Semantic Representation of Physics Research Data

Aysegul Say1a, Said Fathalla1,2b, Sahar Vahdati1,3c, Jens Lehmann1,4d and Sören Auer5,6e

1Smart Data Analytics (SDA), University of Bonn, Germany
2Faculty of Science, University of Alexandria, Egypt
3Department of Computer Science, University of Oxford, U.K.
4Fraunhofer IAIS, Dresden, Germany
5TIB Leibniz Information Center for Science and Technology, Hannover, Germany
6L3S Research Center, University of Hannover, Germany

Keywords: Semantic Web, Domain Ontology, Ontology Engineering, Semantic Publishing, Scholarly Communication, Physics.

Abstract: Improvements in web technologies and artificial intelligence enable novel, more data-driven research practices for scientists. However, scientific knowledge generated from data-intensive research practices is disseminated with unstructured formats, thus hindering the scholarly communication in various respects. The traditional document-based representation of scholarly information hampers the reusability of research contributions. To address this concern, we developed the Physics Ontology (PhySci) to represent physics-related scholarly data in a machine-interpretable format. PhySci facilitates knowledge exploration, comparison, and organization of such data by representing it as knowledge graphs. It establishes a unique conceptualization to increase the visibility and accessibility to the digital content of physics publications. We present the iterative design principles by outlining a methodology for its development and applying three different evaluation approaches: data-driven and criteria-based evaluation, as well as ontology testing.

1 INTRODUCTION

The advent of the Web has led researchers to a new era where research paradigms (empirical, theoretical, and computational) have merged with data-driven science (Hey et al., 2009). Today most of the scientific disciplines, especially physics, require data-driven technologies to integrate large-scale data that is produced by satellites, telescopes, and sensor networks. However, the application of data-intensive practices has produced a vast amount of unstructured data on the Web. Even though most of the scholarly output is meanwhile digitally available, the lack of coverage of digital content for each scientific community is a pervasive barrier to productive research in the physics domain. Since science is multidisciplinary in nature, researchers need resources that cover various subjects related to their research interest; however, with the current search engines, it is difficult to find links with cross-domain publications. As a result, it is inconvenient to discover, reuse, and process published articles with tools or interfaces for researchers.

The traditional system of scholarly communication is improving with recent developments related to the use of semantic and AI technologies. Semantic technologies and knowledge graphs offer new ways for discovering and dissemination of scholarly content, which leads to better collaboration. In this context, ontologies support knowledge extraction and modeling as a specification of a conceptualization. Also, they help to resolve the difficulty of handling the overflow of heterogeneous data by organizing and interlinking the data in a meaningful way.

The objective of this study is to support scholarly communication and fill the research gap by providing an ontology for organizing physics-related sci-
entific contributions. In this study, we target the following research question: How can we facilitate access to the physics research data in a machine-understandable way? Therefore, we applied semantic technologies to represent the outputs of physics research. With the Physics ontology (PhySci), we aim to support the transformation of scholarly communication in physics and provide more effective solutions for scholarly research. To enable a rich representation of scientific data, the FAIR principles (Wilkinson et al., 2016) are applied for rendering data and services. The principles emphasize the capacity of computational systems to find, access, interoperate, and reuse data efficiently. A knowledge graph scheme is designed for physics publications to determine how PhySci can exploit the formal structure of a publication and the details of research that is saved in the author’s mind (see Figure 1). This model has three main contributions. First, the densely interconnected content of scientific publications promotes accessing more convenient scientific collections and proposals. Second, improved reusability of materials enables researchers to advance the production of new knowledge. Third, performing semantic queries on organized knowledge can facilitate the interpretation of the physics content. To apply ontology-based representation, existing ontological resources are aligned with PhySci. In fact, the (PHYSCI) ontology is one of the ontologies of the Science Knowledge Graph Ontologies (SKGO) (Fathalla et al., 2020) suite. Various RDF serializations of the ontology can be found on SKGO’s GitHub repository¹. Furthermore, human-readable documentation of PhySci is available via its Persistent Identifiers (https://w3id.org/skgo/physci#). The prefixes are registered in prefix.cc², a name-space lookup service for RDF developers, under the open CC-BY 3.0 license.

The article is organized as follows: In section 2, we present the fundamental approaches that are applied in the development of the ontology. Section 3 presents specific design patterns and the structure of concepts in PhySci. The evaluation, given in section 4, discusses a set of assessment methods for PhySci. Section 5 introduces state-of-art practices. Section 6 gives a summary of our approach and an outlook of future work.

2 METHODOLOGY

In our approach, we applied ontological engineering practices to systematize the ontology development. Moreover, the application of the modeling methodologies enables to transform the requirements into a formal language that is designed, evaluated, and documented by the ontology. We follow the Ontology development 101 (Noy et al., 2001) and the Systematic Approach for Building Ontologies

¹https://github.com/saidfathalla/Science-knowledge-graph-ontologies
²https://prefix.cc/
(SABiO) (de Almeida Falbo, 2014) for the development of PhySci Ontology. Our modeling methodology is composed of four main phases with support activities and outlined as follows.

Identification of requirements: This phase starts with the selection of the focus area of the ontology, preparing and collecting requirements, and determining input sources and raw data. Requirements are addressed to cover the user issues and to reach a high-quality model. Those requirements are listed as follows.

- Accuracy: All axioms represented in the ontology should be aligned with the domain knowledge of stakeholders.
- Coherence: The ontology must be consistent with the terms related to the physics domain.
- Consistency: The ontology performance should be reliable and be able to process complex queries.
- Data Timeliness: The ontology should offer reliable and accessible data that are related to papers published within a specific period.
- Reusability: The ontology should interoperate with other ontologies.
- Reliability: The ontology should be extended by merging new definitions and information.

Domain Conceptualization and Formalization: Domain concepts that will be integrated into the ontology are determined. Conceptual modeling activity starts with selecting ontological and non-ontological knowledge resources.

Resource: Scientific publications are used as the primary non-ontological data source for capturing the domain concepts while creating the ontology. Those publications are selected from IOP and APS science journals with their whole context (e.g., abstract and introduction).

Topical Coverage: Physics, as a scientific discipline, is divided into different sub-disciplines (Feynman et al., 1965). We defined our corpus with mostly particle physics, and high energy physics since these topics involve popular investigations and have many relations to other sub-disciplines of physics. This activity produces a complete dictionary that includes classes, instances, and properties with dictionary tables. Using tables of classes and properties helps to gather all the useful and potentially usable domain concepts, their meanings, relations, labels, and URI's. Informal axioms extracted from resources are transformed into formal axioms. The formalization phase proposes to have clarity and correctness within the ontology.

Design and Development: Ontology is defined in a formal language. Web Ontology Language (OWL 2) is chosen to formalize the PhySci ontology. OWL provides different elements (e.g., classes, annotations, properties, and instances) that can be used for formalization and development tasks (McGuinness et al., 2004). We used Graffoo Editor to design the formal ontology and to visualize the structure of the ontology. Protégé ontology editor (Musen et al., 2015) offers strong functionality such as modification, querying, and reasoning. Thus, Protégé is used for the development of the PhySci ontology.

Ontology Testing: This phase executes all the requirements that are designed for the behavioral characteristics of the ontology. A set of test cases are formed from competency questions to ensure that the ontology satisfies the expected behavior regarding the competency questions. In addition to the development process listed above, SABiO (de Almeida Falbo, 2014) considers some supporting processes: documentation, reuse (section 3.1), knowledge acquisition, and evaluation.

Knowledge Acquisition: The process of knowledge acquisition starts by extracting terms and their synonyms from the underlying text. For this process, scientific publications about physics are set as a non-ontological resource to form a corpus. This corpus comprises 125,592 words and 5,083 sentences in total. Each section of the papers, rhetorical terms (e.g., conclusion, abstract), and scientific discourse elements (e.g., equations, theories) are identified in this task. We applied statistical techniques; TF-IDF (Term Frequency/Inverse Document Frequency) (Ramos et al., 2003) weighting scheme in combination with Latent Semantic Analysis (LSA) (Landauer et al., 1998) techniques for the knowledge acquisition process to capture latent concepts and discover a coherent knowledge base that devises an effective knowledge representation. Key terms that have the highest scores based on the TF-IDF scores are utilized to create a semantic space where terms are associated with one another. Therefore, most relevant terms are investigated using the LSA method.

3 PhySci ONTOLOGY

The purpose of developing the PhySci Ontology is to increase the value of the physics research data

---

3 https://iopscience.iop.org/
4 https://journals.aps.org/
5 http://www.w3.org/TR/owl2-overview/
6 http://www.essepuntato.it/graffoo
7 https://github.com/aysegulsay/PhySci/blob/master/Datasets
by creating an ontology in regards to FAIR principles (Wilkinson et al., 2016). The PhySci ontology is further refined by reusing other vocabularies and integrating extracted concepts that are determined from knowledge acquisition. Moreover, we demonstrate that PhySci ontology maintains the structure that is represented in Figure 1 by providing classes and relations to link those concepts. Figure 2 bestows the central part of the ontology; it describes the related entities with their object properties.

The development of an ontology requires an analysis of the types of concepts and relations. We applied statistical analysis techniques to the textual data to achieve an effective knowledge representation. Statistical techniques, which are defined in section 2, have been used to extract candidate terms and identify specific sentence patterns from the corpus to conform the triples in the ontology.

3.1 Reuse of Existing Resources

The PhySci ontology imports seven ontologies and provides 110 OWL classes and 77 object properties. We aligned similar and new concepts for the physics domain from existing ontologies to achieve interoperability with other systems and ontologies. This technique is mostly used for constructing domain-specific ontologies in ontology engineering practices (Corcho et al., 2007). Indeed, it helps to provide better coverage of the domain. While adapting new ontologies to PhySci, we have followed the ontological levels (upper, middle, lower) to determine the semantic interoperability of PhySci. We begin defining rhetorical terms that establish the general skeleton of the publication. npg:Thing, npg:Publication, npg:Issue, npg:Person, npg:Journal, and npg:Agent are selected from Nature Publishing Group ontology (NPG) (Hammond and Pasin, 2015). From Dublin Core (Weibel et al., 1998), we selected entities to describe annotations of classes and relations between instances such as terms:creator, terms:publisher and data properties such as terms:Abstract and terms:Date Modified. Semantic Web for Earth and Environmental Terminology (SWEET) (Raskin and Pan, 2003) is developed for the Earth system science domain. Terms relevant to physical models and components of evidence are aligned from SWEET ontology, such as soli:Simulation, mod:ScientificModel, and phen:Phenomena. Extensible Observation Ontology(OBOE) (Madin et al., 2007) arranges semantic subtleties of complex ecological data. To define physical quantities that are used in equations, we reused entities such as oboe-core:Unit, oboe-core:Measurement, and oboe-core:Force from OBOE. Semantic Sensor Network (SSN) (Compton et al., 2012) ontology represents sensors and their observations with related procedures, samples, and actuators. SOSA (Sensor, Observation, Sample, and Actuator) is another module of SSN. We mostly reused concepts such as sosa:Observation and about observations from the SOSA module to define scientific instruments in research. The Ontology of Astronomical Object Types (IVOA) (Cambrésy et al., 2010) introduces astronomy and formation terms. To define characteristics of a formation, ivoao:horizon, physics:collision, and physics:spectrum classes are aligned from this ontology.

3.2 Representing Metadata in PhySci

In this section, we describe classes, instances, and properties that are implicitly developed in PhySci...
Figure 3: Scientific Model. Scientists try to draw their scientific knowledge by using scientific models to understand and define features of specific patterns that occur in the universe. Thus, related classes and object properties to the scientific model have been defined in PhySci ontology with its instances for the publication (Medeiros et al., 2018).

Figure 4: Triple Patterns of instances and relations defined in PhySci for scientific paper (Medeiros et al., 2018).

ontology. Each concept has the following features: a URI, a preferred synonymous label, and a definition. Figure 3 depicts the Scientific Model class and its related classes, instances, and relations described in PhySci. All instances have been directly extracted from research papers. Figure 4 shows relations between instances for a scientific publication (Medeiros et al., 2018). Triple patterns are created by constructing domain and range constraints for each object property such as physci:ResearchWork and physci:considerCase physci:Case. To specify different use cases, relations are asserted for each characteristic of properties (e.g., transitive, asymmetric). For example, physci:isformedfrom and physci:yieldEquation are set as reflexive relations. Other properties are defined as functional relations and inverse functional such as physci:relyOnMeasurement is functional while physci:hasSource, and physci:hasObservation are defined as inverse functional.

Object properties are created concerning clusters of similar instances to other instances. For example, terms:Publication class is defined for the physics articles and it is set as a domain of the object property physci:addressesResearchWork. The range of this property is Research Work class. Another example is the instance physci:CombinedFieldElectricity of physci:Solution class connected to instance physci:GeneralTheoryofrelativity of class ivaoa:theory via the object property physci:usesTheory. Instances of observation class can be related to the observatory, observational data, observer, duration, and measurement classes. For example, the observation class can be related to observatory (Observation ⊑ hasObservatory.Observatory) and formation (Observer ⊑ detectFormation.Formation). Observation class has relations with other classes (Observation ⊑ hasObservationalData.ObservationalData). The instance physci:EHTVLBCampaign Observation belonging to the sosa:Observation class is connected to the instance physci:EHTData.
of the class physci:ObservationalData by the object property physci:hasObservationalData. We specified data properties for the Publication class to define the publication’s title and abstract. Furthermore, the data property npg:publicationYear can be used to associate a published year with a publication. Physical quantities of the equations and scientific properties can be defined by using the data properties such as physci:has Mass, physci:has Condition, physci:hasParameter, physci:has Velocity, physci:hasParameter, physci:has State, physci:has Energy, and physci:has Temperature in PhySci to distinguish the equations.

Example candidate terms captured from scientific publication (Abbott et al., 2016) and their defined classes and instances in PhySci can be seen in Figure 5. Additionally, extracted instances with related class names are listed in the tables Table 1 and Table 2.

### 3.3 Reasoning

The Semantic Web Rule Language (SWRL) (Horrocks et al., 2004) establishes expressive representation formalisms and helps to reveal new inference in an ontology. Therefore, SWRL rules are defined for PhySci by executing Drools reasoner (Proctor, 2011) to infer alternative linking triples, to discover inconsistencies, and to improve the expressivity. The rule-set of PhySci comprises the following SWRL rules.

From Equation 1, we can infer that if two publications relate to each other, then one of them might cite the other. In Equation 2, the research work demonstrates a scientific model to explain a solution by applying specific methods, and thus this solution might include the scientific method. Equation 3 suggests that research work reveals a formation, and it explains a case related to the formation. For instance, research work that investigates black holes might explain some cases, such as the collision of neutrinos and massive star collapsing to describe the occurrence of formations.

\[
\text{Publication}(?x) \land \text{Publication}(?y) \land \\
\text{relation}(?x, ?y) \land \text{addressesResearchWork}(?y, ?z) \land \\
\text{addressesResearchWork}(?x, ?z) \rightarrow 
\]

\[
\text{ScientificModel}(?x) \land \text{provideSolution}(?x, ?y) \land \\
\text{hasScientificMethod}(?y, ?z) \rightarrow 
\text{useScientificMethod}(?x, ?z) 
\]

\[
\text{ResearchWork}(?x) \land \text{Formation}(?y) \land \\
\text{argueFormation}(?x, ?y) \land \text{formedBy}(?y, ?z) \land \\
\text{ConsiderCase}(?x, ?z) 
\]

Figure 5: Terms that are captured from the publication (Abbott et al., 2016) and their matched classes in PhySci.
Table 1: Classes and related data properties in PhySci with captured terms from the research paper (Abbott et al., 2016).

<table>
<thead>
<tr>
<th>Class</th>
<th>Data Property</th>
<th>Literal Value</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication</td>
<td>title</td>
<td>GW150914:Implications for the Stochastic Gravitational-Wave Background from Binary Black Holes</td>
<td>rdfs:Literal</td>
</tr>
<tr>
<td>Publication</td>
<td>publicationDate</td>
<td>2016-03-31</td>
<td>xsd:date</td>
</tr>
<tr>
<td>Publication</td>
<td>doi</td>
<td>DOI:10.1103</td>
<td>rdfs:Literal</td>
</tr>
<tr>
<td>Publication</td>
<td>creator</td>
<td>B.P. Abbott et. al</td>
<td>rdfs:Literal</td>
</tr>
<tr>
<td>Journal</td>
<td>issue</td>
<td>PRL 116, 1311102(2016)</td>
<td>rdfs:Literal</td>
</tr>
<tr>
<td>Observation</td>
<td>ObservationTime</td>
<td>2015-10-14</td>
<td>xsd:dateTimeTime</td>
</tr>
<tr>
<td>Formation</td>
<td>hasScenario</td>
<td>unresolvable events combine to create stochastic background</td>
<td>rdfs:Literal</td>
</tr>
</tbody>
</table>

Table 2: Classes and related instances in PhySci for the research paper (Abbott et al., 2016).

<table>
<thead>
<tr>
<th>Class</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publisher</td>
<td>American Physical Society</td>
</tr>
<tr>
<td>Creator</td>
<td>B.P. Abbott et al.</td>
</tr>
<tr>
<td>ResearchWork</td>
<td>Implications for Gravitational Wave</td>
</tr>
<tr>
<td>Phenomena</td>
<td>Gravitational waves</td>
</tr>
<tr>
<td>Observation</td>
<td>LIGO detection of gravitational waves</td>
</tr>
<tr>
<td>Observatory</td>
<td>The Laser Interferometer Gravitational Wave Observatory (LIGO)</td>
</tr>
<tr>
<td>Observer</td>
<td>Hanford and Livingston detector</td>
</tr>
<tr>
<td>Formation</td>
<td>Binary Black Holes</td>
</tr>
<tr>
<td>Wave</td>
<td>GW150914</td>
</tr>
<tr>
<td>Spectrum</td>
<td>Energy density spectrum</td>
</tr>
</tbody>
</table>

4 EVALUATION

The evaluation is a fundamental task for ontology engineers to verify and validate the quality of the ontology. According to SABIO methodology (de Almeida Falbo, 2014), the evaluation process has two main activities: (1) to check the ontology requirements are being met with specific criteria and (2) to ensure that the verified requirements are compatible with the intended use of the ontology. Therefore, we assessed the content by applying data-driven (Rospocher et al., 2012) and criteria-based approaches (Gangemi et al., 2005). The ontology testing technique is practiced by executing a test case for each competency question to verify the requirements. These approaches have been applied to assess in what dimensions PhySci can bring value to the scholarly community and how high can the impact be on the semantic publishing.

(1) Ontology Content Evaluation: Quality criteria are defined to assess if the content of ontology contains any anomalies or redundant information (Gangemi et al., 2005) (Lovrencic and Cubrilo, 2008). The main goal of this step is to resolve if the ontology defines concepts accurately, does not define, or even defines inaccurately (Gómez-Pérez, 2001). A set of criteria (i.e., consistency, completeness, conciseness, expandability, and sensitiveness) are applied through expert evaluation to assess the quality of the ontology, the rate of its performance, and the definition of the concepts. Each quality metric should conform to associated questions to check if the ontology satisfies conditions or not. Table 3 shows the correspondence between the metric, questions, and results.

(2) Ontology Testing: Competency questions are prepared to determine the behavioral characteristics of the ontology that relate to the knowledge represented in PhySci ontology. They are used to ensure that the ontology implementation compatible with the scope of PhySci. We created these questions from the content of the text corpus. To evaluate the completeness, instantiation queries that represent the competency questions are prepared. Therefore, CQs are transformed into SPARQL queries that can be executable within the framework. Table 4 presents a sample of 10 competency questions out of 20. CQ1 is expressed with SPARQL query in Listing 1 to find publications that contain theories for a specific problem.
Table 3: Quality Criteria for ontology evaluation.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Questions</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency</td>
<td>Does the documentation of ontology meet the specification? Is there any encoding bias related to the transformation from the knowledge level to the encoding? Is the representation be made genuinely for the benefit of implementation?</td>
<td>Yes, ontology is consistent since it does not include any contradictory conclusions. Reasoner shows no error.</td>
</tr>
<tr>
<td>Completeness</td>
<td>Is the domain of interest properly covered? Can the ontology answer all the competency questions? Does the ontology include all related concepts to the domain and their lexical representations?</td>
<td>Yes, the ontology is complete regarding the requirements specifications that are designed in the identification of requirements, and all competency questions are answered.</td>
</tr>
<tr>
<td>Conciseness</td>
<td>Are there any irrelevant axioms concerning the domain to be covered? Does it support a minimal ontological commitment? Are there any weak assumptions regarding the ontology’s underlying domain?</td>
<td>Yes, the ontology is concise since it does not store any unnecessary or useless definitions.</td>
</tr>
<tr>
<td>Expandability</td>
<td>Does the ontology flexible enough to support new definitions and axioms? Is the ontology be expanded by adding new knowledge to classes without altering the already defined concepts?</td>
<td>Yes, the ontology is expandable since adding or modifying the concepts does not influence other axioms and classes.</td>
</tr>
<tr>
<td>Sensitiveness</td>
<td>How is the ontology affected regarding altering the semantics of the ontology?</td>
<td>The ontology is not sensitive since it is expandable, meaning that changes in definitions of different concepts did not affect the other defined concepts.</td>
</tr>
</tbody>
</table>

from PhySci ontology. CQ1. Which publications use relativity theory to solve the particle problem?

The output of query CQ1 gives the publication title that is discussed for the solution of the particle problem. The results are listed as follows; publication: “The Particle Problem in the General Theory of Relativity”, problem: “Particle Problem”, and solution: “A Special Kind of Singularity and Removal”.

(3) Data-driven Approach: The corpus-based terminological ontology approach assesses the coverage and the capacity of the ontology (Rospocher et al., 2012). This approach has many advantages to determine the uncertainty of domain-specific terminology; therefore, it can provide a precise output to rank the relevancy of a knowledge domain.

Listing 1: SPARQL example for query CQ1 (in Table 4).

```sparql
```

It starts with the extraction of concepts from the defined corpus by applying TF-IDF, then each extracted concept is compared with the ontology to find similar class terms. Next, the number of overlapped concepts between the ontology and corpora are listed. After that, metrics (precision, recall, F1) can be applied by using the number of classes in the ontology and the number of matched concepts. Two different corpora are generated to examine how far PhySci covers different topics of physics. The main objective of this approach is to select corpora that contain different topics than the corpus that is used in the development of the PhySci. Additionally, we compared different types of ontologies against the text corpus to see how well PhySci is suitable for the domain to be represented with respect to other ontologies. All ontologies are assessed to check that they adequately define
the terminology and represent the most relevant concepts appropriately. We chose ontologies that are the most current ontologies in their field and closest to the physics domain.

**OM Ontology (Ontology of Units of Measure and Related Concepts)** (Rijgersberg et al., 2013) is an ontology about the science domain and developed to improve the alignment and representation of quantitative research data.

**OPB Ontology (Ontology of Physics for Biology)** (Cook et al., 2008) is a reference ontology of physical principles (classical physics and thermodynamics) that can be applied to the bioinformatics modeling. It is developed to bridge the gap between the biosimulation, biological processes, and physical domains (e.g., fluid dynamics and particle diffusion) to annotate biosimulation models.

**ENVO Ontology (Environment Ontology)** (Buttigieg et al., 2013) is an ontology for defining a broad range of environments related to ecosystems, environmental processes, and habitats.

We established a corpus-based evaluation based on these ontologies and Corpus1 and Corpus2 to evaluate the coverage of each ontology against each other. Corpus1 has produced from the scientific publications that involve topics of atomic, molecular, and optical physics (Physical review A) published in APS. A total of 25 articles are added to the dataset. The search results of Google Scholar generated Corpus2 by performing the keyword “black holes in string theory”. The top 50 keywords are captured by applying TF-IDF to the datasets.

\[
\text{Precision} = \frac{|N_{\text{hits}}|}{N_{\text{class}}} \tag{4}
\]

\[
\text{Recall} = \frac{|N_{\text{hits}}|}{|\text{List}|} \tag{5}
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}
\]

Then, precision, recall, and f1 values are calculated, which are depicted in Equation 4, Equation 5, and Equation 6, respectively. We perform analysis using the Corpus1, Corpus2, OM, OPB, and ENVO ontology. Table 5 presents the output of assessments. It includes the results of precision, recall, F1, the total number of classes, and hits as the number of matched concepts among the ontology and the corpus. Many classes in PhySci have matched with top-ranked concepts extracted from each corpus (17 hits and 15 hits among the top 50 key-concepts) while the compared ontologies have a lower number of hits among the top 50 key-concepts even though they contain more classes than PhySci. Although the precision and recall values of PhySci are still relatively low, it scores significantly higher than the benchmark ontologies. However, to increase the precision and recall values, PhySci would be extended by aligning with different ontologies or adding new concepts to capture more terms from the scientific literature. Also, the results show that the F1 value of PhySci is greater than 0.15, which means that the PhySci covers more knowledge than other ontologies. This method helps to confirm that PhySci is sufficiently aligned with the defined domain of interest.

## 5 RELATED WORK

The availability of encyclopedic and factual knowledge representation in machine-actionable form is increased and resulted in different knowledge graphs, such as DBpedia (Lehmann et al., 2015). However, there is a scarcity of developing science-based ontologies, especially for the physics domain. Thus, most
approaches do not cover scholarly data with physics knowledge; but only interpret articles or scholarly outputs in the light of more general rhetorical elements. For the scholarly domain, semantic publishing is applied as an approach to undertake the challenges of scholarly communication by utilizing the metadata concepts such as Semantic Publishing Referencing Ontologies (SPAR) (Peroni and Shotton, 2018). SPAR covers different ontology modules (e.g., DoCo, FaBiO, and DEO) to support distinctive features of the scholarly publishing domain together with semantic technologies, for example, document description, bibliometric data, and workflow processes. Springer Nature Publishing’s SN SciGraph11 focuses primarily on bibliographic data in the scholarly domain. It provides a rich semantic fabric of bibliographic metadata for the visualization of the scholarly domain. Many attempts have been made (Fathalla et al., 2017; Jaradeh et al., 2019a; Vogt et al., 2020; Say et al., 2020) with the purpose of representing research contributions as knowledge graphs aiming at improving scientific data management and retrieval. The Semantic Survey Ontology (Semsur) (Fathalla et al., 2017; Fathalla et al., 2018) is one of the preliminary attempts to design an ontology for systematizing and linking research findings presented in surveys in computer science. The Open Research Knowledge Graph (ORKG) (Jaradeh et al., 2019b; Jaradeh et al., 2019c) is a semantic publishing platform presenting a knowledge graph to retrieve and explore scientific knowledge that is described in scholarly literature. It aims to represent research in a structured manner for easier access by changing the document-oriented workflows in scholarly communication.

For the science domain, PhySH (Physics Subject Headings)12 (Smith, 2019) is a physics taxonomy that is presented by the American Physical Society to manage subject indexes in physics. The ultimate objective of PhySH is to provide a fully open and high-quality classification scheme in the field of physics. This data model is developed to connect subjects with papers submitted and published in Physical Review journals. There are numerous ontologies presented in the life science domain, such as MeSH (Medical Subject Headings) (Lipscomb, 2000) is a thesaurus developed for indexing articles from the Medline database. OntoBio (Albuquerque et al., 2016) is a biodiversity domain ontology, and it is designed by adopting SABIO methodology (de Almeida Falbo, 2014). It is a formal ontology for biological collection and field data collection of biotic entities. Eventually, concerning its coverage, PhySci incorporates features related to all physics and scholarly domains of research works that are published and found as heterogeneous data. Thus, PhySci, in comparison with all those works, fulfills the uncovered requirements of a physics and scholarly communication domain by representing document-based information in the form of metadata.

6 CONCLUSION

In this study, we examined the possibility of applying linked data principles to physics research data by developing PhySci ontology. We provide an ontology that allows researchers to reuse, access, and find scientific knowledge that will assist their research. The dynamic content of PhySci enables the exploration of non-obvious information found in publications. In this work, we leveraged semantic technologies and knowledge graphs to transform single query patterns of physics research data into a sophisticated ongoing conversation between computers and researchers. Thus, PhySci can support the organization of the content of Physics publications by describing scientific information semantically. PhySci helps to deal with the information overload and facilitates the transformation of the document-oriented workflows in scholarly communication by enabling extensibility, flexibility, and interoperability of scientific data. It also allows the indexing of articles that are used for search, curation, and augmentation. The PhySci ontology satisfies the quality requirements according to the results of the ontology testing, data-driven, and content evaluation.

In future work, we will target the adaptation of
PhySci within the Open Research Knowledge Graph (ORKG)\textsuperscript{13} to facilitate the discoverability of physics-related publications. Besides, the precision and recall of the PhySci ontology will be improved by covering more topics and sub-topics related to physics research such as electricity and magnetism, or mechanics in the future. Furthermore, we will extend this work for other scientific disciplines and envision a science knowledge graph covering various scientific fields (e.g., life science, earth science) for scholarly publishing.

ACKNOWLEDGEMENTS

This work has been supported by ERC project ScienceGRAPH (grant no. 819536).

REFERENCES


de Almeida Falbo, R. (2014). Sabio: Systematic approach for building ontologies. In ONTO.COM/ODISE@ FOIS.


