ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/341679294

Ontology Design for Pharmaceutical Research Outcomes

Preprint · May 2020

DOI: 10.13140/RG.2.2.20218.72644

CITATIONS
0

READS
928

5 authors, including:

Zeynep Say
University of Bonn
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Said Fathalla
University of Bonn
47 PUBLICATIONS **198** CITATIONS
SEE PROFILE

Sahar Vahdati
University of Leipzig
67 PUBLICATIONS **304** CITATIONS
SEE PROFILE

Jens Lehmann
University of Bonn
371 PUBLICATIONS **15,079** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project iASiS: Big Data to Support Precision Medicine and Public Health Policy View project

Project METArchive View project

All content following this page was uploaded by Said Fathalla on 27 May 2020.

The user has requested enhancement of the downloaded file.

Ontology Design for Pharmaceutical Research Outcomes

Zeynep Say[1][0000-0003-4780-6952], Said Fathalla[1,2][0000-0002-2818-5890],
Sahar Vahdati[1,3][000-0002-7171-169X], Jens Lehmann[1,4][0000-0001-9108-4278],
and Sören Auer[5][0000-0002-0698-2864]

[1] Smart Data Analytics (SDA), University of Bonn, Germany
s6zesayy@uni-bonn.de, fathalla@cs.uni-bonn.de, jens.lehmann@cs.uni-bonn.de
[2] Faculty of Science, University of Alexandria, Egypt
[3] Department of Computer Science, University of Oxford, UK
sahar.vahdati@cs.ox.ac.uk
[4] Fraunhofer IAIS, Dresden, Germany
jens.lehmann@iais.fraunhofer.de
[5] TIB Leibniz Information Center for Science and Technology, Hannover, Germany
soeren.auer@tib.eu

Abstract. The network of scholarly publishing includes generating and exchanging ideas, certifying research, publishing and disseminating findings, and preserving outputs. However, the heterogeneous nature and distribution of scholarly data over journal articles, numerous repositories, and libraries make identifying meaningful data a tiresome and manual task. Therefore, transforming document and data structure based on a set of principles and standard recommendations of Semantic Web and Linked Data could reform the data sharing for the scholarly world. In this paper, we present a model (PharmSci) for scholarly publishing in the pharmaceutical research domain with the goal of facilitating knowledge discovery through effective ontology-based data integration. The approach of this paper follows the principles and rules of the ontological engineering development approach. Reasoning and inference based techniques are presented to improve the quality of data integration. Ontology is evaluated with validation and quality verification methods. Our approach represents an agreed model of a particular domain and provides machine-interpretable information to the knowledge discovery process.

Keywords: Semantic Web · Linked Data · OWL Ontologies· Scholarly Communication · Pharmaceutical Research.

1 Introduction

The expansion of the use of digital technologies enables recent developments in academia and shifted the way that scientists research. Figure 1 presents the publication output percentages by field in the world according to the National Science Foundation's (NSF) statistics, and it shows that medical and life science domains produced more publication output than other disciplines of science. Health research disciplines need advances in current big data management approaches [29] since there is a lack of fully Findable, Accessible, Interoperable, and Reusable (FAIR) [32] data resources in the health science domain, especially in pharmaceutical research. Pharmaceutical research has a richness of available sources; however, searching and gaining insight from those resources is not an easy task for a pharmaceutical research scientist. Although structured and well-designed data is easier to be handled by machines, interpreting meaning from unstructured data is not apparent to computers. Thus, there is a need for standards in scholarly communication of pharmaceutical research to prevent these problems. The Semantic Web has emerged to structure and integrate unstructured data on the Web and transform this data into standardized machine-readable formats [2]. Therefore, this work aims to answer research questions: *How can the scholarly pharmaceutical knowledge be supported with a machine-readable and interoperable domain model?* and *How can we increase the reusability and accessibility of pharmaceutical research data more effectively?*

In this paper, we propose an ontology (PharmSci) for modeling pharmaceutical research data. The purpose of the PharmSci ontology is to contribute to pharmaceutical research by making data that is easier to access, reuse, curate, and integrate from documented research towards providing services and unveiling hidden knowledge [27]. Our work helps researchers to find out reliable reference materials, sufficient details of experiments or procedures, and re-investigate experiment results. Besides, this work focuses on solving the challenges of large-scale scholarly data and maximize its usefulness. This model reuses best practices that provide a representation of scientific knowledge to enable interoperability. We followed the rules and principles of Methontology [11] to develop ontology. The ontology coverage is defined with text analysis methods. Ontology reasoning and inference techniques are presented to derive new facts. The developed ontology is evaluated with validation and verification methods. The Pharmaceutical Ontology (PharmSci) is one of the Science Knowledge Graph Ontologies (SKGO) Suite ontologies [8]. The documentation of PharmSci can be found via its PURL(`https://w3id.org/skgo/pharmsci#`) and prefix has been registered at `https://prefix.cc` under the open CC-BY 3.0 license. RDF serializations can be found on GitHub repository[6].

The paper is organized as follows: In section 2, the methodology, and data retrieval technique are presented. The development in section 3 presents the reuse of best practices and the developed conceptual model. Section 4 proposes the evaluation with validation and verification methods. Related work is presented in section 5 for life science and the scholarly domain. Section 6 provides the conclusion and directions for future research work.

2 Methodology

Knowledge Graphs bring enormous opportunities for improving the modern techniques of knowledge discovery. In Figure 2, we envision that pharmaceutical scientist who is investigating the genes involved in multidrug resistance in lung cancer. It shows a knowledge graph of how the publication and research data on

[6] `https://github.com/ZeynepSay/PharmSci`

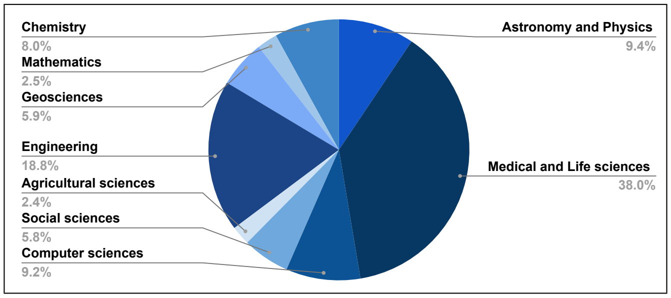


Fig. 1. Scientific publication output percentages by field in the world for the year 2017. (Source: National Science Foundation's (NSF) statistics [31])

the Web can be linked and transformed into a structured domain model. Thus, the goal of PharmSci ontology is to respond to issues of a researcher by interlinking and sharing knowledge of a pharmaceutical research process. We followed the rules and principles of Methontology [11], which is an ontology development method to create domain models. The ontology development lifecycle composes development and supporting activities: specification, conceptualization, development, knowledge acquisition, and evaluation. In the specification phase, we define the domain, data coverage, and tools and techniques for the development of PharmSci. Knowledge acquisition and conceptualization details are explained as follows.

Knowledge Acquisition: The necessary data to create the model can be revealed using text analysis techniques, non-structured or formal interviews with experts, or information acquisition tools. We use text analysis as a knowledge acquisition technique. A corpus[7] is defined with the topic 'multidrug resistance and ABC transporters in cancer'. 200 articles are chosen from pharmaceutical journals in Google Scholar[8] and ScienceDirect[9] related to corpus topic. We reduce 200 articles to 25 articles with a systematic review. We start to choose the most cited articles and eliminate articles that do not include clinical research. Then, we manually analyze the most cited clinical research papers if they cover experimental research. Thus, clinical research papers with experimental research and the highest citation are chosen. First, we identify the common structures in the text. For example, the main parts of the research paper: the objective of the study, the main subjects and subtopics, and the study results. Afterward, we identify the most likely sentence patterns in the article by analyzing its content. For example, "KB-8-5, which is three times as resistant to doxorubicin" [13] sentence in the article, is transformed into "cell line A resistant to drug B". These patterns help us to shape the relations between concepts. Tables, graphs, and figures in articles are analyzed for ascertaining the values of the concept attributes and for identifying certain data regularities.

[7] `https://github.com/ZeynepSay/PharmSci/tree/master/CorpusData`

[8] `https://scholar.google.com/`

[9] `https://www.sciencedirect.com/`

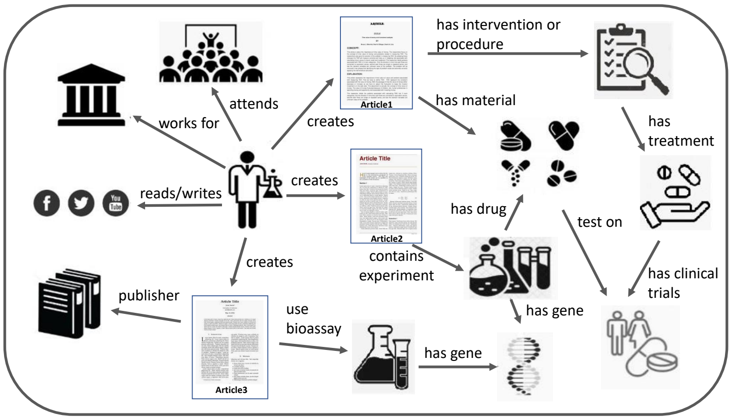


Fig. 2. A knowledge graph of the pharmaceutical research process. In this figure, we can see a scientist who is the creator of three different publications. The content of the articles is transformed into entities and related to other entities with named relationships in a knowledge graph.

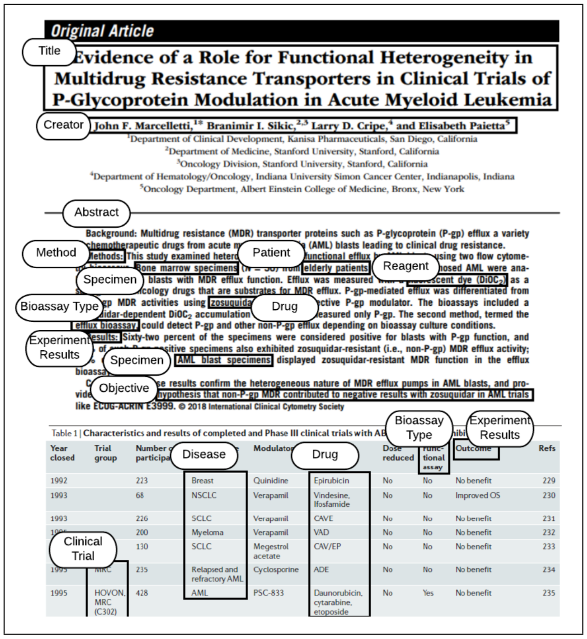
Conceptualization: The conceptual model is designed by organizing and structuring acquired knowledge and converts the informal view of a domain into a semiformal representation by using external representations from external schemas or terminologies. Besides, it includes the analysis of existing data models or ontologies from repositories and the determination of missing classes and properties for the successful formalization of the domain. We build a complete Glossary of Terms as a first thing for the pharmaceutical research domain. All classes and properties gathered in PharmSci glossary of terms to specify usable domain knowledge and its definitions. Figure 3 shows some of the captured classes and instances of the respective classes from pharmaceutical articles.

3 Development

FAIR Principles [32] guide us to increase the value of digital publishings by improving the infrastructure of scientific data. Our aim is to establish an interoperable system by reusing existing best practices. The result of the development is the ontology codified in a formal language. PharmSci is expressed in a W3C standard Web Ontology Language OWL 2[10] and developed by using Protégé v5.5.0[11] [22]. Classes are modelled to represent publication, research activity, clinical study (e.g., clinical trial), material (e.g., reagent), method (e.g., assay), patient, disease, specimen, and informational entities (e.g., objective). Figure 4 represents the main classes, object and data properties, and example instances of PharmSci Ontology. For example, a specific cell line or reagent name used in a particular study is represented as instances of classes in PharmSci ontology.

[10] https://www.w3.org/TR/owl2-overview/

[11] https://protege.stanford.edu/



Original Article

Evidence of a Role for Functional Heterogeneity in Multidrug Resistance Transporters in Clinical Trials of P-Glycoprotein Modulation in Acute Myeloid Leukemia

John F. Marcelletti,[1*] Branimir I. Sikic,[2,3] Larry D. Cripe,[4] and Elisabeth Paietta[5]

[1]Department of Clinical Development, Kanisa Pharmaceuticals, San Diego, California
[2]Department of Medicine, Stanford University, Stanford, California
[3]Oncology Division, Stanford University, Stanford, California
[4]Department of Hematology/Oncology, Indiana University Simon Cancer Center, Indianapolis, Indiana
[5]Oncology Department, Albert Einstein College of Medicine, Bronx, New York

Background: Multidrug resistance (MDR) transporter proteins such as P-glycoprotein (P-gp) efflux a variety chemotherapeutic drugs from acute m a (AML) blasts leading to clinical drug resistance. ethods: This study examined hetero functional efflux b using two flow cytome- Bone marrow specimens from elderly patients hosed AML were ana- blasts with MDR efflux function. Efflux was measured fluorescent dye (DiOC$_2$) as a ncology drugs that are substrates for MDR efflux. P-gp-mediated efflux was differentiated from gp MDR activities using zosuquida ective P-gp modulator. The bioassays included a idar-dependent DiOC$_2$ accumulation easured only P-gp. The second method, termed the efflux bioassay, could detect P-gp and other non-P-gp efflux depending on bioassay culture conditions. esults: Sixty-two percent of the specimens were considered positive for blasts with P-gp function, and of such P-gp-positive specimens also exhibited zosuquidar-resistant (i.e., non-P-gp) MDR efflux activity; AML blast specimens displayed zosuquidar-resistant MDR function in the efflux bioassa

C se results confirm the heterogeneous nature of MDR efflux pumps in AML blasts, and provide hypothesis that non-P-gp MDR contributed to negative results with zosuquidar in AML trials like ECOG-ACRIN E3999. © 2018 International Clinical Cytometry Society

Table 1 | Characteristics and results of completed and Phase III clinical trials with AB hibi

Year closed	Trial group	Number o participa	e	Modulator		Dose reduced	Func-tional assay	Outcome	Refs
1992		223	Breast	Quinidine	Epirubicin	No	No	No benefit	229
1993		68	NSCLC	Verapamil	Vindesine, Ifosfamide	No	No	Improved OS	230
1993		226	SCLC	Verapamil	CAVE	No	No	No benefit	231
199		200	Myeloma	Verapamil	VAD	No	No	No benefit	232
		130	SCLC	Megestrol acetate	CAV/EP	No	No	No benefit	233
1995	MRC	235	Relapsed and refractory AML	Cyclosporine	ADE	No	No	No benefit	234
1995	HOVON, MRC (C302)	428	AML	PSC-833	Daunorubicin, cytarabine, etoposide	No	Yes	No benefit	235

Fig. 3. Captured Entities of PharmSci from scientific publications

3.1 Reuse of Best Practices

In this phase, we consider integrating definitions from already existing semantic models instead of defining them from scratch. We use repositories and open libraries to find terms whose semantic and implementation are coherent with the terms identified in our conceptualization. The repositories and open libraries that we used are Bioportal[12], OntoBee[13], OBOFoundry[14] and Linked Open Vocabularies (LOV)[15] for finding terms in existing ontologies. Table 1 shows the prefixIRI and URL of all reused semantic models in this work. PharmSci follows the National Cancer Institute (NCI) Thesaurus [14] for reusing classes such as `Method(NCIT:C71460)`, `Clinical Study(NCIT:C15206)`, `Material(NCIT:C48187)`, and `Clinical Trial(NCIT:C71104)`. Integrated entities from NCIT and other vocabularies can be seen in Figure 4. `sio:Experiment`, `sio:Specimen`, `sio:sample`, and `sio:investigation` entities were added from Semanticscience Integrated Ontology (SIO) [7] to PharmSci. Besides, PharmSci ontology uses entities that are related to assays, and they were taken from BioAssay Ontology (BAO) [28], such as `experimental setting(bao:BAO_0020005)`, `bioassay(bao:BAO_0000015)`, and `in vitro(bao:BAO_0020008)`. Terms related to chemical substances were imported from Chemical Entities of Biological Interest (ChEBI)[6], for example,

[12] `https://bioportal.bioontology.org/`
[13] `http://www.ontobee.org/`
[14] `http://www.obofoundry.org/`
[15] `https://lov.linkeddata.es/dataset/lov/`

Table 1. Best Practices that are reused in PharmSci Ontology

prefix	URL
NCIT	http://purl.obolibrary.org/obo/ncit.owl
DOID	http://purl.obolibrary.org/obo/doid.owl
bao	http://www.bioassayontology.org/bao#
CHEBI	http://purl.obolibrary.org/obo/chebi.owl
CLO	http://www.ebi.ac.uk/cellline/
terms1	http://ns.nature.com/terms/
foaf	http://xmlns.com/foaf/0.1/
sio	http://semanticscience.org/resource/
terms	https://www.dublincore.org/

drug(CHEBI:23888), reagent(CHEBI:33893), and pharmaceutical(CHEBI:52217). We integrate disease definitions into PharmSci from Human Disease Ontology [21], for example, disease of infectious agent(DOID:0050117) and disease of cellular proliferation(DOID:14566). Cell Line Ontology (CLO) is reused to define cell concepts used in the study. We employ nature publishing group ontologies [16] entities for describing metadata of scholarly domain, such as terms1:Publication, terms1:Publisher, and terms1:Article. DCMI [30] annotations and object properties are reused to link the classes of scholarly publishing domain (dc:title, dc:creator, terms:publisher, etc.). foaf:Person from FOAF Vocabulary [3] is used to define authors of publications and foaf:Organization is used to define publishers in PharmSci. There are also subclass hierarchies in classes, for example, cancer(DOID:162) is subclass of disease of cellular proliferation(DOID:14566).

3.2 Semantic Knowledge Representation in PharmSci

In this section, we attempt to identify distinct "triples" of a publication, clinical study, experiment, methods, and materials classes from the article's sentence structure and then normalize each component to standard terminology. OWL distinguishes properties into two main categories that are object properties and data properties. Object properties and data properties help us to relate entities and transforming data into knowledge. In PharmSci ontology, object properties are used to link individuals to individuals, and datatype properties are used to link individuals to data values. We defined the rdfs:domain and rdfs:range of each property. For example, class terms:Publication describes the pharmaceutical research publications and it is the domain of the object property pharmsci:addressesResearch and the range of this property is Research Activity(NCIT:C15429) class. After specifying the domain and range of all properties, we set the object property relations between the instances. For example, instance pharmsci:ChemotherapyInLungCancer of Treat(NCIT:C70742) class connected to instance pharmsci:Vincristine of class drug(CHEBI:23888) by the object property pharmsci:useDrug.

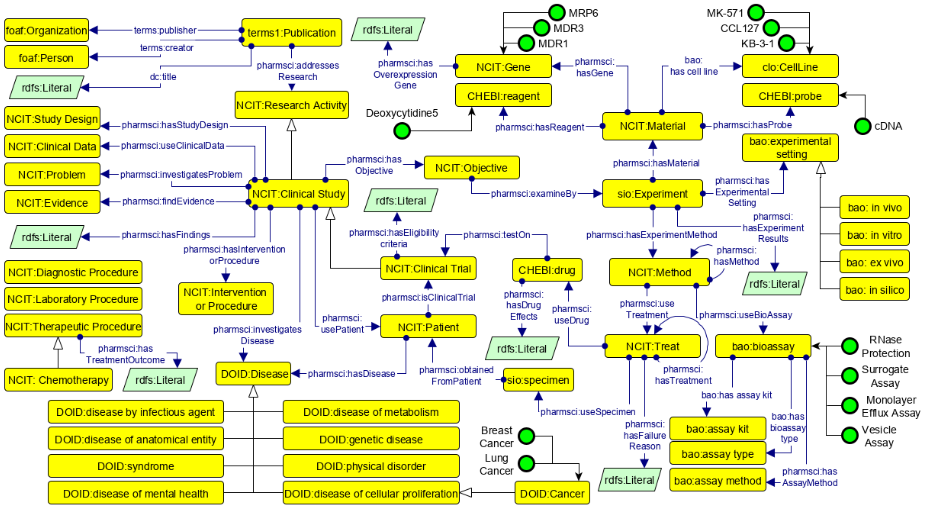


Fig. 4. PharmSci Ontological entities. The integration between the existing ontological entities in PharmSci ontology and relations. In addition, it shows the class Publication and its relation to other classes.

We also defined object properties with different characteristics such as reflexive, irreflexive, inverse, and asymmetric. The class `Material(NCIT:C48187)` is related to other classes with properties which are irreflexive and asymmetric such as `pharmsci:hasGene`, and `pharmsci:hasReagent`. The object properties `pharmsci:hasMethod` is reflexive property because a method can use the same method. The domain `Material(NCIT:C48187)` connects to class `cell line(CLO:0000031)` with object property `has_cell_line(bao:BAO_0002400)` which is inverse of `is_cell_line_of(bao:BAO_0002800)` object property.

Several data properties are defined to link instances to data values. Treatment or therapeutic procedures are an important part of a clinical study. Thus, `pharmsci:treatmentOutcome` and `pharmsci:treatmentFailureReason` are defined with the range `rdfs:Literal`. In addition, `pharmsci:hasDrugEffects` is another data property to define the effects of drug used in the treatment. PharmSci also includes data properties, such as `pharmsci:hasExperiemntResults`, `pharmsci:hasEligibilityCriteria`, and `pharmsci:hasFindings`.

In the knowledge acquisition step, instances are detected as members of the correct target classes and added as instances to the ontology with associated properties. PharmSci includes instances such as the particular patient specimen, treatment types, reagents, genes, probes, assay types, or disease types in research. For example, `pharmsci:daunorubicin`, `pharmsci:vincristine`, `pharmsci:vinblastine`, and `pharmsci:etoposide` are drug types and defined as instances of `drug(CHEBI:23888)` class. Furthermore, `sio:specimen` class has instances in PharmSci, such as `pharmsci:poor_risk_acute_leukemia_samples` and `pharmsci:bone_marrow_aspirates`. Additionally, publication titles, pub-

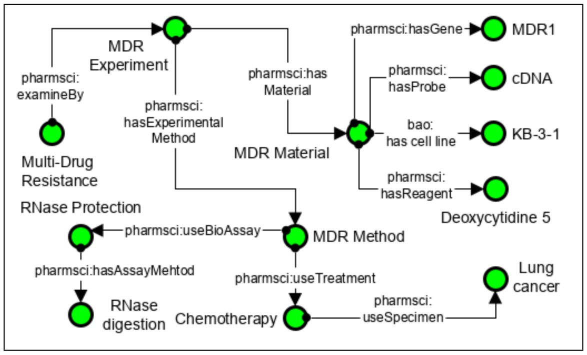


Fig. 5. Instances extracted from the scientific article [13] and their interconnected relationships in PharmSci.

lishers, authors, publication agents, and organizations are added as instances. Some of the instances extracted from the article [13] and their interconnected relations with each other are shown in Figure 5.

3.3 Reasoning and Inference

Reasoning-based approaches are used to derive facts that are not expressed explicitly and increase the expressive power of ontology. We define several SWRL [18] rules in order to infer new logical meanings, and to discover inconsistencies among instances. The rules have been applied with Drools reasoner [24] in Protégé to export new axioms and declarations to instances inside the ontology. The following rule (Equation 1) expresses the fact that clinical study has objective and objective examined by experiment; thus, we can infer that clinical study has an experiment. Drugs used in the treatment can also be part of an experiment material in the studies because treatment is an experimental method as in Equation 2. Recursiveness of the properties is shown Equation 3 and Equation 4, `hasMethod` and `hasTreatment` are reflexive properties.

$$\begin{aligned} ClinicalStudy(?x) \wedge hasObjective(?x, ?z) \wedge examinedBy(?z, ?y) \\ \rightarrow hasExperiment(?x, ?y) \end{aligned} \tag{1}$$

$$\begin{aligned} Experiment(?x) \wedge hasExperimentMethod(?x, ?y) \wedge useDrug(?y, ?z) \\ \rightarrow hasMaterial(?x, ?z) \end{aligned} \tag{2}$$

$$hasMethod(?x, ?z) \wedge hasMethod(?z, ?y) \rightarrow hasMethod(?x, ?y) \tag{3}$$

$$hasTreatment(?x, ?z) \wedge hasTreatment(?z, ?y) \rightarrow hasTreatment(?x, ?y) \tag{4}$$

4 Evaluation

The evaluation step contributes and increases the availability and reusability of our model. This phase includes validation as a first step that guarantees the correctness of an ontology. Verification is the second step to guarantee that the software environment and documentation represent the correct ontology.

4.1 Validation of Ontology

Query Execution of Competency Questions: Competency questions [15] are a list of questions that the knowledge base should be able to answer. We create these questions according to the content of research papers. The results of these questions confirm that the designed model contains enough detail of a particular area. We created 25 competency questions in total, and Table 2 shows the 10 of the competency questions for PharmSci ontology. SPARQL queries have been implemented for each defined competency question. Listing 1.1 is the query for the question Q5 "What is the title of the Publications use the BioAssay 'Efflux Bioassay' as experiment method?" in the PharmSci competency question list.

Listing 1.1. SPARQL query for Q5 in Table 2.

```
SELECT  DISTINCT ?title
WHERE {
?publication   pharmsci:addressesResearch  ?study.
?publication   terms:title                 ?title.
?publication   terms:creator               ?creator.
?study         pharmsci:hasExperiment      ?experiment.
?experiment    pharmsci:hasMethod          ?method.
?method        pharmsci:useBisoassay pharmsci:Efflux_Bioassay
}
```

The possible answer for Listing 1.1 is the publication with the title: *"Different Efflux Transporter Affinity and Metabolism of 99mTc-2-Methoxyisobutylisonitrile and 99mTc-Tetrofosmin for Multidrug Resistance Monitoring in Cancer"*. As a result of this validation phase, competency questions are answered and validated correctly with SPARQL queries.

Comparative Analysis: This approach is used to compare the ontology with the content of a text corpus to check how far an ontology sufficiently covers the given domain. Our approach is to perform an automated term extraction with the latent semantic analysis [5] for the two different corpora. We analyzed the overlapped concepts and counted the number of these words separately for each corpus and ontologies. Then, Precision, Recall, and F1 values are calculated according to the total number of concepts (Keywords) defined in the ontology, most likely terms in analysis results (Class), and the number of matched concepts (Hits) with the corpus.

$$Precision = \frac{|N_{hits}|}{|N_{class}|} \text{ and } Recall = \frac{|N_{hits}|}{|L_{Keywords}|} \text{ and } F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Table 2. Competency questions for PharmSci ontology

Query	Competency Question
Q1	Which Objective examined by Experiment Y for Clinical Study Z?
Q2	Which Clinical Study use the Experiment Method Y for Experimental Material X by using Gene as a material?
Q3	Which Cancer type X is studied by the Clinical Study Y?
Q4	Which Drugs are used in Therapeutic Procedure X and Clinical Study Y for Disease Z?
Q5	What is title of the Publications use the BioAssay Y as Experiment Method?
Q6	Which Cell Lines, Genes, Drugs, Probes are used in Research X?
Q7	Give Publication with Chemotherapy X with drug Y for cancer type Z?
Q8	Give Publication with Experiment Setting X for material Y in Clinical Study Z?
Q9	Which Assay type and assay kit is used for Experiment Method X in Study Y?
Q10	What is the Experiment Setting of Experiment X in Clinical Study Y?

Corpus 1 consists of search results of Google Scholar with the keywords 'multidrug resistance and ABC transporters in cancer'. Corpus 2 gathers 'in vitro evaluation in drug delivery' downloaded from ScienceDirect. Both corpora consist of 25 PDF files and converted to TSV files. (Details of analysis and corpora can be found as .tsv file on Github[16].) We used one of the current ontology in the pharmaceutical domain, which is Drug Interaction Knowledge Base (DIKB) [4], to evaluate how far it satisfies the pharmaceutical research and to compare with PharmSci. We selected 50 most likely words from the latent semantic analysis results (with high TF-IDF weight score) from two corpora, and then we compared these words for both ontologies. Table 3 shows how many words matched with corpora, and the calculations of precision, recall, and F1 value results. The F1 value of PharmSci is greater than DIKB, it is 0.163 for corpus 1 and 0.138 for corpus 2 (see Figure 7). As a result, the PharmSci ontology has more matched concepts than DIKB ontology.

4.2 Verification of Ontology

In this phase, we used the FOCA methodology [1], which has three main steps: Ontology type verification, questions verification, and quality verification. A total of 12 questions should be answered in the question verification step for their respective goal and metric. The expert should score the results of each question. After answering the questions, the expert establishes a grade for each question. Goal 1, 3 and 4 obtain 100% and Goal 2 is 50% for PharmSci (see Figure 6). The result of this evaluation shows that PharmSci received high scores for adaptability, completeness, consistency, clarity, and computational efficiency metrics. However, it needs improvements for the conciseness metric because of moderate abstraction and some reused properties are not used in the model. The total quality of the ontology in the FOCA method is calculated by the beta regression models proposed by Ferrari [12]. The result of the total quality verification

[16] https://github.com/ZeynepSay/PharmSci/tree/master/CorpusData

Table 3. Precision, Recall, and F1 values for PharmSci and DIKB ontology[4]

Corpora	Ontology	Classes	Keywords	Hits	Precision	Recall	F1
Corpus-1	PharmSci	181	50	19	0.10	0.38	0.16
	DIKB	360	50	15	0.04	0.3	0.07
Corpus-2	PharmSci	181	50	16	0.09	0.32	0.14
	DIKB	360	50	12	0.03	0.24	0.06

according to the beta regression model is 0.99423 for PharmSci, which means it has high quality since its value close to 1.

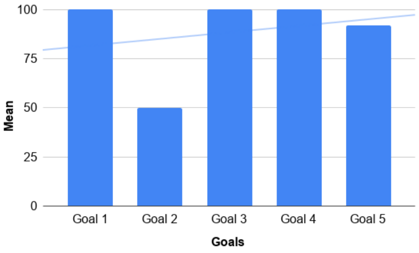


Fig. 6. Goal percentages for PharmSci

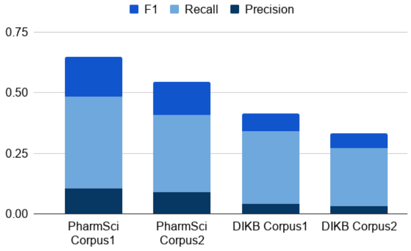


Fig. 7. Comparative Analysis Results

5 Related Work

Much progress in both research and industry has been carried out about representing knowledge in machine-actionable form. Several available vocabularies, platforms, and schemas related to scholarly publishing and life science domain are presented. The Open Research Knowledge Graph [19] is an infrastructure for semantic scholarly knowledge acquisition, publication, processing, and curation. SN SciGraph [17] is the Springer Nature Linked Data platform that provides Linked Open Data as open research. In 2017, an initial step towards representing computer science research data was taken by Fathalla et al. [9]. Subsequently, they developed the Semantic Survey Ontology (Semsur) [10], which semantically captures and represents the knowledge in review and survey articles. SPAR (Semantic Publishing and Referencing) [23] is a set of integral and orthogonal ontologies for defining metadata of the scholarly publication workflow. Another work that deals with information overload is the CSO classifier [25] automatically classifies research papers according to the metadata by using the Computer Science Ontology (CSO). The field of knowledge and data representation in the

[17] https://scigraph.springernature.com/explorer

life science domain is vast. The Open Biomedical Ontologies (OBO) Foundry [26] was founded to address the problem of the proliferation of ontologies. OBO includes over 60 ontologies such as Gene Ontology and Cell Ontology. Medical Subject Headings (MeSH)[18] is a schema vocabulary developed in the field of Medicine. In the pharmaceutical research domain, there are also semantic models, for example, the Drug-Drug Interaction Ontology (DINTO) [17], Drug-drug Interaction Evidence Ontology (DIDEO) [20] and Drug Interaction Knowledge Base (DIKB) [4]. Also, the Drug-Drug Interaction Ontology (DINTO) [17] used the methontology [11] as an ontology development method. However, these models generally focused on drug-drug interaction or drug discovery topics. Thus, PharmSci ontology differs from these domain models and combines scholarly metadata with domain-specific metadata.

6 Conclusion

Science communities start to recognize the significance of semantics in data discovery due to it would provide crucial machine-interpretable information to the knowledge discovery process. The issue of handling, accessing, and representing constant overflow of scientific data can be solved by using Semantic Web-based approaches. Pharmaceutical research data records are one of the most valuable properties for pharmaceutical companies and researchers. Our approach in this work is to structure this knowledge by developing a semantic model that enables us to represent knowledge about a particular study in the pharmaceutical research domain. One of the core impacts of the PharmSci is to add value to the scientific knowledge exploration in pharmaceutical research by describing data with rich metadata. Besides, this work adds value across other research fields because it can be adapted and extended to a vast spectrum of science. Our evaluation results show successful results, and ontology is ready to be used in the implementation of applications. Thus, we envision community-supported semantic models that would enable automated exploration, analysis, understanding, and usage of metadata to gain worthy insight from scientific publications. It is possible to develop a semantic model for other branches of science such as mathematics, physical science, or earth science as future work to allow knowledge extraction from unstructured and structured resources. Furthermore, PharmSci ontology will also be implemented and integrated into a semantic web-based platform Open Research Knowledge Graph (ORKG)[19] as a future work, which is a TIB collaborative project that engages research communities in the development of technologies for open graphs about scientific knowledge.

Acknowledgments

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

[18] https://www.nlm.nih.gov/mesh/meshhome.html

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

References

1. Bandeira, J., Bittencourt, I.I., Espinheira, P., Isotani, S.: Foca: A methodology for ontology evaluation. arXiv preprint arXiv:1612.03353 (2016)
2. Berners-Lee, T., Fischetti, M.: Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor. DIANE Publishing Company (2001)
3. Brickley, D., Miller, L.: Foaf vocabulary specification 0.91 (2007)
4. Brochhausen, M., Schneider, J., Malone, D., Empey, P.E., Hogan, W.R., Boyce, R.D.: Towards a foundational representation of potential drug-drug interaction knowledge. In: First International Workshop on Drug Interaction Knowledge Representation (DIKR-2014) at the International Conference on Biomedical Ontologies (ICBO 2014) (2014)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
6. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: Chebi: a database and ontology for chemical entities of biological interest. Nucleic acids research (2007)
7. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., et al.: The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. Journal of biomedical semantics **5**(1), 14 (2014)
8. Fathalla, S., Auer, S., Lange, C.: Towards the semantic formalization of science. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing. pp. 2057–2059 (2020)
9. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Towards a knowledge graph representing research findings by semantifying survey articles. In: International Conference on Theory and Practice of Digital Libraries. pp. 315–327. Springer (2017)
10. Fathalla, S., Vahdati, S., Auer, S., Lange, C.: Semsur: a core ontology for the semantic representation of research findings. Procedia Computer Science **137**, 151–162 (2018)
11. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering (1997)
12. Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. Journal of applied statistics **31**(7), 799–815 (2004)
13. Fojo, A.T., Ueda, K., Slamon, D.J., Poplack, D., Gottesman, M., Pastan, I.: Expression of a multidrug-resistance gene in human tumors and tissues. Proceedings of the National Academy of Sciences **84**(1), 265–269 (1987)
14. Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J., Parsia, B.: The national cancer institute's thesaurus and ontology. Journal of Web Semantics First Look 1_1_4 (2003)
15. Grüninger, M., Fox, M.S.: Methodology for the design and evaluation of ontologies (1995)
16. Hammond, T., Pasin, M.: The nature. com ontologies portal. In: LISC@ ISWC. pp. 2–14 (2015)
17. Herrero-Zazo, M., Segura-Bedmar, I., Hastings, J., Martínez, P.: Dinto: using owl ontologies and swrl rules to infer drug–drug interactions and their mechanisms. Journal of chemical information and modeling **55**(8), 1698–1707 (2015)

18. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al.: Swrl: A semantic web rule language combining owl and ruleml. W3C Member submission **21**(79), 1–31 (2004)
19. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. pp. 243–246 (2019)
20. Judkins, J., Tay-Sontheimer, J., Boyce, R.D., Brochhausen, M.: Extending the dideo ontology to include entities from the natural product drug interaction domain of discourse. Journal of biomedical semantics **9**(1), 15 (2018)
21. Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., et al.: Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research **43**(D1), D1071–D1078 (2014)
22. Musen, M.A.: The protégé project: a look back and a look forward. AI matters **1**(4), 4–12 (2015)
23. Peroni, S., Shotton, D.: The spar ontologies. In: International Semantic Web Conference. pp. 119–136. Springer (2018)
24. Proctor, M.: Drools: a rule engine for complex event processing. In: Proceedings of the 4th international conference on Applications of Graph Transformations with Industrial Relevance. pp. 2–2. Springer-Verlag (2011)
25. Salatino, A.A., Osborne, F., Thanapalasingam, T., Motta, E.: The cso classifier: Ontology-driven detection of research topics in scholarly articles. In: International Conference on Theory and Practice of Digital Libraries. pp. 296–311. Springer (2019)
26. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11), 1251 (2007)
27. Vahdati, S., Palma, G., Nath, R.J., Lange, C., Auer, S., Vidal, M.E.: Unveiling scholarly communities over knowledge graphs. In: International Conference on Theory and Practice of Digital Libraries. pp. 103–115. Springer (2018)
28. Visser, U., Abeyruwan, S., Vempati, U., Smith, R.P., Lemmon, V., Schürer, S.C.: Bioassay ontology (bao): a semantic description of bioassays and high-throughput screening results. BMC bioinformatics **12**(1), 257 (2011)
29. Wang, X., Williams, C., Liu, Z.H., Croghan, J.: Big data management challenges in health research—a literature review. Briefings in bioinformatics **20**(1), 156–167 (2017)
30. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. Internet Engineering Task Force RFC **2413**(222), 132 (1998)
31. White, K.: Science and engineering publication output trends: 2017 shows us output level slightly below that of china but the united states maintains lead with highly cited publications. National Center for Science and Engineering Statistics, National Science Foundation. Alexandria, VA. (2019)
32. 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.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)

View publication stats