

SUDO: From Scientific Publications to Global Scholarly Discourse

Nipun D. Pathirage¹, Henrique Santos¹ and Deborah L. McGuinness¹

¹*Rensselaer Polytechnic Institute, Troy, NY, USA*

Abstract

Scholarly Knowledge Graphs (SKGs) support literature search, scientific discovery, and agent-assisted research. Yet, most approaches to SKG construction pre-specify the knowledge unit to be represented, typically named scholarly entities or contribution-level statements, and retain only the publication content expressible through that unit. This choice produces two dominant yet distinct paradigms: concept-centric SKGs that preserve entities and local relations, and statement-centric SKGs that capture high-level discourse in sentences. In both cases, the graph reflects a chosen extraction target rather than how a paper constructs and communicates knowledge. We introduce SUDO, the Scholarly Upper Discourse Ontology, which defines how knowledge elements are organized and related to form a paper’s research narrative, namely its scholarly discourse. Using linguistic principles that generalize across fields and tasks, SUDO enables each publication’s scholarly discourse to be constructed from the bottom up as a coherent, source-grounded representation. Applied consistently across papers, the same structure connects their research narratives into a global scholarly discourse without discarding local context or provenance. We demonstrate SUDO’s capabilities across a comprehensive suite of scholarly knowledge-modeling use cases and performance improvements on a downstream Large Language Model (LLM)-based scientific QA task.

Keywords

Scholarly Knowledge Representation, Ontology Engineering, Computational Linguistics,

1. Introduction

Scholarly knowledge representation seeks to make scientific publications machine-actionable. Scholarly Knowledge Graphs (SKGs) operationalize this goal, supporting literature search [1, 2], knowledge discovery [3], scientific question answering [4], and, increasingly, LLM- and agent-assisted scientific workflows [5, 6]. Existing SKGs largely follow two dominant paradigms: concept-centric [7, 8] and statement-centric modeling [9, 6, 10]. Both structures select aspects of publication knowledge, enabling scalable indexing, retrieval, comparison, and aggregation across large scholarly corpora.

Concept-centric SKGs represent scholarly knowledge through named entities and research artifacts, capturing their semantic categories and relations while leaving their roles within a publication’s local context underspecified [1, 7, 8, 11]. For example, when the word “attention” appears in a paper, a concept-centric SKG such as CS-KG [11] may represent it as an instance of the Method category, but without capturing its role in that paper, whether as part of the proposed approach, a theoretical motivation, or a baseline. As a result, attention is represented uniformly across papers, obscuring the differences in how it is applied, adapted, or interpreted in each context. These are differences that are crucial for understanding how the concept develops and shifts across the literature, a central aspect of scholarly discourse.

Statement-centric SKGs provide a complementary view by representing higher-level scholarly assertions (e.g., research problems, approaches, evidence, and claims) and relating them to reconstruct a publication’s research argument [9, 10, 12]. Yet this abstraction creates three limitations. First, statements provide weak anchors for integration across publications, which typically depends on recurring concepts and shared scholarly nomenclature [1]. Second, because these ontologies predefine the scholarly aspects and statement types they capture, they represent only selected parts of a paper,

6th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Oct 2026, Bari, Italy

✉ pathin@rpi.edu (N. D. Pathirage); oliveh@rpi.edu (H. Santos); dlm@rpi.edu (D. L. McGuinness)

id 0000-0002-9313-0925 (N. D. Pathirage); 0000-0002-2110-6416 (H. Santos); 0000-0001-7037-4567 (D. L. McGuinness)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

omitting local details from methodology and evaluation as well as explicit cross-paper discourse in related work [6]. More fundamentally, their statement types are designed to model scholarly functions, not to correspond to directly expressed units in the source. Instantiating them may therefore require combining information distributed across several passages or deriving a higher-level abstraction not stated in any one location [6]. This weakens source grounding and is particularly problematic for LLM-based knowledge graph (KG) construction, where plausible statement-level descriptions can be generated without clearly identifiable textual support.

These limitations reveal a broader representational gap in which existing SKGs specify in advance which knowledge should be extracted, rather than how scholarly knowledge is built within a publication. Within a paper, knowledge is organized from the bottom up: established scholarly nomenclature anchors the work in prior knowledge; clauses and sentences use and relate that nomenclature, introduce new terms, and express pragmatics; and their progression across sentences and sections develops the paper’s research contribution. This organization constitutes its local scholarly discourse. Modeling that discourse preserves both the structural context and source grounding of knowledge before it is transformed into domain- or task-specific abstractions.

We operationalize this bottom-up view in SUDO, the Scholarly Upper Discourse Ontology. It draws on the theoretical linguistics of Jackendoff’s conceptual semantics, developed in the 1980s, which models how linguistic expressions map to structured conceptual representations of meaning [13]. SUDO extends this insight from individual expressions to scholarly publications, taking each publication’s lexical, sentential, and discourse organization as the basis of its knowledge representation: concepts are introduced and related within local linguistic contexts; discourse relations organize these expressions into the publication’s research argument; and recurring concepts connect publication-level arguments across the literature to form global scholarly discourse. Because these structures reflect general properties of language rather than field- or task-specific extraction categories, SUDO provides a representation that can be applied across domains and tasks.

We evaluate SUDO by constructing an SKG covering the International Semantic Web Conference’s (ISWC) main-track publications from 2012 to 2025, assessing its support for a comprehensive collection of scholarly knowledge-modeling use cases, and demonstrating performance gains on downstream LLM-based scientific question answering. This work contributes a new computational-linguistic view of scholarly knowledge representation, its realization in the SUDO ontology, and the resulting ISWC scholarly knowledge graph.

2. Related Work

2.1. The Data-Source Paradigm and Information Loss

Formal scholarly knowledge representation has developed over nearly three decades, with increasing emphasis on scalable construction and access. Early systems such as ScholOnto [14] represented scholarly claims through expressive, manually constructed ontologies. Although these representations captured rich scholarly semantics, their reliance on author or expert input limited their adoption at scale. Subsequent work therefore shifted toward automated and centralized scholarly infrastructures, including ArnetMiner [15], AMiner [16], the Microsoft Academic Graph (MAG) [17], and OpenAlex [2]. By integrating metadata, citation links, author identities, and topical indicators across large publication collections, these web-scale SKGs provided substantial value for literature discovery, filtering, bibliometric analysis, and scholarly search.

Advances in named-entity recognition and relation extraction enabled concept-centric SKGs, such as Semantic Scholar [1], AI-KG [7], and CS-KG [8, 11], to be constructed automatically over web-scale publication collections and integrated into infrastructures such as MAG [18] and AMiner [16]. By linking recurring concepts and relations across publications, these graphs naturally support cross-publication knowledge exploration. However, they provide limited information about the role a concept plays within a specific paper. For example, whether it denotes a research problem, method, experimental variable, or prior work. This limits search and analysis that depend on publication-specific meaning

and argumentative context, and makes concept-centric representations less compatible with LLM-based workflows that require connected, context-rich information.

Statement-centric SKGs were introduced to represent this publication-level information more explicitly. Systems such as ORKG [10], SemSur [9], and NLP-AKG [6] model research contributions, experimental descriptions, or other scholarly information as natural-language statements. Compared with concept-centric graphs, these representations retain more of the meaning conveyed in the source publication and are therefore better suited to describing what a study investigates, proposes, observes, or concludes. Their construction, however, is more complex, particularly when statements must be extracted or generated automatically from information distributed across a paper. Consequently, statement-centric SKGs have not yet been widely adopted across web-scale publication collections.

Existing statement-centric SKGs typically represent only selected dimensions of scholarly content, such as research arguments [19], contributions [10, 6], or experiments [9]. They therefore provide only partial coverage of a publication, commonly model statements as isolated units without preserving the discourse relations that connect information across the paper, and, particularly in LLM-based construction, must ensure that generated statements remain faithful to and traceable to their source passages. These limitations have constrained their adoption in scientific question answering [20, 21, 4], knowledge discovery [22, 23, 24, 25], and literature-review workflows [26], which continue to rely largely on full-text representations even when operating over web-scale publication collections. To bridge this gap, we introduce SUDO as a source-grounded representation for structurally modeling complete publications, connecting their statements through scholarly discourse, and supporting LLM-based workflows over web-scale scholarly knowledge graphs.

2.2. Speech Acts and the Positioning of SUDO

Linguistic theories have long examined how human expressions encode meaning and relate to one another within larger communicative structures [27, 28]. Rhetorical Structure Theory (RST), for example, models relations between discourse units and distinguishes their functional roles, such as central statements (*nuclei*) and supporting statements (*satellites*) [29]. Broader discourse theories extend these local relations to explain how individual statements collectively form coherent discourse [28, 30], while theories of illocution and commitment characterize the communicative intent an utterance expresses [31, 27, 32]. Building on these foundations, Speech Act Theory provides a well-established framework for analyzing how utterance-level intentions contribute to conversation progression [33].

We bring these established linguistic foundations into scholarly knowledge modeling to address the limitations discussed above. SUDO is designed to structurally represent an entire publication while preserving its broader discourse organization. From this perspective, the global body of scientific literature can be understood as a longitudinal, multi-party dialogue. Within this dialogue, an individual paper functions as a complex linguistic artifact (a formal scholarly utterance) that responds to, builds upon, or challenges prior work through citations and shared conceptual vocabularies.

3. Conceptual semantic view of scholarly knowledge

We use the example in Fig. 1 to illustrate the linguistic view adopted in this work. Consider the following sentence from the abstract of Attention Is All You Need [34]: “*We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolution entirely.*” A reader recognizes that this sentence proposes Transformer as a new method, identifies it as a type of network architecture, and states that it is based on attention mechanisms rather than recurrence or convolution. The preceding sentence, “*The best performing models also connect the encoder and decoder through an attention mechanism.*” plays a different role. It provides background for attention mechanism applications. Human readers infer these distinctions from linguistic cues expressed through noun phrases, clauses, lexical choices, and relations between sentences. Computational-linguistic theories such as Rhetorical Structure Theory [29] and discourse-structure models [28] seek to make such relations explicit.

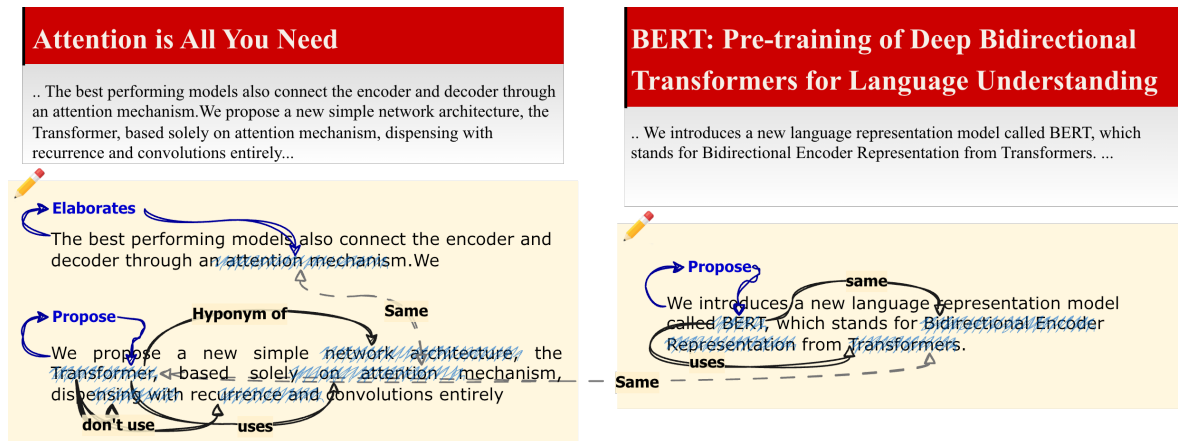


Figure 1: Illustration of how humans understand a publication’s meaning and its relationships to other scholarly works from its linguistic structure.

This linguistic view provides a basis for modeling both local and global scholarly discourse. Consider a sentence from the abstract of the BERT paper [35]: “We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers.” Like the previous example, this sentence propose BERT as a new method. However, Transformer plays a different role here: it describes BERT’s architectural basis rather than the principal contribution being introduced. Thus, the role of a concept within a publication is determined by how it participates in a clause and how that clause contributes to the broader discourse of the paper. Relations between concepts and clauses, and between clauses and publication-level functions, constitute the local discourse of a publication. Recurring concepts such as Transformer then connect these local discourse structures across papers, forming a global scholarly discourse.

This linguistic view allows the modeling of an entire publication by following how information is presented throughout the text. For example, when Transformer is introduced in one sentence and described further in later methodological sentences, the recurring named entity connects these distant clauses: the first proposes the approach, while the later clauses elaborate it with additional detail. The scattered information is thereby assembled into a coherent account of the approach without requiring a predefined contribution template. Importantly, lexical form varies far more than the linguistic properties that signal a scholarly function. An issue statement may be worded in many ways, but the cues marking it as the presentation of a research problem remain comparatively uniform. The same regularity holds for ideas, approaches, and claims, allowing any number of them to be represented consistently without prescribing how they must be expressed. Because each modeled element remains linked to the source text that expresses it, the resulting representation provides a fully source-grounded model of scholarly discourse.

Jackendoff’s **conceptual semantics** provides the theoretical mechanism for making this linguistic view computational [13]. It treats meaning as a structured conceptual representation assembled from a finite inventory of primitives and rules of combination. Lexical and syntactic forms remain distinct from this representation but are connected to it through explicit interface rules, allowing varied surface expressions to map to comparable linguistic structures. Within linguistic structure, **propositional structure** encodes predicate–argument content (roughly, who did what to whom) whereas **information structure** records how that content is situated communicatively, including distinctions such as topic and focus or given and new information. We adapt this architecture in SUDO: lexical and clausal analysis recovers scholarly objects and propositions, while discourse analysis assigns their scholarly functions and relations. These mappings operationalize the preceding linguistic view as a compositional knowledge representation constructed directly from publication text.

4. SUDO

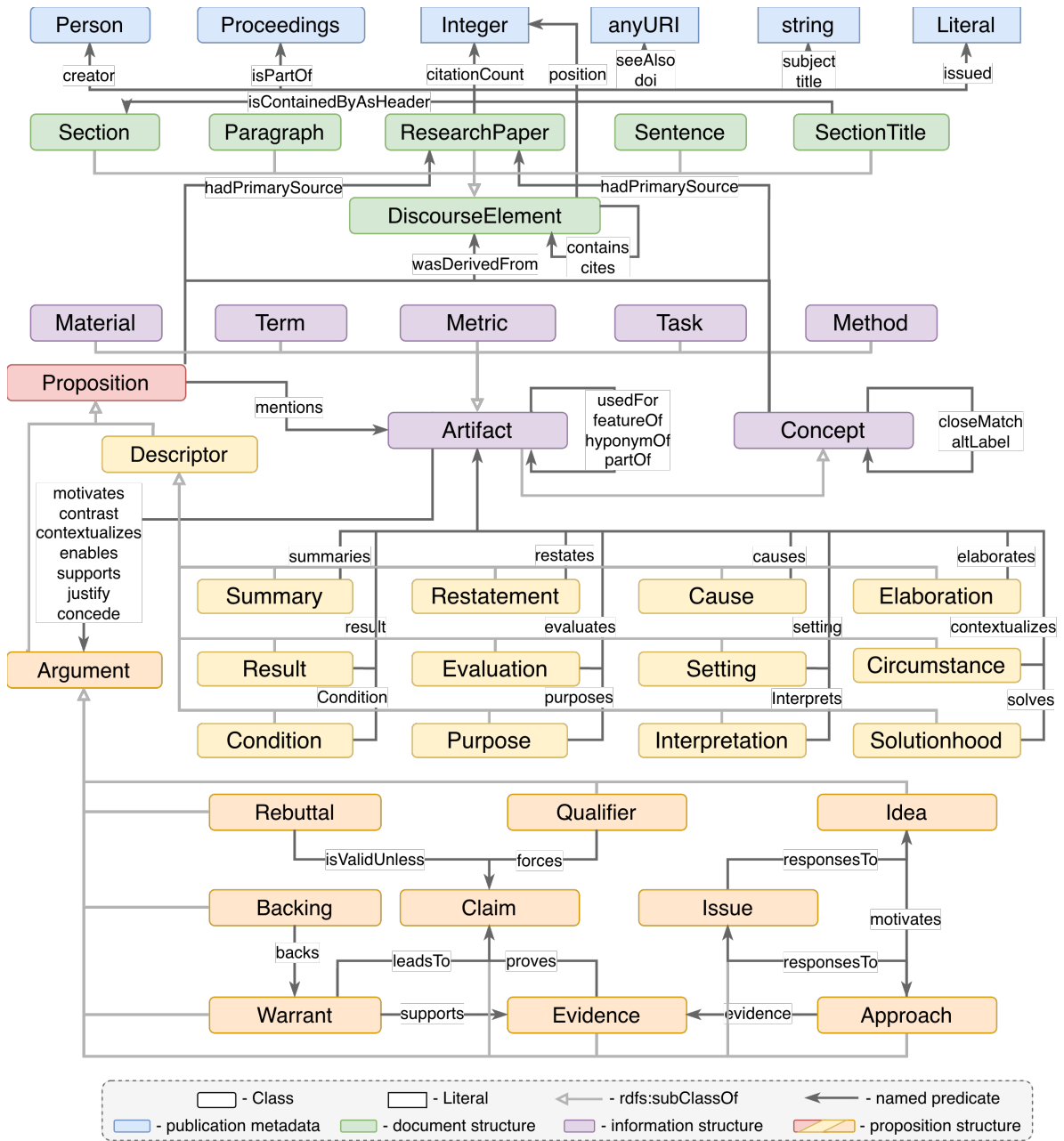


Figure 2: SUDO ontology diagram. A.4

4.1. Overview

The **Scholarly Upper Discourse Ontology** (SUDO) formalizes the conceptual semantic view of scholarly knowledge introduced in Section 3. As illustrated in Figure 2, SUDO represents a publication through its complementary *information structure* and *propositional structure*, together with the metadata and document structure required to locate this knowledge within the source publication. The propositional structure is further divided into **Arguments** and **Descriptors**. Arguments capture propositions that constitute the paper’s research argument, whereas Descriptors capture propositions that describe, elaborate, or contextualize the scholarly concepts involved. This distinction draws on Jackendoff (2002) [36] work on propositional segregation and Simone Teufel’s argumentative-zoning

theories [37, 38, 39], which differentiate statements according to their roles in scholarly communication.

The following subsections describe how SUDO models and materializes the information and propositional structures in detail. We then show how these complementary structures are integrated to form the complete conceptual semantic representation of a scholarly publication.

4.2. Information structure

In SUDO, the information structure models the scholarly nomenclature through which a publication establishes what it discusses and organizes its propositions. It therefore represents noun-phrase expressions as named scholarly entities and captures the semantic relations among them. To remain applicable across fields, these entities are classified using broad scholarly semantic categories rather than field-specific subject matter. For this purpose, SUDO adapts the cross-domain SciERC scientific named entity recognition (NER) schema, using **Task**, **Method**, **Metric**, **Material**, and **Term**. Relations are defined at the same level of generality: **used-for**, **feature-of**, **hyponym-of**, **part-of**, **compare**, and **conjunction** capture functional, compositional, and taxonomic associations [Fig. 2]. Crucially, SUDO distinguishes a nomenclature’s textual realization from its canonical identity. Each noun-phrase occurrence is represented as an **Artifact**, preserving its source location, lexical form, abbreviation or coreferential expression, and context-specific semantic category. The corresponding **Concept** represents the canonical identity shared across occurrences. Artifact instances can subsequently be reconciled with Concepts through mappings such as **closeMatch**, while lexical alternatives are retained through **altLabel**. This separation preserves publication-specific usage while supporting concept alignment across the literature. Further definitions of the entity and relation types are provided in Section A.3.

Figure 3 materializes this design through the running Transformer and BERT example in Figure 1. The occurrences *The Transformer in Attention Is All You Need* and *Transformers* in the BERT paper are represented as separate Method Artifacts linked to their respective source sentences. These lexical realizations remain distinct in their publication contexts, while **closeMatch** maps both to the canonical Transformer Concept. Their semantic relations contribute, through this canonical alignment, to a shared cross-publication concept structure. Because the Artifacts remain distinct, each can also participate in the Approach proposition that captures its role within the corresponding publication. The information structure thereby connects recurring nomenclature across papers without erasing the lexical and contextual distinctions required by SUDO’s propositional structure.

4.3. Propositional Structure

The **propositional structure** models how a publication turns the scholarly objects in its information structure into a research narrative. Each **Proposition** is anchored to a source clause and linked to the **Artifacts** that participate in it, preserving what the publication states about each locally realized concept. Some propositions perform the principal scholarly acts of the paper. For example, posing an issue, proposing an approach, presenting evidence, or making a claim, establishing argumentative backbones. Other propositions supply the semantic detail through which those acts are contextualized, qualified, or summarized, connecting locally expressed detail. This separation aligns with argumentative-zoning theories, which distinguish scholarly text by its rhetorical function within the publication [37, 39]. Together, they represent both the publication-level research argument and the source-grounded information from which it is constructed.

Argument Structure. The **Argument** structure models the propositions and relations that form a publication’s argumentative backbone and develop its research contribution. Indirect argumentation is a natural fit because this backbone follows the scientific method: an **Issue** motivates an **Idea** or **Approach**, its evaluation produces **Evidence**, and **Warrants** connect that Evidence to resulting **Claims**. The justificatory roles of Evidence, Warrant, and Claim follow Toulmin’s model of practical argumentation [40], as formalized in the Argument Model Ontology (AMO) [41]. SUDO complements

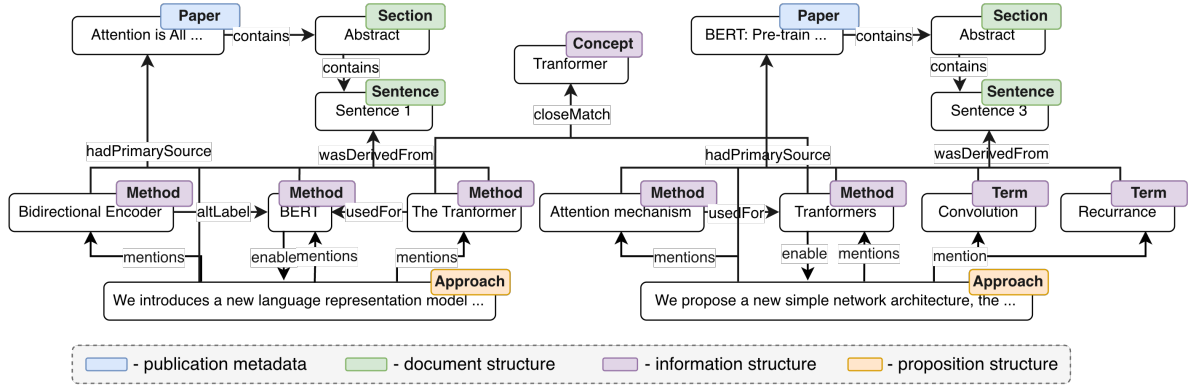


Figure 3: SUDO realization of the scenario introduced in Fig. 1, emphasizing the constructed information and propositional structures while illustrating how their elements remain grounded in the source publications through document structure and metadata.

them with the Issue, Idea, and Approach roles used by SWiM/SIOC to adapt argumentation modeling to scientific discourse [19]. Rhetorical Structure Theory and discourse-structure models provide the linguistic basis for recognizing how these functions and their connections are expressed in text [29, 28]. Together, these elements relate propositions through **responds-to**, **motivates**, **supports**, and **leads-to** [Figure 2], allowing each contribution to form its own argument structure within the publication.

In Figure 3, the proposal in *Attention Is All You Need* forms an **Approach** proposition centered on Transformer. The introductory statement in the BERT paper forms a separate Approach centered on BERT, with Transformer appearing as its architectural basis. Although the two Transformer Artifacts align with the same canonical Concept, their proposition-level connections preserve these different argumentative roles. Connecting each Approach to its associated Issues, Evidence, Warrants, and Claims then reconstructs the distinct research argument each publication develops.

Descriptor Structure. The **Descriptor** structure models propositions that supply semantic detail about an Artifact or another Proposition. Rhetorical Structure Theory provides a natural framework by classifying interactions between rhetorical elements according to how they contribute to the overall statement [29]. In *subject-matter relations*, a rhetorical statement supplies semantic detail about another element; in *presentational relations*, a rhetorical statement uses another element to advance its communicative purpose. Descriptors therefore correspond to RST’s subject-matter relations. RST defines fourteen such relations, of which *sequence*, *contrast*, and *joint* are multinuclear and may involve more than two discourse units. Representing these relations through binary RDF/OWL object properties would require an additional n-ary reification pattern [42]. SUDO consequently adopts the remaining eleven binary relations, realizing each as a Descriptor type paired with a property that links the Descriptor to the Artifact or Proposition it describes [Figure 2].

Returning to the running example, later methodological clauses describing the architecture or internal layers of Transformer become Elaboration Descriptors. Through their shared Transformer Artifact, these Descriptors attach the dispersed methodological detail to the Approach that introduced the method, without treating each description as a separate contribution. Because every Descriptor corresponds to the clause that expresses it, the resulting detail remains source-grounded while contributing to the publication’s broader discourse.

4.4. Conceptual Semantic View of Global Scholarly Discourse

SUDO constructs a publication’s local scholarly discourse by joining the structures defined above. Artifact instances and their semantic relations form its publication-specific information structure. **Argument** propositions organize **Issue**, **Idea**, **Approach**, **Evidence**, **Warrant**, and **Claim** acts into one or more argumentative backbones, while **Descriptor** propositions attach semantic detail to those

arguments and their Artifacts. Artifact–Proposition links bind the two structures by recording which locally realized nomenclature participates in each proposition. Global scholarly discourse is then constructed through canonical **Concepts**: `closeMatch` aligns recurring Artifacts across publications, connecting their surrounding proposition structures without merging the source-specific Artifacts themselves.

Concept graphs can encode only relations anticipated by their predicate vocabulary. When a publication expresses a more nuanced connection, the graph must either omit it or reduce it to a generic predicate, making semantically distinct relations indistinguishable and losing how the concepts are actually connected in the source. Alongside discrete Concept–Concept relations, SUDO links Artifacts to the source Propositions in which they participate. These Propositions preserve the specific meaning underlying generalized Concept relations and capture connections for which no discrete relation is instantiated. Because this meaning remains accessible through the Proposition’s linguistic representation in latent semantic space, the Proposition acts as a semantic reification of the relation, a behavior we call **latent semantic reification**.

4.5. Implementation

SUDO was developed following the Ontology Development 101 methodology [43] and has undergone three major revisions across 19 update cycles. Version 3.5.2 is the current stable release, with backward compatibility maintained throughout the 3.x series. SUDO extensively reuses established ontologies for publication metadata, document structure, and provenance, as detailed in Section A.4. Publication metadata is represented in its most general reusable form, allowing derivative ontologies to introduce more specialized models without losing information during mapping. The ontology is version-controlled through Git and published under the persistent W3ID namespace <https://w3id.org/twc/sudo/> in RDF/XML, OWL/XML, Turtle, N-Triples, and JSON-LD. Each release includes WIDOCO-generated documentation [44], maintained competency questions, and an OOPS! evaluation for common ontology-modeling pitfalls [45].

4.6. User Case Scenario

To demonstrate SUDO’s utility for scholarly knowledge modeling, we construct a knowledge graph from 898 ISWC main-track papers published between 2012 and 2025, including their authorship, publication year, and track metadata. For each publication, triples describing publication metadata, document structure, local scholarly discourse, and provenance are stored in four dedicated named graphs, with provenance statements linking them into a unified publication representation. The complete ISWC corpus and resulting knowledge graph are publicly available through the resources listed in Section A.2. Graph construction uses a neuro-symbolic pipeline whose methodology is treated separately, allowing this paper to focus on SUDO’s ontology design and evaluation.¹ The competency-question evaluation in ?? is conducted over this graph, with the complete question set provided in the appendix.

Using the same pipeline, we additionally construct and publish a SUDO knowledge graph for 413 publications from the QASPER benchmark [21]. This graph supports the LLM-based usability evaluation described in Section 5.2.

5. Evaluation

Because SUDO is intended both to structurally represent scholarly knowledge and to support its practical use, its evaluation must extend beyond any single dataset or task. We therefore combine two complementary forms of assessment: a use-case-based analysis of representational coverage and a downstream evaluation of practical utility in LLM-assisted scientific question answering. These assessments address the following research questions:

¹The construction pipeline is described in a separate manuscript currently under review; its reference will be added once that work is published.

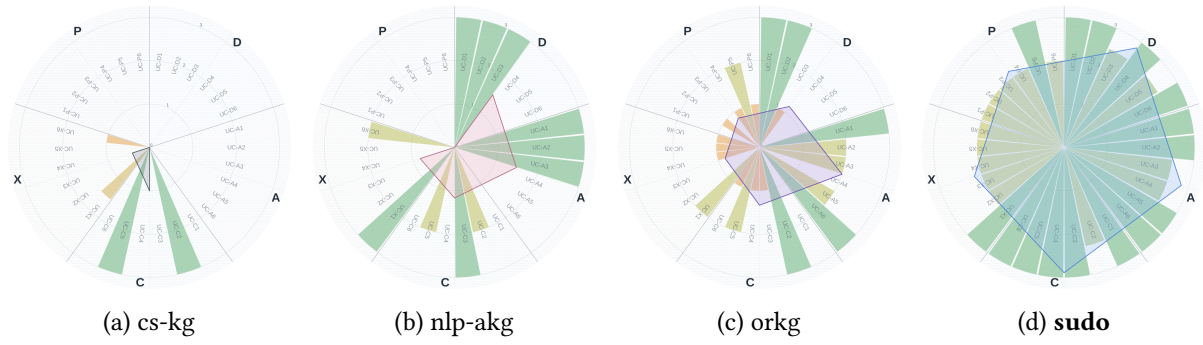


Figure 4: Comparative use-case coverage across SKGs. Nightingale rose sectors show individual use-case satisfiability (S_{UC}) on a four-level scale (0–3), with no bar for Level 0 and red, yellow, and green denoting Levels 1, 2, and 3, respectively; the overlaid radar plot shows family-level coverage computed from the mean S_{UC} . Sector labels denote the family IDs in Table 3, with use cases ordered clockwise from 1 to 6 within each family.

Table 1

Answer Capability Levels (ACLs) used to evaluate competency-question answerability.

ACL	Capability	Interpretation
3	Direct resolution	The answer is explicitly represented and directly retrievable.
2	Computable resolution	The graph contains sufficient structure to derive the answer through traversal, rules, or deterministic computation.
1	Structural retrieval	The graph identifies a constrained set of source material from which the answer can be obtained through additional human or language-model processing.
0	No support	The schema or populated graph lacks the information required to assist with the question.

- **RQ1:** To what extent does SUDO structurally represent scholarly knowledge throughout a publication compared with existing SKG models?
- **RQ2:** Does SUDO’s discourse-aware representation improve information access in LLM-based scientific question-answering workflows?

For each research question, we first define the corresponding evaluation objective and procedure, then report results and discuss the evidence-supported conclusions.

5.1. RQ1: Structural Representation Coverage

Competency questions are the established means of testing whether an ontology or populated knowledge graph can satisfy its intended information needs. Because they are formulated for a particular schema and dataset, however, competency-question collections do not readily generalize across SKGs or support direct coverage comparisons. RQ1 therefore requires a shared evaluation framework rather than a SUDO-specific question set. We introduce a collection of ontology-independent scholarly knowledge-modeling use cases, each of which can be operationalized as a schema-specific competency question for every SKG under evaluation. This preserves a common information requirement while allowing each model to answer it through its own representational structure. We derived the collection by examining nearly two decades of scholarly knowledge-modeling research and published use cases, identifying five recurring families: *publication metadata and document structure*, *paper argument and discourse*, *concepts and paper-local use*, *cross-publication concept dynamics*, and *paper-local pragmatic descriptions*. Each family contains six use cases, yielding 30 in total. The complete collection is provided in Table 3.

Competency-question evaluations are typically binary, but binary answerability understates information coverage in SKGs. An answer may be explicitly represented in the graph or implicitly supported when the graph preserves and retrieves the source information from which a downstream computa-

tional component can derive it. We therefore introduce four *Answer Capability Levels* (ACLs), ranging from no support (0) to direct resolution (3), as defined in Table 1. Because a single use case may be operationalized through multiple competency questions, following Ontology Development 101 [43], we define its cumulative satisfiability as the mean ACL across its question set Q_U :

$$S_{UC}(U) = \frac{1}{|Q_U|} \sum_{q_i \in Q_U} ACL(q_i). \quad (1)$$

Averaging retains the 0–3 ACL scale, allowing the cumulative use-case score to preserve the same interpretation while supporting fair comparison across SKGs. Family-level coverage is obtained by macro-averaging the six use-case scores within each family.

5.1.1. Experimental Setup

We constructed model-specific competency-question sets for four SKGs under the proposed use-case collection. CS-KG [8] represents concept-centric modeling, while ORKG [10] and NLP-AKG [6] represent statement-centric modeling through manual and automated population, respectively. These models cover the dominant SKG representation units and construction strategies. The CS-KG² and ORKG³ questions were evaluated against their publicly available knowledge-graph dumps. Because no comparable NLP-AKG dump was available, we generated one by applying its implementation to the ISWC publication collection used for SUDO. The resulting competency questions were evaluated using the ACL procedure defined above.

5.1.2. Results and Discussion

As shown in Figure 4, SUDO achieves the highest and most balanced coverage, with most use cases reaching ACL 2 or ACL 3 and substantial support maintained across all five families. CS-KG reaches ACL 3 for several explicitly modeled concept relations but falls to ACL 0 across most document-structure, argument, and pragmatic-description requirements, producing narrow but direct coverage. ORKG and NLP-AKG broaden coverage to contribution- and statement-oriented use cases, particularly those concerning claims and research contributions. Both reach ACL 3 when these elements are explicitly modeled, but decline to ACL 0 or ACL 1 when answering requires explicit connections among statements, paper-local concept roles, detailed document structure, or cross-publication context. ORKG retains intermediate support through its manually instantiated contribution templates, whereas NLP-AKG provides direct support where a requirement matches one of its predefined statement types but preserves little relevant information outside them.

SUDO’s stronger ACL profile follows from how its structures complement one another. Publication metadata and document structure directly resolve their corresponding requirements; Argument and Descriptor propositions capture research narratives and pragmatic detail; and the distinction between Artifacts and canonical Concepts supports both contextualized concept use and cross-publication connections. These structures produce ACL 3 when the requested answer is explicitly represented, ACL 2 when it can be derived from graph structure, and ACL 1 when source-grounded propositions preserve the information required for subsequent interpretation. SUDO’s coverage therefore reflects not only a broader explicit vocabulary but also its ability to retain answer-bearing information when no dedicated graph relation exists. These results support RQ1 by showing that SUDO represents scholarly knowledge more comprehensively across the full publication than the evaluated concept- or statement-centric models.

5.2. RQ2: Utility for Scholarly Downstream Tasks

RQ2 examines whether SUDO’s discourse-aware representation improves access to answer-bearing information in scientific question answering, specifically in LLM-based scientific workflows. We

²<https://zenodo.org/records/14167682>

³<https://orkg.org/data>

Table 2

QASPER performance comparison across retrieval modalities and knowledge-graph systems.

Modality	Model	Test Size	Precision	Recall	F1
Text [6]	BM25	200	67.94	73.86	70.84
	Embedding retrieval	200	68.29	73.83	70.69
	GPT-4	200	67.09	72.20	69.29
NLP-AKG (Graph) [6]	KAG	200	68.73	73.87	70.97
	MindMap	200	69.38	74.91	71.73
	NLP-AKG (Agent)	200	69.43	75.50	72.04
CS-KG	–	–	–	–	–
ORKG	–	–	–	–	–
SUDO-KG	Embedding retrieval	600	75.47	81.11	78.16
	gpt-oss-20b	600	82.57	82.39	82.44

evaluate 600 question–answer pairs from the QASPER dataset [21], spanning 413 scientific publications and grounded in their source content. Because QASPER provides document-level questions with varying complexity, it suits evaluating scientific evidence retrieval without conflating results with open-ended agent planning. A SUDO knowledge graph is constructed for these publications using the same neuro-symbolic pipeline applied to the ISWC collection. We evaluate the retrieved information using BERTScore [35] precision, recall, and F1: precision reflects the exclusion of unrelated content, recall reflects the retention of answer-relevant information, and F1 balances both.

5.2.1. Results and Discussion

As reported in Table 2, SUDO’s embedding-based retrieval achieves an F1 score of 78.16, exceeding the optimized NLP-AKG agent by approximately six percentage points. Query-guided retrieval further increases the F1 score to 82.44, with BERTScore precision and recall of 82.57 and 82.39, respectively, improving by about 10 percentage points over NLP-AKG. These results suggest that SUDO’s publication-local concepts and propositions provide more focused access to answer-bearing information than coarser statement units. The additional gain from query-guided retrieval further indicates that language models can use SUDO’s typed structure to refine information access.

Directly comparable scores could not be established for CS-KG or ORKG under the same evaluation protocol. CS-KG yielded too few reliable question-to-concept alignments for aggregate evaluation, reflecting the limited contextual information available from entity-level retrieval for publication-grounded questions. The available ORKG graph covered only 34 of the 413 QASPER publications, making a corpus-level comparison unrepresentative. These constraints reflect different modeling priorities: CS-KG emphasizes scalable concept-level organization, whereas ORKG provides richer contribution-level representations through manual curation, which constrains comprehensive population across large corpora. SUDO combines these strengths through scalable concept organization and source-grounded propositions, demonstrating the value of publication-centric, discourse-aware modeling for large-scale scientific question answering.

6. Future Work

Future work will investigate SUDO as an intermediary representation between publication text and specialized SKGs. Rather than extracting information directly into a predefined domain- or task-specific schema, a construction pipeline can first represent the publication through SUDO and derive specialized representations from that shared layer. Because the derived knowledge remains connected to SUDO’s Artifacts, Propositions, source locations, and provenance, specialized SKGs can reorganize or extend it without forfeiting source grounding. Concept-centric models can align their entities with SUDO

Concepts while inheriting the publication-local roles captured by associated Artifacts and Propositions. Statement-centric models can construct higher-level claims and contributions over SUDO’s propositional structure while retaining traceable connections to the expressions from which they are derived. Future work will implement and evaluate the ontology alignments, identity mappings, and transformation pipelines required to materialize these complementary views.

This layered architecture also enables downstream applications that depend on how knowledge is expressed and connected. Within publications, it can support evidence tracing, limitation analysis, and claim verification; across publications, it can enable contribution comparison, method-lineage analysis, and investigation of how concepts and claims develop through the literature. For LLM-assisted workflows, SUDO can provide structured, source-grounded contexts for scientific question answering, knowledge discovery, literature-based hypothesis generation, and scholarly analysis. We will investigate domain- and task-specific extensions for these applications and evaluate their construction and use across larger scholarly collections. The broader goal is to enable scholarly systems to operate not only over extracted information, but also over how that information is introduced, justified, related, and propagated.

7. Conclusion

In this paper, we introduced SUDO, a domain-agnostic, discourse-aware ontology for representing scholarly knowledge through the complementary information and propositional structures of conceptual semantics. SUDO connects scholarly concepts, source-grounded propositions, research arguments, descriptive context, document structure, and provenance, preserving both how knowledge is communicated within a publication and how recurring concepts connect that knowledge across the literature to build global scholarly discourse structure. Evaluation across comprehensive scholarly knowledge-modeling use cases showed broader representational coverage than the examined concept- and statement-centric SKGs, while the QASPER experiment demonstrated a measurable advantage for retrieving focused, answer-bearing information in an LLM-based scientific question-answering workflow. By combining scalable concept organization with semantically rich propositions and precise source grounding, SUDO provides a shared foundation for specialized SKG extensions, cross-publication analysis, and future web-scale scholarly AI workflows.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5.5 and Gemini in order to: Grammar and spelling checks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- [1] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, et al., Construction of the literature graph in semantic scholar, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, 2018, pp. 84–91.
- [2] J. Priem, H. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, *arXiv preprint arXiv:2205.01833* (2022).
- [3] S. Henry, B. T. McInnes, Literature based discovery: models, methods, and trends, *Journal of biomedical informatics* 74 (2017) 20–32.
- [4] Y. Wan, Y. Liu, A. Ajith, C. Grazian, B. Hoex, W. Zhang, C. Kit, T. Xie, I. Foster, Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation, *arXiv preprint arXiv:2405.09939* (2024).
- [5] B. Mo, K. Yu, J. Kazdan, P. Mpala, L. Yu, C. Kanatsoulis, S. Koyejo, Kggen: Extracting knowledge graphs from plain text with language models, *Advances in Neural Information Processing Systems* 38 (2026) 30092–30115.
- [6] J. Lan, J. Li, B. Wang, M. Liu, D. Wu, S. Wang, B. Qin, Nlp-akg: Few-shot construction of nlp academic knowledge graph based on llm, *arXiv preprint arXiv:2502.14192* (2025).
- [7] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, H. Sack, Ai-kg: an automatically generated knowledge graph of artificial intelligence, in: *International semantic web conference*, Springer, 2020, pp. 127–143.
- [8] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg: A large-scale knowledge graph of research entities and claims in computer science, in: *International Semantic Web Conference*, Springer, 2022, pp. 678–696.
- [9] S. Fathalla, S. Vahdati, S. Auer, C. Lange, Semsur: a core ontology for the semantic representation of research findings, *Procedia computer science* 137 (2018) 151–162.
- [10] M. Y. Jaradeh, A. Oelen, M. Prinz, M. Stocker, S. Auer, Open research knowledge graph: a system walkthrough, in: *International Conference on Theory and Practice of Digital Libraries*, Springer, 2019, pp. 348–351.
- [11] D. Dessì, F. Osborne, D. Buscaldi, D. Reforgiato Recupero, E. Motta, Cs-kg 2.0: A large-scale knowledge graph of computer science, *Scientific Data* 12 (2025) 964.
- [12] J. Chan, M. Akamatsu, D. Vargas, L. Kawerau, M. Gartner, Steps towards an infrastructure for scholarly synthesis, *arXiv preprint arXiv:2407.20666* (2024).
- [13] C. Maienborn, K. von Heusinger, P. Portner (Eds.), *Semantics - Theories*, Mouton Reader, De Gruyter Mouton, Berlin, Boston, 2019. doi:10.1515/9783110589245.
- [14] S. Buckingham Shum, E. Motta, J. Domingue, Scholonto: an ontology-based digital library server for research documents and discourse, *International Journal on Digital Libraries* 3 (2000) 237–248.
- [15] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: extraction and mining of academic social networks, in: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 990–998.
- [16] Y. Zhang, F. Zhang, P. Yao, J. Tang, Name disambiguation in aminer: Clustering, maintenance, and human in the loop., in: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1002–1011.
- [17] M. Färber, The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data, in: *International semantic web conference*, Springer, 2019, pp. 113–129.
- [18] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft academic graph: When experts are not enough, *Quantitative Science Studies* 1 (2020) 396–413.
- [19] C. Lange, T. Hastrup, S. Corlosquet, Arguing on issues with mathematical knowledge items in a semantic wiki., in: *LWA*, 2008, pp. 41–48.
- [20] S. Pramanick, R. Chellappa, S. Venugopalan, Spiqa: A dataset for multimodal question answering on scientific papers, *Advances in Neural Information Processing Systems* 37 (2024) 118807–118833.

- [21] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, M. Gardner, A dataset of information-seeking questions and answers anchored in research papers, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 4599–4610.
- [22] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, D. Ha, The ai scientist: Towards fully automated open-ended scientific discovery, *arXiv preprint arXiv:2408.06292* (2024).
- [23] Y. Yamada, R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune, D. Ha, The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, *arXiv preprint arXiv:2504.08066* (2025).
- [24] A. Novikov, N. Vű, M. Eisenberger, E. Dupont, P.-S. Huang, A. Z. Wagner, S. Shirobokov, B. Kozlovskii, F. J. Ruiz, A. Mehrabian, et al., Alphaevolve: A coding agent for scientific and algorithmic discovery, *arXiv preprint arXiv:2506.13131* (2025).
- [25] A. H. Artilles, M. Weiss, L. Brinkmann, I. Rahwan, B. Schölkopf, C. Pal, H. Larochelle, A. Goyal, N. Rahaman, The alien space of science: Sampling coherent but cognitively unavailable research directions, *arXiv preprint arXiv:2603.01092* (2026).
- [26] A. Ajith, M. Xia, A. Chevalier, T. Goyal, D. Chen, T. Gao, Litsearch: A retrieval benchmark for scientific literature search, in: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 15068–15083.
- [27] P. Grice, *Studies in the Way of Words*, Harvard University Press, 1991.
- [28] J. L. Austin, *How to do things with words*, volume 5, Harvard university press, 1975.
- [29] W. C. Mann, S. A. Thompson, *Rhetorical structure theory: A theory of text organization*, Technical Report, University of Southern California, Information Sciences Institute Los Angeles, 1987.
- [30] A. Lascarides, N. Asher, *Segmented discourse representation theory: Dynamic semantics with discourse structure*, in: *Computing meaning*, Springer, 2007, pp. 87–124.
- [31] J. R. Searle, D. Vanderveken, *Foundations of illocutionary logic*, CUP Archive, 1985.
- [32] R. B. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, 4th ed., Harvard University Press, Cambridge, Mass., 1994.
- [33] M. Sbisà, K. Turner (Eds.), *Pragmatics of Speech Actions*, De Gruyter Mouton, Berlin, Boston, 2013. URL: <https://doi.org/10.1515/9783110214383>. doi:10.1515/9783110214383.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [35] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (long and short papers), 2019, pp. 4171–4186.
- [36] R. Jackendoff, *Précis of foundations of language: Brain, meaning, grammar, evolution*, *Behavioral and brain sciences* 26 (2003) 651–665.
- [37] I. Zelch, M. Hagen, B. Stein, J. Kiesel, Segmentation of Argumentative Texts by Key Statements for Argument Mining from the Web, in: E. Chistova, P. Cimiano, S. Haddadan, G. Lapesa, R. Ruiz-Dolz (Eds.), *Proceedings of the 12th Argument Mining Workshop*, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 228–242. doi:10.18653/v1/2025.argmining-1.22.
- [38] S. Teufel, et al., *Argumentative zoning: Information extraction from scientific text*, Ph.D. thesis, University of Edinburgh Edinburgh, 1999.
- [39] S. Teufel, A. Siddharthan, C. Batchelor, Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics, in: P. Koehn, R. Mihalcea (Eds.), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2009, pp. 1493–1502.
- [40] J. Karbach, Using toulmin’s model of argumentation, *Journal of Teaching Writing* 6 (1987) 81–92.
- [41] S. Peroni, F. Vitali, The Argument Model Ontology (AMO), <https://sparontologies.github.io/amo/current/amo.html>, 2011. Accessed: 2026-09-16.
- [42] W. W. W. Consortium, et al., Owl 2 web ontology language document overview (2012).
- [43] E. F. Kendall, D. L. McGuinness, *Ontology engineering*, Morgan & Claypool Publishers, 2019.
- [44] D. Garijo, Widoco: a wizard for documenting ontologies, in: *International Semantic Web*

Conference, Springer, Cham, 2017, pp. 94–102. URL: <http://dgarijo.com/papers/widoco-iswc2017.pdf>. doi:10.1007/978-3-319-68204-4_9.

- [45] M. Poveda-Villalón, A. Gómez-Pérez, M. C. Suárez-Figueroa, OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation, *International Journal on Semantic Web and Information Systems (IJSWIS)* 10 (2014) 7–34.

A. Appendices

A.1. Scholarly Publication Knowledge-Representation Use Cases

Table 3: Scholarly Publication Knowledge-Representation Use Cases

ID	Title	Description
Family D: Publication Metadata and Document Structure		
UC-D1	Identity	Which bibliographic attributes, such as title, DOI, publication date, venue, publication type, and license, identify a given publication?
UC-D2	Agent Role	Which agents contributed to a publication, and what explicit structural or scholarly roles, such as lead author, corresponding author, editor, reviewer, or data curator, did they perform?
UC-D3	Associations	What directional relationships connect a publication to external scholarly artifacts, such as preprints, datasets, software, supplementary materials, errata, or citations, and how are these relationships typed?
UC-D4	Structural Composition	What hierarchical and sequential document components, such as sections, subsections, paragraphs, sentences, and figures, compose a publication, and what is their absolute linear ordering?
UC-D5	Rhetorical Function	What structural or rhetorical block type, such as background, methodology, results, or limitations, characterizes each document component of a publication?
UC-D6	Source Grounding	Which exact document component and character- or byte-offset span grounds a given extracted entity, claim, or relation within a publication?
Family A: Paper Argument and Discourse		
UC-A1	Objectives	What specific research problems, questions, or objectives are explicitly declared as the target of investigation in a publication?
UC-A2	Claims	What distinct claims, hypotheses, findings, or core contributions are advanced by a publication, and what are their fine-grained propositional types?
UC-A3	Evidence Grounding	What specific empirical evidence, observations, results, figures, tables, or secondary citations ground or substantiate a given claim?
UC-A4	Discourse Relations	What rhetorical or argumentative relations, such as supporting, challenging, qualifying, explaining, motivating, or elaborating, exist between distinct propositions within a publication?
UC-A5	Constraints	What epistemic modal markers, assumptions, qualifications, or boundary limitations constrain the validity of a given claim?
UC-A6	Argument Paths	What directed path of discourse relations structurally connects the initial research problem of a publication through its methods and evidence to its final conclusions?
Family C: Concepts and Paper-local Use		
UC-C1	Mentions	What distinct textual expressions, surface forms, abbreviations, or coreferential expressions mention a concept within a publication?
UC-C2	Global Typing	To what canonical domain entity or global knowledge-base URI does a concept mention resolve, and what is its fixed ontological type, such as model, dataset, task, metric, theory, material, or algorithm?
UC-C3	Local Operational Role	What dynamic, paper-local role, such as proposed method, baseline approach, evaluation dataset, experimental variable, or output, does a concept fulfill within the specific methodology of a publication?
UC-C4	Propositional Argument Roles	In which specific semantic triples or n-ary propositions within a publication does a concept participate, and what structural position does it occupy?
UC-C5	Local Concept Alignment	How is a concept contextually paired with or related to other concepts, such as mapping a specific evaluation metric to a specific dataset, within the experimental scope of a publication?
UC-C6	Pragmatic Disposition	What is the authors' explicit disposition toward a concept, such as introducing, adopting, reusing, extending, modifying, comparing, evaluating, criticizing, or replacing it?
Family X: Cross-publication Concept Dynamics		

Continued on the next page

Table 3 continued from the previous page

ID	Title	Description
UC-X1	Cross-paper Co-occurrence	Which distinct publications reference the same canonical domain concept, irrespective of local naming variations?
UC-X2	Role Shift Matrix	How does the localized operational role of a single concept vary, transition, or contrast across a collection of different publications?
UC-X3	Definitional Variance	How do the formal definitions or textual characterizations of a specific concept differ across distinct publications?
UC-X4	Dialectical Claims	What epistemic alignments or conflicts, such as agreement, contradiction, refinement, or qualification, exist between claims made about the same concept across different papers?
UC-X5	Inter-paper Influence	How does the treatment of a concept in one publication directly influence, inherit from, reproduce, extend, or critique its treatment in an earlier publication?
UC-X6	Evolutionary Trajectory	What is the chronological trajectory of a concept's roles, usage frequencies, and domain shifts across a temporal or community-based graph projection?

Family P: Paper-local Pragmatic Descriptions

UC-P1	Intensional Definition	How does the text of a publication intensionally define or describe the essential nature of a concept?
UC-P2	Functional Intent	What explicit objective or goal-driven motivation does the text of a publication assign to the selection or deployment of a concept?
UC-P3	Operationalization Parameters	How is a concept configured, parameterized, operationalized, or structured into the systemic workflow described in a publication?
UC-P4	Attributed Affordances	What explicit capabilities, advantages, performance benefits, or unique affordances does a publication attribute to a concept?
UC-P5	Boundary Conditions	Under what explicit environmental constraints, domain dependencies, assumptions, or operational limits does a publication declare a concept to be functional?
UC-P6	Pragmatic Contrast	How does a publication contrast the utility, trade-offs, or suitability of a concept against alternative baselines or rival concepts?

A.2. Online Resources

SUDO ontology: <https://w3id.org/twc/sudo/>

SUDO documentation: <https://tetherless-world.github.io/sudo-ontology/>

ISWC publication corpus: <https://zenodo.org/records/21812666>

ISWC SUDO knowledge graph: <https://zenodo.org/records/21813306>

QASPER SUDO knowledge graph: <https://zenodo.org/records/21815806>

A.3. Concept-Structure Classes and Relations

SUDO defines five domain-agnostic subclasses of **Artifact** to capture the broad semantic categories through which scholarly nomenclature is organized. Their definitions and representative examples are summarized in Table 4.

Relations among Artifacts capture domain-independent functional, taxonomic, and compositional associations. SUDO also reuses SKOS properties to align canonical concepts with related instances and lexical variants, as summarized in Table 5.

A.4. Ontology Reuse and Vocabulary

SUDO reuses established vocabularies for publication metadata, document structure, knowledge organization, and argumentation. Table 6 lists these ontologies, while Table 7 summarizes the classes, datatypes, and properties used in each representational layer.

Table 4

Artifact classes used in the SUDO concept structure.

Class	Definition	Examples
Task	A problem to solve, application to support, or system to construct.	Information extraction, machine reading, image segmentation.
Method	A method, model, system, tool, framework, or system component.	Language model, CoreNLP, POS parser, kernel method.
Metric	A measure used to express the quality, efficiency, or behavior of a method or system.	F1, BLEU, precision, recall, robustness, time complexity.
Material	Data, datasets, corpora, resources, or knowledge bases used in research.	Image data, bilingual dictionary, CoNLL corpus, paraphrased questions.
Term	A significant scholarly notion that does not fall under the preceding categories.	Hallucination, faithfulness, Rhetorical Structure Theory, named entity.

Table 5

Relations used in the SUDO concept structure.

Relation	Interpretation	Example
used-for	Indicates that one Artifact is used to perform, support, or construct another.	TISPER is used for text-processing applications .
feature-of	Indicates that one Artifact is a feature, property, or characteristic of another.	Genre-specific regularities are features of discourse structure .
hyponym-of	Expresses a taxonomic relation in which one Artifact is a more specific type of another.	TUIT is a type of software library .
part-of	Expresses a compositional relation in which one Artifact forms a component of another.	A natural-language understanding module is part of a dialogue system .
closeMatch	Connects publication-specific Artifacts to canonical Concepts with sufficiently similar meanings, without asserting strict identity.	The Transformer Artifact in each publication closely matches the canonical Transformer Concept .
altLabel	Associates a canonical Concept with an alternative lexical form, abbreviation, or spelling variant.	The BERT Concept has the alternative label Bidirectional Encoder Representations from Transformers .

Table 6: Ontologies and vocabularies reused or referenced by SUDO.

Prefix	Ontology or vocabulary	Specification
foaf	Friend of a Friend	https://xmlns.com/foaf/spec/
bibo	The Bibliographic Ontology	https://dcmi.github.io/bibo/
xsd	XML Schema Datatypes	https://www.w3.org/TR/swbp-xsch-datatypes/
dcterms	Dublin Core Metadata Terms	https://www.dublincore.org/specification/s/dublin-core/dcmi-terms/
cito	The Citation Typing Ontology	https://sparontologies.github.io/cito/current/cito.html
fabio	The FRBR-aligned Bibliographic Ontology	https://sparontologies.github.io/fabio/current/fabio.html
doco	The Document Components Ontology	https://sparontologies.github.io/doco/current/doco.html
po	The Pattern Ontology	https://sparontologies.github.io/po/current/po.html

Continued on the next page

Table 6 continued

Prefix	Ontology or vocabulary	Specification
pav	Provenance, Authoring and Versioning Ontology	https://pav-ontology.github.io/pav/
prov	The Provenance Ontology	https://www.w3.org/TR/prov-o/
rdfs	RDF Schema	https://www.w3.org/TR/rdf-schema/
schema	Schema.org	https://schema.org/
skos	Simple Knowledge Organization System	https://www.w3.org/TR/skos-reference/
amo	The Argument Model Ontology	https://sparontologies.github.io/amo/current/amo.html
swan	Semantic Web Applications in Neuromedicine Ontology	https://www.w3.org/TR/hcls-swan/
semsur	Semantic Survey Ontology	https://saidfathalla.github.io/SemSur/doc/
sudo	Scholarly Upper Discourse Ontology	https://w3id.org/twc/sudo/ontology

Table 7: Classes, datatypes, and properties used in the SUDO ontology, grouped by representational layer.

Source	Classes and datatypes	Properties
Publication metadata		
foaf	Person	—
bibo	Proceedings	doi
xsd	integer, anyURI, string	—
dcterms	—	creator, isPartOf, subject, title, issued
cito	—	cites
rdfs	Literal	seeAlso
sudo	—	citationCount
Document structure		
fabio	ResearchPaper	—
doco	DiscourseElement, Section, SectionTitle, Paragraph, Sentence	—
po	—	isContainedByAsHeader, contains
schema	—	position
Information structure		
skos	—	closeMatch, altLabel
sudo	Concept, Artifact, Material, Method, Metric, Task, Term	usedFor, featureOf, hyponymOf, partOf
Propositional structure: argumentation		
amo	Argument, Backing, Claim, Evidence, Qualifier, Rebuttal, Warrant	backs, forces, isValidUnless, leadsTo, proves, supports
swan	Idea, Issue	responseTo
semsur	Approach	—
sudo	—	motivates, evidence, contextualizes, enables, justify, concede
Propositional structure: description		

Continued on the next page

Table 7 continued

Source	Classes and datatypes	Properties
sudo	Proposition, Descriptor, Elaboration, Circumstance, Solutionhood, Cause, Setting, Interpretation, Restatement, Evaluation, Purpose, Summary, Result, Condition	contrast, contextualizes, enables, summarizes, result, condition, restates, evaluates, purposes, causes, setting, interprets, elaborates, solves
Provenance		
prov	Agent	hadPrimarySource, wasDerivedFrom
pav	—	createdBy