance of citation context for each reference, as it is valuable for tasks such as citation recommendation and citation-based paper summarization [13]. Consequently, there is an urgent need to develop a minimal set of cognitive-based citation contexts and reasons in the form of a dedicated ontology.

5. Conclusion

A scientific research paper contains vital information that prompts its citation by authors and researchers for various reasons. These citation reasons are crucial for uncovering cognitive relationships between research papers. Automated processing of citation data necessitates a formal and semantic definition of these reasons. This study reveals that while numerous attempts have been made to record citations in a semantically meaningful manner through ontologies, most of these do not adequately address citation reasons. One ontology, the Citation Type Ontology (CiTO), offers a formal semantic definition of citation reasons; however, it has several limitations. These include the prevalence of less frequently used neutral properties, lower inter-rater agreement, non-taxonomic organization of properties, misinterpretation of properties, and a focus on the annotator's perspective rather than that of the author. Overall, CiTO reflects the viewpoint of annotators, which may not align with the authors' intentions.

Funding: No Funding.

Conflicts of Interest: “The authors declare no conflict of interest.”

References

1. G. N. Gilbert and S. Woolgar, "The quantitative study of science: An examination of the literature," Social Studies of Science, vol. 4, no. 3, pp.279--294, 1974.
2. B. Motik, B. Parsia, and P. F. Patel-Schneider, OWL 2 Web Ontology Language XML Serialization (Second Edition). W3C Recommendation, 2012.
3. FaBiO and CiTO: Ontologies for describing bibliographic resources and citations," Journal of Web Semantics, vol. 17, pp. 33--43, 2012.
4. A. Di Iorio, A. G. Nuzzolese, S. Peroni, D. Shotton, and F. Vitali, "Describing bibliographic references in RDF," in Proceedings of 4th Workshop on Semantic Publishing - SePublica, vol. 1155. CEUR, 2014, pp. 41--56.
5. B. Tillett, "What is FRBR? A conceptual model for the bibliographic universe," Australian Library Journal, vol. 54, no. 1, pp. 24--30, 2005.
6. A. Constantin, S. Peroni, S. Pettifer, D. Shotton, and F. Vitali, "The Document Components Ontology (DoCO)," Semantic Web, vol. 7, no. 2, pp. 167--181, 2016.
7. Ahmad, R., Salahuddin, H., Rehman, A. U., Rehman, A., Shafiq, M. U., Tahir, M. A., & Afzal, M. S. (2024). Enhancing Database Security through AI-Based Intrusion Detection System. Journal of Computing & Biomedical Informatics, 7(02).
8. PRO – Publishing Role Ontology. Available Online: https://sparontologies.github.io/pro/current/pro.html
9. PSO – Publishing Status Ontology. Available Online: https://sparontologies.github.io/pso/current/pso.html
10. P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, "The SWAN biomedical discourse ontology," Journal of Biomedical Informatics, vol. 41, no. 5, pp. 739{751, 2008.
11. P. Ciancarini, A. Di Iorio, A. G. Nuzzolese, S. Peroni, and F. Vitali, "Evaluating citation functions in CiTO: Cognitive issues," Lecture Notes in Computer Science, vol. 8465, no. LNCS, pp. 580--594, 2014.
12. C. Biemann, "Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems," in Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing. Association for Computational Linguistics, 2006, pp. 73--80.
13. S. Teufel, a. Siddharthan, and D. Tidhar, "An annotation scheme for citation function," in Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. ACM, 2006, pp. 80--87.
14. F. Michael, "The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data," The Semantic Web - ISWC, vol. 11779, pp. 113--129, 2019.
15. Khan, R., Iltaf, N., Shafiq, M. U., & Rehman, F. U. (2023, December). Metadata based Cross-Domain Recommender Framework using Neighborhood Mapping. In 2023 International Conference on Sustainable Technology and Engineering (i-COSTE) (pp. 1-8). IEEE.
16. Shafiq, M. U., & Butt, A. I. (2024). Segmentation of Brain MRI Using U-Net: Innovations in Medical Image Processing. Journal of Computational Informatics & Business, 1(1).