Semantic Representation of Physics Research Data

Aysegul Say[1][a], Said Fathalla[1,2][b], Sahar Vahdati[1,3][c],
Jens Lehmann[1,4][d] and Sören Auer[5,6][e]

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

[a] https://orcid.org/0000-0002-8803-7355
[b] https://orcid.org/0000-0002-2818-5890
[c] https://orcid.org/0000-0002-7171-169X
[d] https://orcid.org/0000-0001-9108-4278
[e] https://orcid.org/0000-0002-0698-2864

Say, A., Fathalla, S., Vahdati, S., Lehmann, J. and Auer, S.
Semantic Representation of Physics Research Data.
DOI: 10.5220/0010111000640075
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 2: KEOD*, pages 64-75
ISBN: 978-989-758-474-9

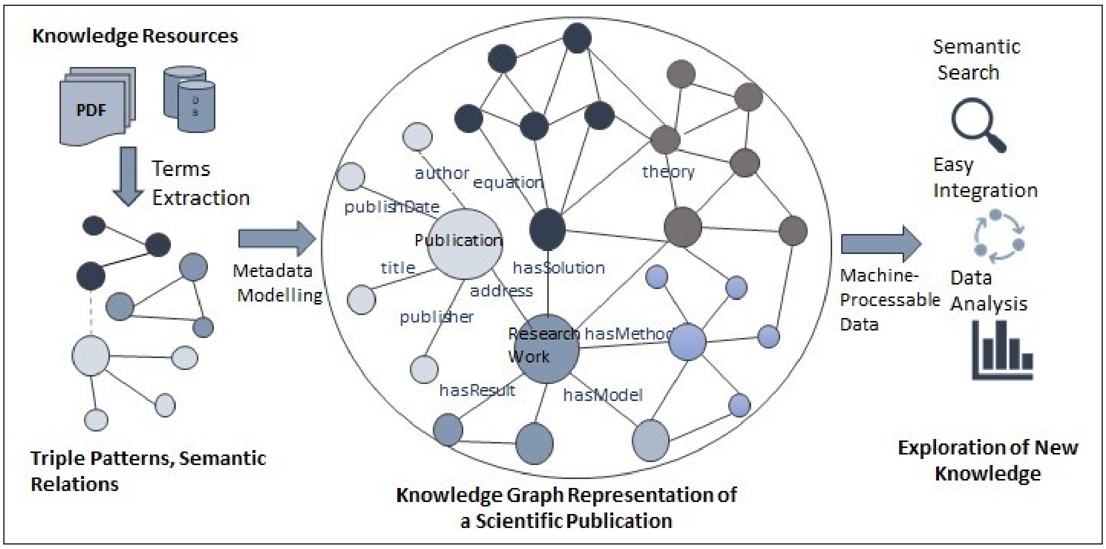


Figure 1: Knowledge capturing for physics publications. (1) Candidate terms are detected from unstructured text (i.e., physics publications). (2) Networks of scholarly entities and science discourse elements are created. (3) All concepts are structured and combined in a knowledge graph. (4) Discover and analyze complex or hidden relations. Enable collaboration between different scientific communities.

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 repository[1]. Furthermore, human-readable documentation of PhySci is available via its Persistent Identifiers (https://w3id.org/skgo/physci#). The prefixes are registered in *prefix.cc*[2], 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

[1] https://github.com/saidfathalla/Science-knowledge-graph-ontologies

[2] https://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 unified with the terms related to the physics domain. *Consistency:* The ontology must be consistent without having any contradiction about input data. *Extensibility:* The ontology should be extended by merging new definitions and information. *Reliability:* 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.

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[3] and APS[4] 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)[5] 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[6] 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[7]. 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]https://www.w3.org/TR/owl2-overview/
[6]http://www.essepuntato.it/graffoo
[7]https://github.com/aysegulsay/PhySci/blob/master/Datasets

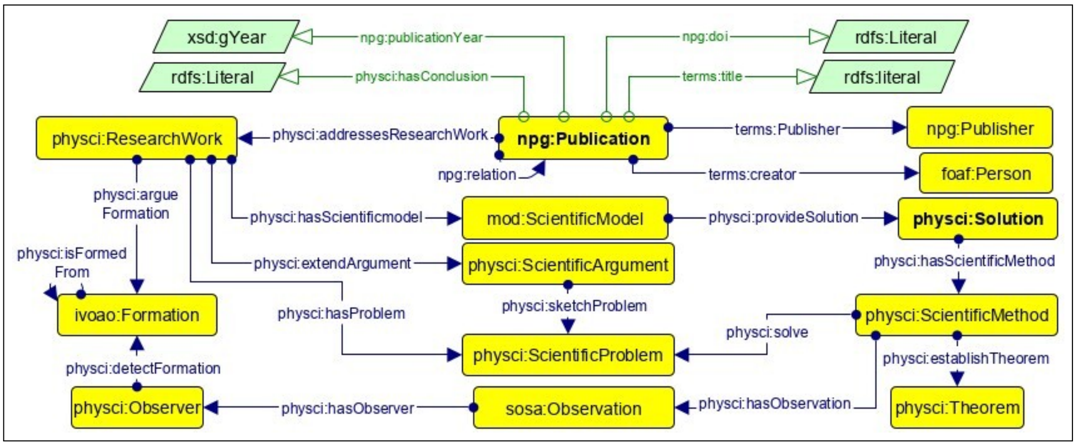


Figure 2: Linked Scholarly Entities. The interpretation of main classes, object properties and data properties in PhySci ontology.

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 `sol: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

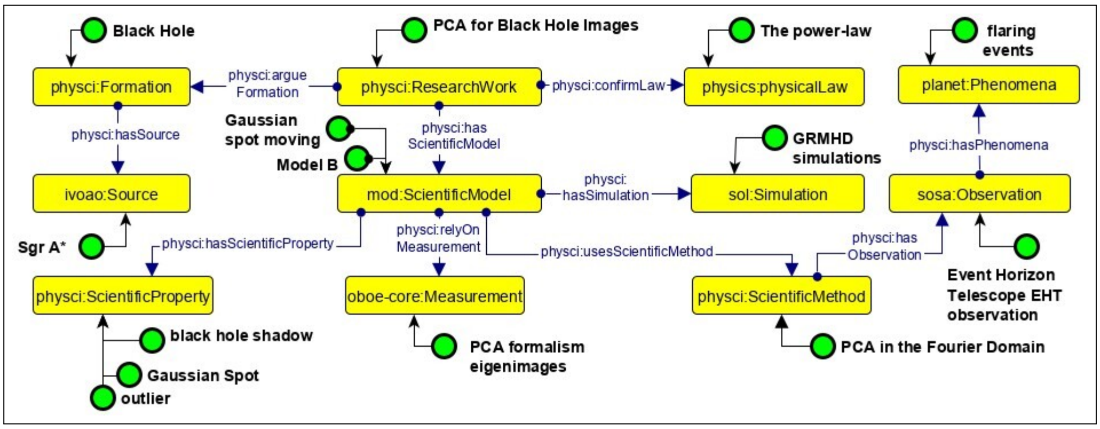


Figure 3: Scientific Model. Scientists try to draw their scientific knowledge by using scientific models to understand and define features of specific patterns that occur in the universe. Thus, related classes and object properties to the scientific model have been defined in PhySci ontology with its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

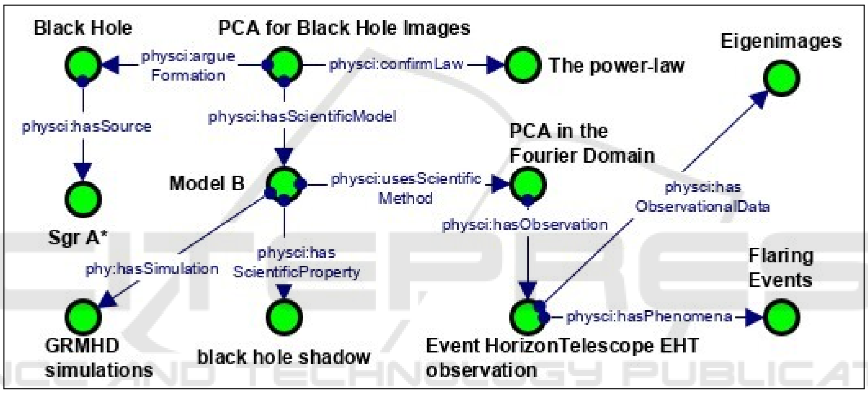
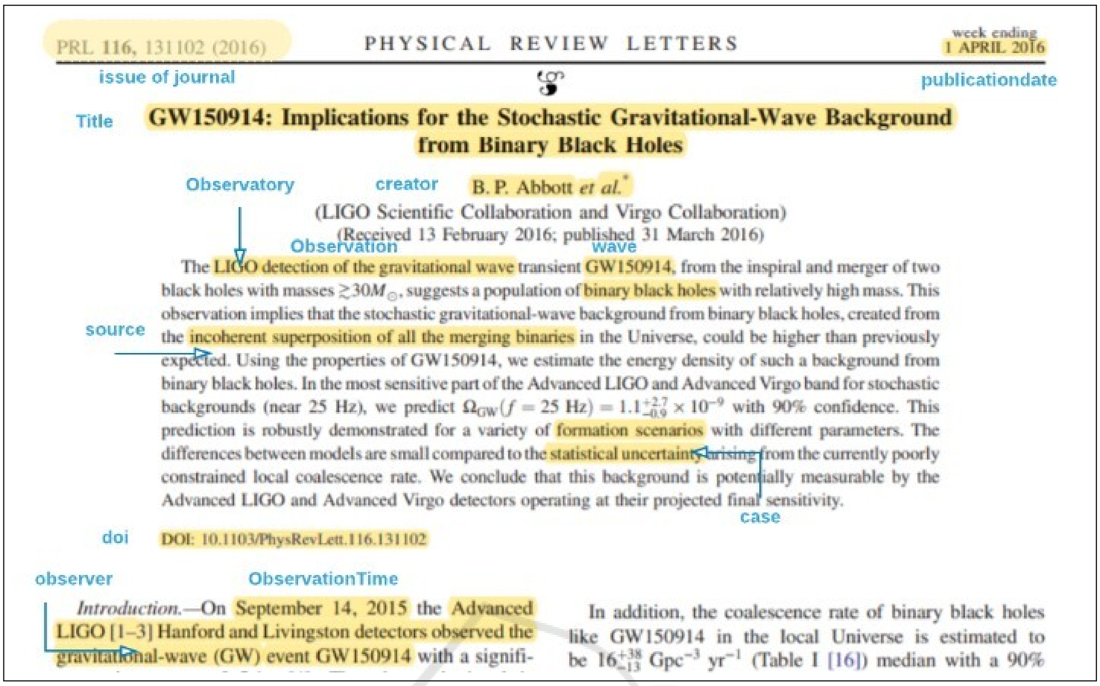


Figure 4: Triple Patterns of instances and relations defined in PhySci for scientific paper (Medeiros et al., 2018).

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 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:CombinedField Electricity` 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 ($Observation \sqsubseteq \exists hasObservatory.Observatory$) and formation ($Observer \sqcap \exists detectFormation.Formation$). `Observation` class has relations with other classes ($Observation \sqsubseteq \exists hasObservationalData.ObservationalData$). The instance `physci:EHTVLBICampaign Observation` belonging to the `sosa:Observation` class is connected to the instance `physci:EHTData`


PRL 116, 131102 (2016) PHYSICAL REVIEW LETTERS week ending 1 APRIL 2016

GW150914: Implications for the Stochastic Gravitational-Wave Background from Binary Black Holes

B. P. Abbott *et al.**

(LIGO Scientific Collaboration and Virgo Collaboration)

(Received 13 February 2016; published 31 March 2016)

The LIGO detection of the gravitational wave transient GW150914, from the inspiral and merger of two black holes with masses $\gtrsim 30 M_\odot$, suggests a population of binary black holes with relatively high mass. This observation implies that the stochastic gravitational-wave background from binary black holes, created from the incoherent superposition of all the merging binaries in the Universe, could be higher than previously expected. Using the properties of GW150914, we estimate the energy density of such a background from binary black holes. In the most sensitive part of the Advanced LIGO and Advanced Virgo band for stochastic backgrounds (near 25 Hz), we predict $\Omega_{GW}(f = 25\ \text{Hz}) = 1.1^{+2.7}_{-0.9} \times 10^{-9}$ with 90% confidence. This prediction is robustly demonstrated for a variety of formation scenarios with different parameters. The differences between models are small compared to the statistical uncertainty arising from the currently poorly constrained local coalescence rate. We conclude that this background is potentially measurable by the Advanced LIGO and Advanced Virgo detectors operating at their projected final sensitivity.

DOI: 10.1103/PhysRevLett.116.131102

Introduction.—On September 14, 2015 the Advanced LIGO [1–3] Hanford and Livingston detectors observed the gravitational-wave (GW) event GW150914 with a signifi-

In addition, the coalescence rate of binary black holes like GW150914 in the local Universe is estimated to be 16^{+38}_{-13} Gpc^{-3} yr^{-1} (Table I [16]) median with a 90%

Figure 5: Terms that are captured from the publication (Abbott et al., 2016) and their matched classes in PhySci.

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:hasState`, `physci:hasEnergy`, 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 ruleset 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.

$$Publication(?x) \wedge Publication(?y) \wedge relation(?x,?y) \wedge addressesResearchWork(?y,?z) \rightarrow addressesResearchWork(?x,?z) \quad (1)$$

$$ScientificModel(?x) \wedge provideSolution(?x,?y) \wedge hasScientificMethod(?y,?z) \rightarrow useScientificMethod(?x,?z) \quad (2)$$

$$ResearchWork(?x) \wedge Formation(?y) \wedge argueFormation(?x,?y) \wedge formedBy(?y,?z) \rightarrow ConsiderCase(?x,?z) \quad (3)$$

Table 1: Classes and related data properties in PhySci with captured terms from the research paper (Abbott et al., 2016).

Class	Data Property	Literal Value	Data Type
Publication	title	GW150914:Implications for the Stochastic Gravitational-Wave Background from Binary Black Holes	`rdfs:Literal`
Publication	publicationDate	2016-03-31	`xsd:date`
Publication	doi	DOI:10.1103	`rdfs:Literal`
Publication	creator	B.P. Abbott et. al	`rdfs:Literal`
Journal	issue	PRL 116, 131102(2016)	`rdfs:Literal`
Observation	ObservationTime	2015-10-14	`xsd:dateTime`
Formation	hasScenario	unresovable events combine to create stochastic background	`rdfs:Literal`

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

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

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.

Criteria	Questions	Results
Consistency	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?	Yes, ontology is consistent since it does not include any contradictory conclusions. Reasoner shows no error.
Completeness	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?	Yes, the ontology is complete regarding the requirements specifications that are designed in the identification of requirements, and all competency questions are answered.
Conciseness	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?	Yes, the ontology is concise since it does not store any unnecessary or useless definitions.
Expandability	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?	Yes, the ontology is expandable since adding or modifying the concepts does not influence other axioms and classes.
Sensitiveness	How is the ontology affected regarding altering the semantics of the ontology?	The ontology is not sensitive since it is expandable, meaning that changes in definitions of different concepts did not affect the other defined concepts.

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).

```
SELECT  DISTINCT ?pbl ?prblm ?sol
{
?pbl physci:addressesResearchWork ?wrk.
?wrk physci:hasProblem ?prblm.
?sol physci:solve ?prblm.
?sol physci:usesTheory  ?theory.
?theory rdfs:label  ?label.
?prblm  rdfs:label  ?plabel.
FILTER (regex(?label,"relativity")
&&regex(?plabel ,"Particle"))
}
```

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

Table 4: Competency questions for PhySci ontology.

Query	Competency Question
CQ1	Which publications use theory X to solve problem Y?
CQ2	Which concepts is used in publications for the theory Y?
CQ3	What are the results of Research Work X?
CQ4	List all the equations that are used in publication X?
CQ5	Which publications use scientific model X?
CQ6	Who are the authors of publications that use research work X for problem Y?
CQ7	Which equations, theories are used in Scientific Method X?
CQ8	Which observational data is used in observation X with observatory Y?
CQ9	Which phenomena are observed in Observation X with scientific method Y?
CQ10	Give me the velocity, energy of the signal that is found in the horizon X of the Formation Y formed from Formation Z?

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)[8] (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 [9] and Corpus2 [10] to evaluate the coverage of each ontology against each other. Corpus1 has produced from the scientific publications that involve the 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.

$$Precision = \frac{|N_{hits}|}{N_{class}} \tag{4}$$

$$Recall = \frac{|N_{hits}|}{|List|} \tag{5}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

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

[8] http://www.environmentontology.org

[9] https://github.com/aysegulsay/PhySci/blob/master/Datasets/DatasetCorpus1.tsv

[10] https://github.com/aysegulsay/PhySci/blob/master/Datasets/DatasetCorpus2.tsv

Table 5: Data-driven approach, corpus-based evaluation results.

	Corpus1				Corpus2			
	PhySci	OM	OPB	ENVO	PhySci	OM	OPB	ENVO
Classes	110	808	846	6,240	110	808	846	6,240
Hits	17	13	15	15	15	14	16	14
Precision	0.15	0.01	0.01	0.002	0.13	0.01	0.01	0.002
Recall	0.34	0.26	0.30	0.30	0.30	0.28	0.32	0.28
F1	0.20	0.03	0.03	0.004	0.18	0.03	0.03	0.004

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 SciGraph[11] 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.

[11]https://www.springernature.com/gp/researchers/scigraph

[12]https://physh.aps.org

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)[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

Abbott, B. P., Abbott, R., Abbott, T., Abernathy, M., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R., et al. (2016). Gw150914: Implications for the stochastic gravitational-wave background from binary black holes. *Physical review letters*, 116(13):131102.

Albuquerque, A. C. F., Dos Santos, J. L. C., and de Castro Júnior, A. N. (2016). Ontobio: Designing new features to improve modeling and implementation. In *ONTOBRAS*, pages 169–174.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., Consortium, E., et al. (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics*, 4(1):43.

Cambrésy, L., Derriere, S., Padovani, P., Preite Martinez, A., and Richard, A. (2010). Ontology of astronomical object types. *International Virtual Observatory Alliance. Available at: http://www. ivoa. net/Documents/Notes/AstrObjectOntology/20100117/ NOTE-AstrObjectOntology-1.3-20100117. html. Accessed on*, 11(06):2011.

Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al. (2012). The ssn ontology of the w3c semantic sensor network incubator group. *Web semantics: science, services and agents on the World Wide Web*, 17:25–32.

Cook, D. L., Mejino Jr, J. L., Neal, M. L., and Gennari, J. H. (2008). Bridging biological ontologies and biosimulation: the ontology of physics for biology. In *AMIA Annual Symposium Proceedings*, volume 2008, page 136. American Medical Informatics Association.

Corcho, O., Fernandez-Lopez, M., and Gomez-Perez, A. (2007). Ontological engineering: what are ontologies and how can we build them? In *Semantic web services: Theory, tools and applications*, pages 44–70. IGI Global.

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

Fathalla, S., Auer, S., and Lange, C. (2020). Towards the semantic formalization of science. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2057–2059.

Fathalla, S., Vahdati, S., Auer, S., and Lange, C. (2017). Towards a knowledge graph representing research findings by semantifying survey articles. In *International Conference on Theory and Practice of Digital Libraries*, pages 315–327. Springer.

Fathalla, S., Vahdati, S., Auer, S., and Lange, C. (2018). Semsur: a core ontology for the semantic representation of research findings. *Procedia Computer Science*, 137:151–162.

Feynman, R. P., Leighton, R. B., and Sands, M. (1965). The feynman lectures on physics; vol. i. *American Journal of Physics*, 33(9):750–752.

Gangemi, A., Catenacci, C., Ciaramita, M., and Lehmann, J. (2005). Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. *On-line: http://www. loa-cnr. it/Files/OntoEval4OntoDev_Final. pdf*.

Gómez-Pérez, A. (2001). Evaluation of ontologies. *International Journal of intelligent systems*, 16(3):391–409.

Hammond, T. and Pasin, M. (2015). The nature. com ontologies portal. In *LISC@ ISWC*, pages 2–14.

Hey, A. J., Tansley, S., Tolle, K. M., et al. (2009). *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al. (2004). Swrl: A semantic web rule language combining owl and ruleml. *W3C Member submission*, 21(79):1–31.

Jaradeh, M. Y., Auer, S., Prinz, M., Kovtun, V., Kismihók, G., and Stocker, M. (2019a). Open research knowledge graph: Towards machine actionability in scholarly communication. *arXiv preprint arXiv:1901.10816*.

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., and Auer, S. (2019b). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In Kejriwal, M., Szekely, P. A., and Troncy, R., editors, *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, pages 243–246. ACM.

Jaradeh, M. Y., Oelen, A., Prinz, M., Stocker, M., and Auer, S. (2019c). Open research knowledge graph: A system walkthrough. In Doucet, A., Isaac, A., Golub, K., Aalberg, T., and Jatowt, A., editors, *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceed-*

[13]https://projects.tib.eu/orkg/

ings, volume 11799 of *Lecture Notes in Computer Science*, pages 348–351. Springer.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Lovrencic, S. and Cubrilo, M. (2008). Ontology evaluation-comprising verification and validation. In *Central European Conference on Information and Intelligent Systems*, page 1. Faculty of Organization and Informatics Varazdin.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological informatics*, 2(3):279–296.

McGuinness, D. L., Van Harmelen, F., et al. (2004). Owl web ontology language overview. *W3C recommendation*, 10(10):2004.

Medeiros, L., Lauer, T. R., Psaltis, D., and Özel, F. (2018). Principal component analysis as a tool for characterizing black hole images and variability. *The Astrophysical Journal*, 864(1):7.

Musen, M. A. et al. (2015). The protégé project: a look back and a look forward. *AI matters*, 1(4):4.

Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101: A guide to creating your first ontology.

Peroni, S. and Shotton, D. (2018). The spar ontologies. In *International Semantic Web Conference*, pages 119–136. Springer.

Proctor, M. (2011). Drools: a rule engine for complex event processing. In *Proceedings of the 4th international conference on Applications of Graph Transformations with Industrial Relevance*, pages 2–2. Springer-Verlag.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ.

Raskin, R. and Pan, M. (2003). Semantic web for earth and environmental terminology (sweet). In *Proc. of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, volume 25.

Rijgersberg, H., Van Assem, M., and Top, J. (2013). Ontology of units of measure and related concepts. *Semantic Web*, 4(1):3–13.

Rospocher, M., Tonelli, S., Serafini, L., and Pianta, E. (2012). Corpus-based terminological evaluation of ontologies. *Applied Ontology*, 7(4):429–448.

Say, Z., Fathalla, S., Vahdati, S., Lehmann, J., and Auer, S. (2020). Ontology design for pharmaceutical research outcomes. In *International Conference on Theory and Practice of Digital Libraries*, pages 119–132. Springer.

Smith, A. (2019). Physics subject headings (physh). *ISKO Encyclopedia of Knowledge Organization*.

Vogt, L., D'Souza, J., Stocker, M., and Auer, S. (2020). Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 107–116.

Weibel, S., Kunze, J., Lagoze, C., and Wolf, M. (1998). Dublin core metadata for resource discovery. *Internet Engineering Task Force RFC*, 2413(222):132.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.