Towards the Semantic Formalization of Science

Said Fathalla
Faculty of Science, University of Alexandria
Alexandria, Egypt
Smart Data Analytics (SDA), University of Bonn
Bonn, Germany
fathalla@cs.uni-bonn.de

Sören Auer
soeren.auer@tib.eu
TIB Leibniz Information Centre for Science and Technology
L3S Research Center, University of Hannover, Hannover, Germany

Christoph Lange
Information Systems, RWTH Aachen University, Germany
Fraunhofer FIT, Germany
lange@cs.rwth-aachen.de

ABSTRACT

The past decades have witnessed a huge growth in scholarly information published on the Web, mostly in unstructured or semi-structured formats, which hampers scientific literature exploration and scientometric studies. Past studies on ontologies for structuring scholarly information focused on describing scholarly articles' components, such as document structure, metadata and bibliographies, rather than the scientific work itself. Over the past four years, we have been developing the Science Knowledge Graph Ontologies (SKGO), a set of ontologies for modeling the research findings in various fields of modern science resulting in a knowledge graph. Here, we introduce this ontology suite and discuss the design considerations taken into account during its development. We deem that within the next years, a science knowledge graph is likely to become a crucial component for organizing and exploring scientific work.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; • **Information systems** → **Information extraction**;

KEYWORDS

Knowledge Graphs, Knowledge Capture, Semantic Metadata Enrichment, Scholarly Communication

ACM Reference Format:
Said Fathalla, Sören Auer, and Christoph Lange. 2020. Towards the Semantic Formalization of Science. preprint.

1 INTRODUCTION

This plethora of scientific literature makes it intractable to obtain an overview of the current state of research results in different science disciplines. Currently, the unstructured or semi-structured representation of research data published on the Web still has deficiencies – the content is not represented in a formal, machine-comprehensible way, which prevents conceptualization problems as well as the building of intelligent search, exploration and browsing applications on top. Modern information systems could support knowledge discovery applications such as scientific exploration, thanks to highly structured ontologies [1]. However, most existing scholarly communication infrastructures use traditional, keyword-based information retrieval. Subsequently, knowledge-based representation of scholarly data, which motivates the development of data models, ontologies and knowledge graphs, yields a richer representation of this data, thus making it easier to query and process [2]. The vision of establishing scholarly communication in a knowledge-based way makes analysis and exploration of scientific data in digital libraries both easier and more efficient than it is now. Our efforts aim at increasing the impact that researchers can make, by enabling them to immediately contribute to a common knowledge base comprising comprehensive descriptions of their research, thus making research contributions transparent and directly comparable. In 2017 we made an initial step [3] towards this goal with the comprehensive Semantic Survey Ontology (SemSur) for capturing the content of computer science survey articles. SemSur [4] generally defines how surveys for research fields can be represented semantically, resulting in a knowledge graph that represents individual research problems, approaches, implementations and evaluations in a structured and comparable way.

Here, we introduce the Science Knowledge Graph Ontologies (SKGO), which capture the knowledge of scientific information typically presented in publications by interlinking domain-specific information in a highly structured way, thus enabling access to these data in a machine-readable, transparent and comparable manner. Currently, SKGO comprises ontologies for five fields of science, including computer science (SemSur), chemistry (ChemSci), physics (PhySci), dentistry (DentSci), and pharmaceutics (PharmSci), as well as an upper ontology on top, called Modern Science Ontology (ModSci). This will support the digital transformation of scholarly communication from documents to a knowledge-oriented representation in the form of structured and interlinked knowledge graphs, aiming at analyzing, exchanging and exploiting scholarly knowledge efficiently. The SKGO ontologies have been made publicly available in standard formats, at permanent URLs, following best practices for publishing ontologies [5] and FAIR data [6]. This paper presents the first complete overview of the SKGO ontology suite as well as the ModSci ontology.

2 RELATED WORK

Most previous approaches, such as the Semantic Publishing and Referencing (SPAR) Ontologies [7] and the Journal Article Tag Suite [8], have focused on particular aspects of scholarly articles,

such as their structural components or bibliographic information within them, rather than the scientific results themselves. In this section, we present research efforts on developing ontologies for modeling research findings in different fields of science, including biology, computer science, dentistry and chemistry.

Computer Science-related ontologies: *SemSur* is a comprehensive ontology for capturing the content of computer science articles including research problems, implementations used, experiment setup, etc. [3]. This supports efficient exploration and comparison of research findings based on an explicit semantic representation of the knowledge contained in articles. The Computer Science Ontology [9] is an ontology for characterizing higher-level Computer Science research areas and their corresponding sub-topics and related terms. *Semantic Web for Research Communities* (SWRC) is an ontology for representing knowledge about researchers and research communities such as persons, organizations, publications and their relationships [10]. The *scientific EXPeriments Ontology* (EXPO)[1] formalizes the generic concepts of scientific experiments, such as experiment setup, experimental design, goal and results [11].

Natural Science-related ontologies: The chemical information ontology (cheminf) represents chemical entities and richly describes chemical properties, intrinsic or computed, such as chemical descriptor, boiling point, and molecular descriptor [12]. CombeChem [13] is a chemical ontology capturing some aspects of chemical structure, such as organic molecules, molecular properties, as well as scientific units. Konyk et al. [14] represented chemical knowledge such as types of objects, e.g., molecules, atoms, or rings, as well as their connectivity and qualities. Their knowledge base integrates several data sources, including DrugBank, OpenBabel and DBpedia. The Oral Health and Disease Ontology (OHD)[2] is used to describe dental practice health records and designed for use in translation medicine. The Ontology for Biomedical Investigations (OBI) [15] is an ontology for describing all aspects of how investigations in the biological and medical domains are conducted. Semantic Web for Earth and Environmental Terminology (SWEET) [16] is used to model knowledge about earth system science and related concepts.

3 SCIENCE KNOWLEDGE GRAPH ONTOLOGIES: THE SUITE

Modern science is commonly divided into three major branches: natural sciences, social sciences, and formal sciences. Each of them comprises various specialized yet overlapping disciplines, which often possess their own nomenclature and expertise [17]. For instance, *ecology* is a new branch of biology dealing with the relations of organisms to one another and to their physical surroundings, i.e., overlapping with earth sciences. Modern science follows a set of core procedures or rules to determine the nature and underlying natural laws of the universe. This requires collaborations between scientists from different fields of science. For example, biologists require mathematics to process, analyze and report experimental research data and to represent relationships between some biological phenomena. Statistics is used in economics to measure correlation, to analyze demand and supply, and to forecast through regression, interpolation and time series analysis. The Science Knowledge Graph Ontologies (SKGO) are a suite of OWL ontologies to capture the knowledge of scientific information from publications, by interlinking domain-specific information, and to make such data accessible in a machine-readable, transparent and comparable way.

3.1 Development Methodology

The Systematic Approach for Building Ontologies (SABiO) [18] has been followed when developing SKGO. It comprises five phases: 1) ontology purpose identification and requirements elicitation, 2) ontology capture and formalization, 3) ontology design, 4) ontology implementation, and 5) ontology testing. SKGO is being created and validated through cross-disciplinary interaction between ontology experts and researchers belonging to the respective fields of science. Several Ontology Design Patterns [19] have been applied in SKGO, such as the OWL patterns of Gangemi [20], which are used to capture inverse relations and composition of relations. SKGO ontologies are available in multiple RDF serializations, including Turtle, JSON-LD, RDF/XML and N-Triples, from a GitHub repository.

3.2 Characteristics

The following design considerations have been taken into account in the development of SKGO:

- *Publication.* The SKGO ontologies are published with dereferenceable URIs, including human-readable HTML content, using the recipes provided in [5].
- *Availability:* All SKGO ontologies have been published under a persistent URL (https://w3id.org/skgo/{mod,dent, chem,phy,pharm}sci# and http://purl.org/semsur/owl/) under the open CC-BY 4.0 license. The source is available from a *GitHub* repository (http://tiny.cc/4l22dz).
- *Readability:* The Widoco wizard for documenting ontologies [21] is used to create HTML documentation, thus facilitating human understandability of the ontologies. The documentation for each SKGO ontology is accessible through its PURL.
- *Sustainability:* The SKGO ontologies are planned to be integrated into the Open Research Knowledge Graph[3]. The issue tracker on GitHub helps to get feedback, suggestions for improvement, e.g., re-using related ontologies that may appear in future, and to report any problems.
- *Metadata:* A checklist has been used to complete the ontologies' metadata [22].

3.3 Ontologies

In this section, we briefly describe those SKGO ontologies that have not been published before. For the generic metadata of research articles, e.g., title, description, and creators, the Dublin Core Metadata Initiative and schema.org vocabularies are used in all SKGO ontologies.

The pivotal concepts of the **Modern Science Ontology (ModSci)** are the branches and sub-branches of modern science. Further concepts include (following the definitions of [23]): modern science, scientific discovery, phenomenon, scientific organization, scientist, and scientific instrument. The **Chemistry Ontology (ChemSci)** models high-level descriptions of experiments, such as experiment material,

[1] http://expo.sourceforge.net/

[2] https://github.com/oral-health-and-disease-ontologies/ohd-ontology

[3] http://orkg.org/

instruments, and reaction type. The **Dentistry Ontology (DentSci)** models abstract descriptions of the main dentistry research procedures, including treatment phases, eligibility criteria of involved patients, and experiment equipment. The Pharmaceutical Science ontology (PharmSci) combines a broad range of scientific concepts related to Drug development, including clinical study, drug, experiment, material, drug effect, and disease. The **Physics Ontology (PhySci)** is designed to characterize the description of the scientific data based on the Physics research domain for both experimental and theoretical studies, including metadata about scientific observations and measurements.

4 EVALUATION

To evaluate SKGO ontologies, we performed ontology Verification & Validation (V&V) following the SABiO guidelines [18]. Verification has been performed by human evaluation, i.e., by means of expert judgment, in which the concepts, relations and axioms defined in the ontology have been checked regarding whether they are able to answer a predefined set of competency questions (CQs) [24]. This approach enabled us not only to check whether the ontology can answer the CQs, but also whether there are irrelevant (should be removed) or missing (should be added) terms in the ontology. Therefore, we performed this evaluation step in parallel with the ontology development in an iterative manner, which significantly helped in improving the ontology. After several iterations (development-to-evaluation and vice versa), we obtained the final version of the ontologies. In parallel, we verified that the SKGO ontologies are able to answer all competency questions defined by determining terms that matched the CQs by querying the ontologies.

5 CONCLUSIONS AND FUTURE WORK

This paper is the first that presents a full view of the suite of Science Knowledge Graph Ontologies (SKGO). Currently, SKGO comprises five ontologies for modeling scientific work in various fields of science, including computer science, chemistry, physics, dentistry, and pharmaceutical science as well as the Modern Science Ontology. The SKGO ontologies have been made publicly available in standard formats, at a permanent URL, following best practices for ontology publication and vocabulary metadata completion. Several design principles have been taken into consideration in the development of the SKGO ontologies, such as publication and configuration to support semantic web applications, registration in online services for ontology visualization and exploration, validation, creation of human readable documentation, and sustainability. The SABiO methodology has been followed when developing the ontologies, as well as FAIR principles for data publication. Ontology verification by experts and the ontology development have been performed in parallel in an iterative manner, which significantly helped in improving the ontology. Finally, we hope that SKGO constitutes a significant step towards facilitating the representation and analysis of scholarly data in various fields of science, thus supporting the transition from document- to knowledge-based scholarly communication.

Our future work has three main directions: refining the formal representation of science in SKGO, covering further fields of science by dedicated ontologies, and realizing services on top of these ontologies. We are planning to complement the ontologies in SKGO by integrating more related data models and adding more ontologies to model research data in other fields of science, such as earth sciences, biology, as well as mathematics. Finally, to boost real-world applications of SKGO, we are planning to realize knowledge management and e-research services on top of the Open Research Knowledge Graph[3], into which we will integrate the SKGO ontologies.

ACKNOWLEDGMENTS

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

REFERENCES

[1] William Pike and Mark Gahegan. 2007. Beyond ontologies: toward situated representations of scientific knowledge. *Human-Computer Studies*, 65, 7.
[2] Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. 2018. Towards a knowledge graph for science. In *WIMS*.
[3] Said Fathalla, Sahar Vahdati, Sören Auer, and Christoph Lange. 2017. Towards a knowledge graph representing research findings by semantifying survey articles. In *TPDL*. Springer.
[4] Said Fathalla, Sahar Vahdati, Sören Auer, and Christoph Lange. 2018. Semsur: a core ontology for the semantic representation of research findings. *Procedia Computer Science*, 137.
[5] Diego Berrueta, Jon Phipps, Alistair Miles, Thomas Baker, and Ralph Swick. 2008. Best practice recipes for publishing rdf vocabularies. *Working draft, W3C*. http://www.w3.org/TR/swbp-vocab-pub/.
[6] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3.
[7] Silvio Peroni and David Shotton. 2018. The SPAR ontologies. In *ISWC*.
[8] Paul Donohoe, Jenny Sherman, and Ashwin Mistry. 2015. The long road to JATS. In *JATS-Con*.
[9] Angelo A Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. 2018. The computer science ontology: a large-scale taxonomy of research areas. In *ISWC*. Springer, 187–205.
[10] York Sure, Stephan Bloehdorn, Peter Haase, Jens Hartmann, and Daniel Oberle. 2005. The SWRC ontology – semantic web for research communities. In *EPIA*.
[11] Larisa N Soldatova and Ross D King. 2006. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3, 11.
[12] Janna Hastings, Leonid Chepelev, Egon Willighagen, Nico Adams, Christoph Steinbeck, and Michel Dumontier. 2011. The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PloS one*, 6, 10.
[13] Kieron R Taylor, Jonathan W Essex, Jeremy G Frey, Hugo R Mills, G Hughes, and EJ Zaluska. 2006. The semantic grid and chemistry: experiences with CombeChem. *Web Semantics*, 4, 2.
[14] Mykola Konyk, Alexander De Leon, and Michel Dumontier. 2008. Chemical knowledge for the semantic web. In *DILS*.
[15] Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H Brush, Bill Bug, Marcus C Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, Michel Dumontier, et al. 2016. The ontology for biomedical investigations. *PloS one*, 11, 4.
[16] Robert G Raskin and Michael J Pan. 2005. Knowledge representation in the semantic web for earth and environmental terminology (SWEET). *Computers & geosciences*, 31, 9.
[17] K Boyack, D Klavans, WB Paley, and K Börner. 2007. Scientific method: relationships among scientific paradigms. *Seed Magazine*, 9.
[18] Ricardo de Almeida Falbo. 2014. SABiO: systematic approach for building ontologies. In *1st Joint Workshop Onto.Com/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering*.
[19] Valentina Presutti and Aldo Gangemi. 2008. Content ontology design patterns as practical building blocks for web ontologies. In *ER*.
[20] Aldo Gangemi and Valentina Presutti. 2009. Ontology design patterns. In *Handbook on ontologies*. Springer.
[21] Daniel Garijo. 2017. Widoco: a wizard for documenting ontologies. In *ISWC*.
[22] Daniel Garijo and María Poveda-Villalón. 2017. A checklist for complete vocabulary metadata. (2017). https://w3id.org/widoco/bestPractices.
[23] Wikipedia contributors. 2019. Science — Wikipedia, the free encyclopedia. [Online; accessed 7-October-2019]. (2019). https://en.wikipedia.org/w/index.php?title=Science&oldid=918085492.
[24] Janez Brank, Marko Grobelnik, and Dunja Mladenic. 2005. A survey of ontology evaluation techniques. In *Data Mining and Data Warehouses*.