Toward Representing Research Contributions in Scholarly Knowledge Graphs Using Knowledge Graph Cells

Conference Paper · August 2020

DOI: 10.1145/3383583.3398530

4 authors:

Lars Vogt
Leibniz Information Centre for Science and Technology University Library
150 PUBLICATIONS **1,615** CITATIONS
SEE PROFILE

Jennifer D'Souza
Leibniz Information Centre for Science and Technology University Library
41 PUBLICATIONS **192** CITATIONS
SEE PROFILE

Markus Stocker
Leibniz Information Centre for Science and Technology University Library
76 PUBLICATIONS **941** CITATIONS
SEE PROFILE

Sören Auer
Leibniz Universität Hannover
508 PUBLICATIONS **15,377** CITATIONS
SEE PROFILE

Toward Representing Research Contributions in Scholarly Knowledge Graphs Using Knowledge Graph Cells

Lars Vogt
TIB Leibniz Information Centre for Science and Technology
Hannover, Germany
Lars.Vogt@tib.eu

Jennifer D'Souza
TIB Leibniz Information Centre for Science and Technology
Hannover, Germany
Jennifer.DSouza@tib.eu

Markus Stocker
TIB Leibniz Information Centre for Science and Technology
Hannover, Germany
Markus.Stocker@tib.eu

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

ABSTRACT

There is currently a gap between the natural language expression of scholarly publications and their structured semantic content modeling to enable intelligent content search. With the volume of research growing exponentially every year, a search feature operating over semantically structured content is compelling. Toward this end, in this work, we propose a novel semantic data model for modeling the *contribution* of scientific investigations. Our model, i.e. the Research Contribution Model (RCM), includes a schema of pertinent concepts highlighting six core information units, viz. Objective, Method, Activity, Agent, Material, and Result, on which the *contribution* hinges. It comprises bottom-up design considerations made from three scientific domains, viz. Medicine, Computer Science, and Agriculture, which we highlight as case studies. For its implementation in a knowledge graph application we introduce the idea of building blocks called Knowledge Graph Cells (KGC), which provide the following characteristics: (1) they limit the expressibility of ontologies to what is relevant in a knowledge graph regarding specific concepts on the theme of research contributions; (2) they are expressible via ABox and TBox expressions; (3) they enforce a certain level of data consistency by ensuring that a uniform modeling scheme is followed through rules and input controls; (4) they organize the knowledge graph into named graphs; (5) they provide information for the front end for displaying the knowledge graph in a human-readable form such as HTML pages; and (6) they can be seamlessly integrated into any existing publishing process that supports form-based input abstracting its semantic technicalities including RDF semantification from the user. Thus RCM joins the trend of existing work toward enhanced digitalization of scholarly publication enabled by an RDF semantification as a knowledge graph fostering the evolution of the scholarly publications beyond written text.

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/08…$15.00
https://doi.org/10.1145/3383583.3398530

CCS CONCEPTS

• **Computing methodologies** → **Semantic networks**; *Ontology engineering*; • **Information systems** → *Semantic web description languages*; **Content analysis and feature selection**; **Ontologies**; Digital libraries and archives.

KEYWORDS

open science, semantic publishing, digital libraries, ontology, scholarly infrastructure, machine actionability, FAIR data principles

ACM Reference Format:
Lars Vogt, Jennifer D'Souza, Markus Stocker, and Sören Auer. 2020. Toward Representing Research Contributions in Scholarly Knowledge Graphs Using Knowledge Graph Cells. In *ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3383583.3398530

1 INTRODUCTION

Since the appearance of the first scientific journals in the late 17th century, document-centric communication represented the main mode of scholarly publications for centuries. In other words, scholarly knowledge in the form of new ideas and methods, standards and best practices, the description of new phenomena and data remain predominantly *unstructured* w.r.t. their machine interpretability, despite their long-lived existence as digital records on the web. To the computer, scholarly content is semantically just an index of keywords, which clearly leaves buried layers of their rich content. In the scope of current semantic technologies, i.e. an evergrowing network of ontologies and the better expressivity in semantic modeling languages, there is ample opportunity to develop an enriched semantic structure of the scholarly record thereby fueling scholarly data access applications where machines more intelligently assist the researcher.

Researchers today are faced with a deluge of scientific literature– in 2009, the 50 millionth mark of the total number of science papers published since 1665 was passed, and approximately 2.5 million new scientific papers are published each year [21]. In this present

scenario, the task of systematic literature reviews [25], which involve summarising vast amounts of investigations on a specific topic, even in one's own narrow discipline, is becoming practically impossible. The problem essentially lies in having to scan through the content of dozens, sometimes hundreds, of articles to glean insights into the scientific inquiry [24]. In such systematized reviews, one often looks for the *contribution* of an investigation. In other words, one looks for *the result of the investigation that contributes towards the advancement of scientific human knowledge by adding something new.*

Along the lines of the preceding discussion, within the broad spectrum of the content in the research record that are candidates for semantic structuring, in this paper, we show how semantic technologies can be leveraged to semantically structure the *contribution* of an investigation which is presently conveyed in the abstract of a scientific article. Specifically, we propose an expert-designed semantic model, the **Research Contribution Model (RCM)**, that targets the structured recording of *contributions* of an investigation, which were exemplified over three scientific disciplines, viz. Agriculture, Computer Science, and Medicine, where their models are presented in this article narrative as case studies. We selected these domains since they are significantly disparate in content and hence their analysis offers insights about their common semantic knowledge units for structuring contributions. While within the framework of the RCM, standard semantic technology in the form of ontology terms are incorporated, we also introduce a novel concept of building blocks for knowledge graph applications, called **Knowledge Graph Cells (KGC)** since they inform the creation of knowledge graphs, and show how the RCM can be implemented in a knowledge graph using an ontology together with a set of defined KGCs.

In the broader context of scholarly data, the FAIR guiding principles for scientific data management and stewardship [40] identify general guidelines for making data and metadata machine-actionable by making them maximally Findable, Accessible, Interoperable, and Reusable for machines and humans alike. Semantic Web technologies such as the W3C recommendations Resource Description Framework (RDF) and Web Ontology Language (OWL) are the most widely-accepted choice for implementing the FAIR guiding principles [19]. Ontologies and other controlled vocabularies are important because they enable providing data and metadata in the standardized semantic structure that eScience and the FAIR guiding principles require [10, 36, 38, 39]. The FAIR principles were followed in the specification of the set of KGCs for implementing the RCM in a strict interpretation.

In the rest of the article, we present the **RCM** and introduce the concept of building blocks for knowledge graph applications, the **KGC**. The combination of RCM as a model for structuring semantic representations of *contributions* from scholarly publications and a set of defined KGCs that facilitates the creation of Scholarly Knowledge Graphs (SKG) will inform the development of the **Open Research Knowledge Graph**[1] **(ORKG)** [6, 7, 20] for the representation of research contributions.

[1]https://www.orkg.org/orkg/

2 MOTIVATING EXAMPLE

We use a scenario to motivate our approach. *Sarah* is a fresh graduate student in Computer Science and is looking for the research frontier around Computer Vision techniques for classifying MRI images of the brain. Specifically, she needs to compare results around similar metrics, tools, experimental datasets, and a methods analysis in terms of constituent functional building blocks, and the training and testing parameters of the algorithms.[2] After several weeks or months of search and assimilation of existing literature on the topic, she has all the needed information. She collects together and publishes this information in semantically structured form using ontologized concept templates, i.e. our KGCs, that are specialized to model research contributions, which generate a knowledge graph within the RCM formalism, so that others can 1) curate the information that *Sarah* has obtained; 2) query the semantically structured data from her survey at various granularities of information; 3) access the data and tools through the web links that *Sarah* has found; and 4) reuse the comparison that *Sarah* created in a research paper with supporting edit and direct export functionality in the paper's format (e.g., LaTeX) fostering other features such as citations to her survey. While *Sarah* could have alternatively published her survey as a research article in the traditional unstructured form, with the new structured model she guarantees that the machine interpretability of her survey content will not be restrained only to a set of a keywords, but will now tap into the entire rich survey data given its wholesome semantic structured form. Further, by the enhanced querying now enabled over the semantically structured data, computers can now more intelligently assist others who access the survey by offering different views of the data and at different granularities. Finally, *Sarah* embeds links to the knowledge graph in other digital publishing portals such as scientific blogs for wider access audience.

In the future, other researchers are no longer faced with the daunting obstacle of scouring through an overwhelming number of papers on the topic. With *Sarah*'s existing knowledge graph, they can reuse her survey results. They can then deconstruct *Sarah*'s graph, tap into the aspects they are interested in, and can enhance it for their purposes.

3 RESEARCH CONTRIBUTION SEMANTICS

For the remainder of the paper, we relegate our focus to discussing our proposed structured semantic units that translate aspects of *contributions* of scholarly publications from unstructured text into equivalent machine-actionable structured representations–as opposed to the contributors themselves as was the inclination in prior work (e.g., https://casrai.org/credit/). The units we propose, by their design characteristics, serve as intermediaries for the end-vision of building and organizing a Scholarly Knowledge Graph (SKG).

Our toolbox comprises the following three semantic constructs: 1) a set of defined concepts relevant for modeling the contributions of scholarly publications, organized in an ontology; 2) a semantic construct that facilitates building an SKG, which we call Knowledge Graph Cells (KGC), that consists of a set of quad templates that

[2]Consider that the scale of such research is not small. OpenNeuro (https://openneuro.org/) itself lists over 300 datasets on brain scans. This factor is only expanded given the plethora of applicable machine learning methods.

organize and structure the process of instantiating the defined concepts or to add new concepts to our ontology; and finally, 3) resulting from the specification of KGCs, the semantic Research Contribution Model (RCM) for modeling research contributions from scholarly publications.

3.1 Set of Defined Concepts

For semantically modeling research contributions, we require a set of defined concepts that we either reuse from existing ontologies or create, if none exist, for our specific aim of modeling a *contribution* of a scientific investigation. All our selected concepts for modeling the *contribution* are then placed in our ontology, referred as ***skg*** ontology.[3] Almost all top-level concepts in the *skg* ontology are derived from the Basic Formal Ontology (BFO) [5, 32] since it provides concepts that are abstract and generically applicable for classifying and describing entities from all kinds of domains. BFO provides an upper ontology upon which all ontologies of the Open Biomedical Ontologies Foundry (OBO) [33] are built, where the latter provides the most comprehensive set of ontologies for the life sciences and beyond, that are to a certain degree interoperable and comparable. The top-level concepts in our current schema are: *iao:information content entity*; *bfo:material entity*; *foaf:agent*; *bfo:process*; *bfo:quality*; *bfo:relational quality*; *bfo:disposition*; *bfo:role*; *bfo:site*; *bfo:temporal region*.[4] The definitions of these concepts are the same as in their original source ontology. While we do not exhaustively list all the defined concepts in our ontology, we define in Table 1 those that are predominant. To model specifically the *contribution* of a publication, we introduced six core concepts, viz. ResearchActivity, ResearchObjective, ResearchAgent, ResearchMaterial, ResearchMethod, and ResearchResult. While all concepts, even those that are derived from existing ontologies are ontologized to the *skg* ontology, the newly introduced concepts, such as the six we listed, only pertain to *skg*.

3.2 Knowledge Graph Cell (KGC)

We now introduce the idea of a Knowledge Graph Cell (KGC) which is a building block for the modular specification of an SKG, and thus, can serve as a semantic model for knowledge graph applications such as the ORKG.[5] A KGC is defined by specifying 1) the defined concept it pertains to, which includes all the concept's sub-concepts in the ontology, 2) the set of allowed properties, and 3) specifications facilitating the generation of human-readable interfaces for populating the knowledge graph and for representing its contents in, e.g., HTML pages. 2) and 3) are specified as a set of quad templates. Quad templates provide template specifications for a KGC that model all contents in which a specific concept or an individual instance thereof takes the *Subject* position in a quad statement. This way, a KGC restricts the possible space of statements about a given concept and its instances to the set of attributes that can be recorded for them. A Quad is an RDF triple consisting of *Subject*, *Predicate* and *Object* to which an IRI is added in its fourth position. This fourth position specifies the named graph to which the triple belongs, turning the triple into a quad. A set of triple statements with the same IRI in the fourth position constitute a named graph. We use quads instead of triples for modeling contents of scholarly publications because named graphs enable partitioning data in an RDF store, which facilitates searching for specific contents in the store, and allow for making (i) statements about statements comparable to RDF reification, but outperforming it for more complex queries [14] and (ii) statements about collections of statements.

We distinguish two types of KGCs which are closely related to each other: **ABox KGCs** for creating instance-based semantic graphs and **TBox KGCs** for creating class-based semantic graphs. The former are used for populating the ontology with instances and thus for representing assertions about individuals in the form of descriptive information (i.e., facts or empirical data), whereas the latter are used for adding new classes to the ontology and thus for representing universal statements about kinds or types of things such as definitions and explanatory hypotheses or theories. We posit that generating an SKG would need both.

In the context of knowledge graph applications, KGCs are called by the front end, e.g. via user input. Calling a KGC always results in adding quads to a graph or modifying existing quads. We call this process *instantiating a KGC*. Instantiating a KGC requires the IRI of the defined concept (i.e. named ontology class) for TBox KGCs or that of an instance of a defined concept for ABox KGCs (from here on we refer to this IRI as variable *C*). *C* takes the *Subject* position in the quads to be added or modified. *C* is provided through the application or through user input.

In the following, we will present ABox KGCs since TBox KGCs are not significantly different from their ABox counterpart. Each quad template within an ABox KGC follows the formalism:

$$S \ / \ P \ / \ O / \ NG \ / \ [x..y] \ / \ inv$$

where, ***S*** constitutes the *Subject* of a quad and is occupied by *C*; ***P*** constitutes one or more mutually exclusive properties; ***O*** is either a datatype (in case *P* is a data property), or one or more IRIs (if *P* is an object property)–*O* thus restricts the possible value space of the *Object* of the quad and its actual value is provided via user input; ***NG*** specifies one or more IRIs, whereby the triple defined in this quad template must be stored separately at each specified IRI, with *ir* being an identified resource that is either *C* or an IRI that has been explicitly forwarded through the instantiation of another KGC; ***[x..y]*** constitutes the cardinality of the quad template with *x* specifying if the statement is mandatory or not, i.e. as value 1 or 0, and *y* specifying how many times this quad template can be instantiated for the same *C*, or in other words, how many quads with the same *S* and *P* may be created (i.e., 1=once; m=multiple times)–this usage of cardinality is comparable to that in SHACL; ***inv*** specifies the inverse property of *P* and at the same time points to the quad template that is used for tracking the inverse relationship–this other quad template belongs to the KGC that is associated with the IRI that has been used in *O*.

It bears mention how modeling by KGCs compares with modeling by ontologies. While ontologies enable modeling a domain by nature of their properties and classes, using KGCs provide additional means for organizing the interactions between (i) ontologies,

[3] The name ***skg*** refers to our ontology, where its concepts inform the nodes of a Scholarly Knowledge Graph.
[4] Each term designates its ontology via the prefixed acronym (see our Github site for a list of all ontologies we use); 'iao' is the Information Artifact Ontology, which is part of OBO; 'foaf' is the friend of a friend ontology commonly used for information relating to persons and organizations.
[5] https://www.orkg.org/

Table 1: The predominant concepts for a Scholarly Knowledge Graph (SKG), with the bold concepts representing the core concepts in the Research Contribution Model proposed in this work. ICE is *iao:information content entity*

Concept Name	Definition
MainDocument	An ICE that is an RDF document containing all information about a specific scholarly publication.
ResearchPaper	An ICE that is a scholarly publication, i.e. a document that has been accepted by a publisher (cf. *iao:publication*) and has content relevant to research.
ResearchField	An ICE that is an area of knowledge and research that refers to a specific part of reality. Research fields have no clearly defined borders between each other.
ResearchObjective	An ICE that describes an intended process endpoint for some research activity (cf. *iao:objective specification*).
ResearchResult	An ICE that is intended to be a truthful statement about something and is the output of some research activity. It is usually acquired by some research method which reliably tends to produce (approximately) truthful statements (cf. *iao:data item*).
ResearchMethod	An ICE that specifies how to conduct some research activity. It usually has some research objective as its part. It instructs some research agent how to achieve the objectives by taking the actions it specifies (cf. *iao:plan specification*).
ResearchMaterial	A material entity that functions as input or output of some research activity.
ResearchActivity	A process that has been planned and executed by some research agent and that has some research result. The process ends if some specific research objective is achieved (cf. *obi:investigation*).
ResearchAgent	A material entity that is a person, group of persons, or an organization who is directly involved in research. Research agents participate in research activities either as the study subject or as investigating agents.
Assertion	An ICE that is a proposition from some research paper and that is asserted to be true, either by the authors of the paper or by a third party referenced in the paper.

(ii) knowledge graph applications using these ontologies, and (iii) users of the application in the following ways: (1) a set of KGCs can be specified to implement a uniform modeling scheme and a particular data model such as the RCM in a knowledge graph application so that input added by users complies with the model, resulting in semantically more consistent contents without the users having to be experts in semantics; (2) ABox KGCs organize and structure the population of ontology classes with instances through user input; (3) TBox KGCs organize and structure the process of adding new ontology classes in a bottom-up approach by the users of the knowledge graph (user-driven ontology evolution)–a functionality that is required for any knowledge graph with a broader scope, since important concepts are sometimes missing in ontologies and the terminologies of research fields are evolving too; (4) KGCs restrict the expressibility of the ontologies used in a knowledge graph application to what is actually needed or relevant to be said about a given instance or concept in a particular context and for the purpose of that particular knowledge graph–ontologies often allow to record more information about a resource but this additional information may be of a type that is not relevant in the given context and for the knowledge graph and therefore the KGC does not enable adding this information to the knowledge graph.

Next, we present three ABox KGCs with their set of quad templates to show concretely the newly introduced notions. For the sake of clarity, each KGC's quad template has been simplified as follows: (i) they omit the *S* specification since in all cases it is *C*; (ii) when property *P* takes a value from a mutually exclusive list, the properties are separated by '|' logical operator; (iii) if the *O* specification pertains to the IRI of a specific type of concept or instances thereof, then its possible value space is enclosed in double angle brackets as follows «$CONCEPT_1$, ..., $CONCEPT_n$», otherwise its data type is stated. Concepts written in uppercased strings have a corresponding KGC which will be called if the quad template is instantiated, while those in lowercase do not; (iv) in some cases *inv* remains unspecified (indicated by "-") because the overall model that underlies the KGC does not require tracking the inverse relationship. The title of each KGC corresponds to the defined concept it is associated with. Note, the association of a particular KGC to a defined concept of the *skg* ontology is inherited to all of its subclasses within *skg*. The inheritance lineage is interrupted if a subclass has an explicitly assigned KGC, in which case the subclass's KGC is then inherited.

We start with the ABox KGC associated with the top-level concept *iao:information content entity*, which has 25 quad templates:

INFORMATION CONTENT ENTITY (ICE)

rdfs:label / string / SKG TERM MAPPING / [1..1] / -

rdf:type / «information content entity term» / SKG TERM MAPPING / [1..1] / -

dcterms:identifier / string / SKG TERM MAPPING / [0..m] / -

doi:hasDOI / string / SKG TERM MAPPING / [0..1] / -

sio:name / string / ir / [0..m] / -

vo:tradeName / string / ir / [0..m] / -

dcterm:creator / «PERSON,ORGANIZATION» / ir / [0..m] / sio:isCreatorOf

bfo:hasPart / «ICE,ASSERTION» / ir / [0..m] / bfo:partOf

bfo:partOf / «ICE,ASSERTION» / ir / [0..m] / bfo:hasPart

ro:bearerOf / «QUALITY, DISPOSITION» / ir / [0..m] / ro:inheresIn

edam:hasFormat / «format» / ir / [0..m] / -

skg:hasDataType / «data type» / ir / [0..m] / -

obi:isSpecifiedInputOf / «PROCESS» / ir / [0..m] / obi:hasSpecifiedInput

obi:isSpecifiedOutputOf / «PROCESS» / ir / [0..m] / obi:hasSpecifiedOutput

iao:isAbout / «MATERIAL ENTITY, DISPOSITION, QUALITY, AGENT, PROCESS, ICE» / ir / [0..m] / -

ro:hasEvidence / «ICE,ASSERTION» / ir / [0..m] / ro:isEvidenceFor

ro:isEvidenceFor / «ICE,ASSERTION, MATERIAL ENTITY, PROCESS» / ir / [0..m] / ro:hasEvidence

ro:axiomContradictedByEvidence / «ICE,ASSERTION» / ir / [0..m] / ro: evidenceContradictsAxiom

ro:evidenceContradictsAxiom / «ICE,ASSERTION» / ir / [0..m] / ro: axiomContradictedByEvidence

ero:hasDocumentation / «DOCUMENT» / ir / [0..m] / iao:isAbout

iao:denotes / «DOCUMENT» / ir / [0..m] / -

iao:hasMeasurementUnitLabel / «MEASUREMENT UNIT LABEL» / ir / [0..1] / -
iao:hasMeasurementValue / string / ir / [0..1] / -
obi:specifiesValueOf / «QUALITY» / ir / [0..1] / -
dcterms:description / string / ir / [0..m] / -

For this KGC, consider quad template: ***ro:hasEvidence*** */ «ICE, ASSERTION» / ir / [0..m] / ro:isEvidenceOf*. This quad template semantically links a newly created instance of the top-level concept *iao:information content entity* or of one of its subclasses, with instances of either the ICE or ASSERTION concept (or one of its respective sub-classes) by the **ro:hasEvidence** property. All quads created with this quad template are stored at the named graph 'ir', which is either the IRI of the newly created instance or some IRI that has been forwarded by another KGC. Since the cardinality of this quad template is [0..m], it implies that this quad template is optional for the KGC and that, if used, it can be instantiated multiple times resulting in quads with the same *Subject* and *Predicate*. Furthermore, the quad template specifies the inverse relation **ro:isEvidenceOf**. All the remaining quad templates in this KGC can be interpreted similarly and collectively offer a view of the attributes that this KGC allows to be described for any instance of *iao:information content entity* or any of its subclasses that are in the inheritance lineage.

Second, we show the ABox KGC for the predominant concept MAINDOCUMENT from the contribution schema. Its specification involves 5 quad templates as shown below:

SKG MAIN DOCUMENT
rdf:type / «fabio:digital manifestation» / SKG TERM MAPPING / [1..1] / -
dcterms:identifier | skg:hasDOI / string / SKG TERM MAPPING / [1..m] / -
dcterm:creator / «PERSON, ORGANIZATION» / ir / [1..m] / sio:isCreatorOf
skg:isEmbodimentOf / «SKG RESEARCH PAPER» / ir / [1..1] / frbr:hasEmbodiment
bfo:hasPart / «SKG TERM MAPPING» / ir / [1..1] / bfo:partOf

We see that the KGC associated with MAINDOCUMENT has only mandatory quad templates. The quad template pertaining to **dcterms:identifier | skg:hasDOI** specifies two named graphs, which means that the quad created by instanitating this quad template will be stored twice. Further, since named graphs arise from the instantiation of quad templates, we illustrate the MAINDOCUMENT-instance named graph in Fig. 1, the contents of which result from the instantiation of quad templates belonging to several KGCs that used the IRI of an instance of MAINDOCUMENT as their **NG** value, viz. RESEARCHPAPER with the paper's identifier as a DOI string, RESEARCHFIELD, and MAINDOCUMENT.

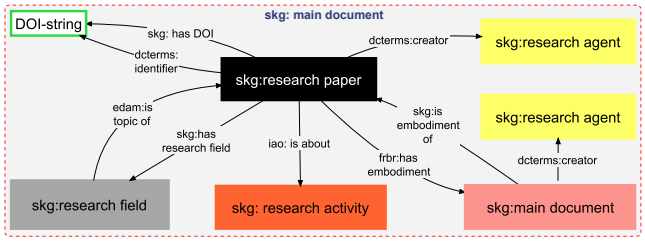

Figure 1: The resulting named graph at the MAINDOCUMENT IRI comprising instantiated parts of the KGCs of MAINDOCUMENT, RESEARCHFIELD, and RESEARCHPAPER with a DOI.

Finally, as a third example, we show the ABox KGC for RESEARCHPAPER, that is specified by the 7 quad templates shown below.

SKG RESEARCH PAPER
dcterms:title / string / SKG TERM MAPPING/ [1..1] / -
rdf:type / «iao:document» / SKG TERM MAPPING / [1..1] / -
dcterms:identifier | doi:hasDOI / string / SKG TERM MAPPING/ [0..m] / -
dcterms:creator / «PERSON, ORGANIZATION» / SKG MAIN DOCUMENT/ [1..m] / sio:isCreatorOf
frbr:hasEmbodiment / «SKG MAIN DOCUMENT» / SKG MAIN DOCUMENT / [1..1] / skg:isEmbodimentOf
skg:hasResearchField / «SKG RESEARCH FIELD» / SKG MAIN DOCUMENT / [1..m] / edam:isTopicOf
iao:isAbout / «SKG RESEARCH ACTIVITY» / SKG MAIN DOCUMENT / [1..m] / ero:hasDocumentation

This KGC has only one optional quad template, i.e. the one defined by the property **dcterms:identifier | doi:hasDOI**, since not all papers have assigned persistent identifiers. This quad template also gives a choice between two mutually exclusive properties.

Thereby, by the illustrated KGCs we demonstrate how three different aspects relevant to research contributions of scholarly publications can be semantically modeled leveraging KGCs as building blocks. The instantiation of an ABox KGC results in the generation of an instance-based semantic graph that relates instances of different defined concepts to each other and how this graph can be organized into different named graphs for storage within a tuple store framework. Further, different KGCs link to each other through the ***O*** specifications in their quad templates. When respective quad templates are instantiated, the KGC that is referenced in the respective ***O*** specification will be called and instantiated as well. For specifications of further KGCs see https://github.com/LarsVogt/ResearchContributionModel.

3.3 Research Contribution Model (RCM)

Finally, we present the **Research Contribution Model (RCM)**. We used the basic functionality of KGCs like it is described above and defined the set of KGCs associated with the remaining core concepts to model a *Research Contribution* as the unit of content of a scholarly publication. We model a *Research Contribution* as the relationship between instances of six core concepts, connected to each other through the instance of SKG RESEARCHACTIVITY. In other words, each instance of RESEARCHACTIVITY relates all instances of the remaining SKG core concepts that are at the same level in the concept schema hierarchy as itself. When users document a *Research Contribution*, they thus can instantiate the KGCs of RESEARCHMATERIAL, RESEARCHAGENT, RESEARCHMETHOD, RESEARCHOBJECTIVE, and RESEARCHRESULT for modeling the content of a scholarly publication, resulting in a named graph for each instantiated KGC. The union of these named graphs models a *Research Contribution* (Fig. 2) and the set of respective KGCs and their quad templates specifies the **RCM**.

Like a Matryoshka, the russian stacking doll, each particular *Research Contribution* can have another particular *Research Contribution* as its part (connected through parthood relations between its corresponding instances), modeling the same set of relations between SKG concepts belonging to a *Research Contribution*, but at a finer level of detail. This modularity of the RCM allows applying the same model across different levels of granularity, resulting in a knowledge graph whose contents will be better maintainable and

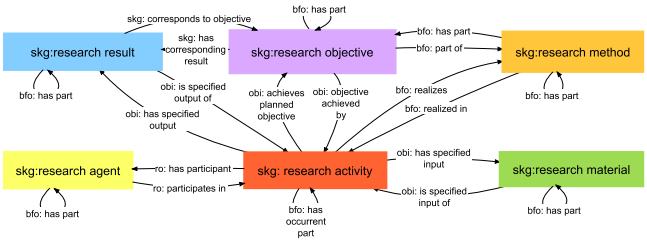

Figure 2: The SKG Research Contribution Model (RCM). It models the relation between six SKG core concepts, with ResearchActivity connecting them. Note that each core concept can have parts, which can relate to the parts of the other concepts, resulting in the same set of relations at a finer level granularity

easier understandable. Moreover, in future, users could perform faceted searches without having to write SPARQL code for them by utilizing (i) the organization of the contents into different named graphs, each of which can be identified by the IRI of the instances of these six SKG core concepts, and (ii) the pre-defined SPARQL queries provided by the KGCs. Users could, e.g., search for the combination of a specific research objective and research method, in order to find all research results documented in the knowledge graph that are associated with this combination.

4 BOTTOM-UP RCM DESIGN PROCESS

Generally, in the Life Sciences and, particularly, in Medicine, the development of ontologies and their use in practical applications is most advanced. Our RCM design is based on the experiences made there and the standards they developed.

Designing the model involved the following two steps: 1) selecting an abstract from a scholarly publication in one of the three domains we consider, i.e. Medicine, Computer Science, and Agriculture;[6] and 2) identifying sentences or paraphrases of contents of the abstract that pertain to an aspect of the contribution.

4.1 Medicine

Case 1: Self-referencing KGCs enable nested RCM and thus flexible granularity for modeling research contributions. In this section, we present the design considerations made for the KGCs in the RCM by including a 'parthood' quad template in all KGCs associated with the six core concepts. Each 'parthood' quad template *self-references* its KGC by calling it when being instantiated. This self-referencing of KGTs enables modeling contents at different levels of granularity starting from most general to more specific, depending on the scholarly contents that the user wants to represent (for levels of granularity see [35, 37]).

We show this with a sentence from the abstract of the research paper: "Reorganisation of brain networks in frontotemporal dementia and progressive supranuclear palsy" [17], in the medical domain. The sentence we model is: *"We assessed adults with behavioural variant frontotemporal dementia and progressive supranuclear palsy using magnetoencephalography during an auditory oddball paradigm."* It explains that a group of adults with two different types of dementia were subjected to an auditory oddball paradigm experiment during which they got a magnetoencephalography treatment that recorded their reaction to the oddball stimulus. As such they were study subjects.

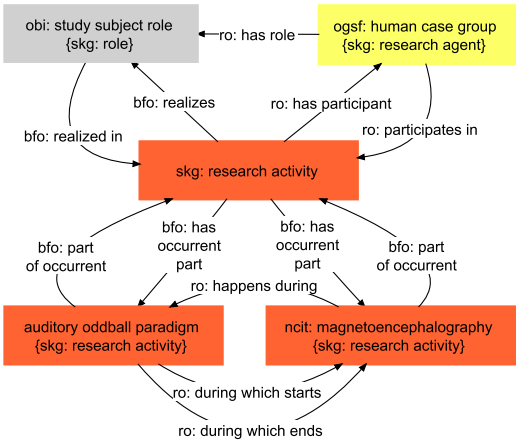

Figure 3: Illustration of nesting of the Research Contribution Model with three instances of ResearchActivity being related to each other using 'bfo: has occurrent part' property, resulting in a two-level granularity

The nested model for this sentence is depicted in Figure 3 for three instances of ResearchActivity being connected by **bfo: part of occurrent** properties (see red boxes). In the figure, another context from the sentence is modeled by an instance each of the Role and ResearchAgent concept (see gray and yellow boxes), respectively, that reflect the information in the sentence about humans as study subjects. For focus on the design case at hand, their details are not shown. Both 'auditory oddball paradigm' and 'magnetoencephalography' are a ResearchActivity. Our semantic model therefore had to take into account that a research activity may have other research activities as parts, resulting in a nested modeling of ResearchActivity. This is done via its **bfo: has occurrent part** property. With this nested application feature enabled, the RCM is applied three times (Fig. 3), for modeling: 1) auditory oddball paradigm; 2) magnetoencephalography; and 3) a research activity that has 1) and 2) as its parts. This results in a graph that represents ResearchActivity at two levels of granularity. The other core concepts in the RCM adopt the same nested model, using the property **bfo:has part** (refer RCM core concepts in Fig. 2).

Case 2: Exemplifying the overall expressivity of the RCM. In this section, we present the design considerations made for KGCs to allow users to model propositions from a scholarly publication either in a semantically formal and detailed way or by using a mixture of natural language description and specifying relations to ontology classes. We further point at the possibility to use KGCs not only for modeling assertional statements (i.e., ABox expressions) but also universal statements such as hypotheses and class

[6]We randomly select the abstracts from the Elsevier Labs OA-STM corpus https://github.com/elsevierlabs/OA-STM-Corpus.

definitions, by allowing them to be defined within the semantic construct of a newly defined class using TBox KGCs.

We show this with a sentence taken from the abstract of a different research paper: "Vascular risk status as a predictor of later-life depressive symptoms: a cohort study" [22], in the medical domain. "*We examined whether standard clinical risk profiles developed for vascular diseases also predict depressive symptoms in older adults.*" The selected sentence describes the research objective of the investigation as attempting to test the following hypothesis: "Standard clinical risk profiles developed for vascular diseases predict depressive symptoms in older adults."

Since hypotheses represent universal statements (i.e., they claim to be universally true for all instances of the referenced types), we cannot model the hypothesis as an instance-based semantic graph but must instead use a class-based graph which we present using Manchester Syntax below.[7]

```
'obi: is specified output of' some
    ('process ncit: profile'
     and ('ro: has participant' some
        ('material entity foaf: person'
         and ('ro: has disposition' some 'disposition efo: vascular disease')))))
 and ('ro: correlated with' some
    ('ogms: symptom'
     and ('iao: is about' some
        ('disposition ncit: depression'
         and ('ro: inheres in' only
            ('material entity ncit: adult'
             and ('ro: participates in' some 'process uberon: late adult stage')))
         and ('ro: realized in' only 'process uberon: late adult stage')))).
```

In cases like this, where propositions exhibit a complex relational structure that requires some effort to describe them semantically in detail, the RCM allows for modeling them in a mixture of unstructured text and semantic relations. The following Manchester Syntax expression models the same sentence, but in considerably lesser detail [8]. Instead of the complex relational structure, in this alternative, terms from the sentence that seem to be relevant to the hypothesis are added via the property '*iao:is about*.'[9]

```
'iao:is about' some
    ('efo:vascular disease'
     and 'ncit:depression'
     and 'ncit:adult'
     and 'skg:depressive symptom')
```

Coming back to our selected sentence: The authors conclude that the results of their study support the hypothesis mentioned above. However, how to model ResearchResults has not yet been discussed. This can be done as shown in Figure 4. Regarding the hypothesis, it could be described as an ontology class using either of the two alternative Manchester Syntax expressions, and unstructured text could be linked to an instance of that class via the property *dcterms: description*.

Fig. 4 also depicts a named graph as a gray box inside dotted lines. This named graph was triggered by the instantiation of the 'has part' quad template of the KGT that is associated with ResearchResult. Instantiating this quad template calls for two different KGCs, one of which is the Assertion KGC. All quad templates of the Assertion KGC forward their IRI as the ir-value for all KGCs they subsequently call, resulting in all quads being created through the called Assertion KGCs to be stored in the same named graph. In Fig. 4 this named

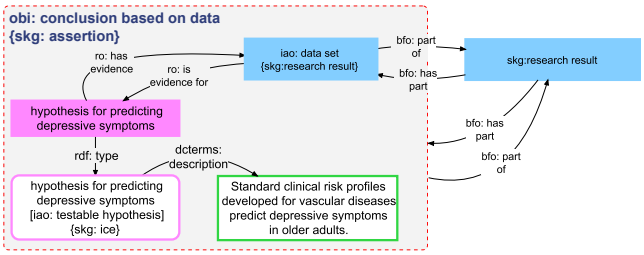

Figure 4: Illustration of using unstructured text and a class for modeling a universal statement such as a hypothesis and of how to model a research conclusion within a named graph

graph is an instance of 'obi: conclusion based on data'. This way we can describe the conclusion of the result within a single named graph and link it to the result using the property '*bfo:part of*.'

4.2 Computer Science

Case 3: Modeling research results in the RCM. Here we present design considerations made in the modeling of research results in the RCM by showing two representative instances, viz. a result in the form of a performance score, and a result that is not a score.

Our first example is selected from the research paper: "Coherent clusters in source code" [18]. For the paper's result, we select the following sentence: *We introduce an approximation to efficiently locate coherent clusters and show that it has a minimum precision of 97.76%*, and, in addition, rely on surrounding sentences for context. Basically, the authors discuss the performance of their software, which they call "approximation software," to locate patterns in code in the form of coherent clusters–where "locating patterns in code" is a task in software engineering–, and they find that their method for the task obtains a precision score of 97.76%.

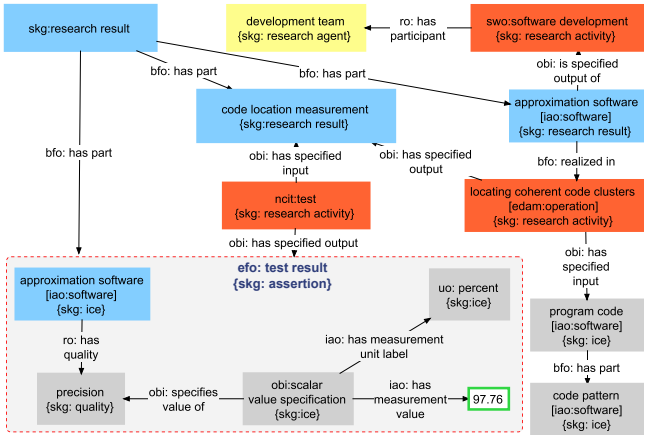

Figure 5: Illustration of the modeling of a numeric result (specifically, a precision measurement) as a part of the ResearchResult by storing the measurement graph in its own Assertion named graph

[7] https://www.w3.org/TR/owl2-manchester-syntax/
[8] this is done using respective TBox KGCs–due to page restrictions and clarity of the paper's narrative, we do not further introduce and discuss TBox KGCs.
[9] This Manchester Syntax expression can be translated into a class-based semantic graph representation–see our Github site for an additional example.

Our model for the result is depicted in Fig. 5. Starting at the result value itself (i.e. 97.76), it is represented as a float value (box with green border) and is linked by the property **iao: has measurement value** to an instance of *obi:scalar value specification*, to which also an instance of **uo: percent** as the measurement unit is linked. Further down the network, the value is qualified as a measurement for an instance of the quality type 'precision' for the approximation software. This part of the graph has been created using the ICE and the QUALITY KGC. Finally, this graph is stored in its own named graph and represents the test result. It has been created as a result of the instantiation of the KGC **Assertion**, which forwards the named graph IRI to all subsequently called KGCs. The contents of this named graph are part of the research result. Essentially, the just discussed semantic graph reads in natural language as: *Part of the research result is the measurement of the precision of the approximation software, which is 97.76 percent.*

Next, we depict the modeling of a result that is not numeric. It is from the research paper: "Using SPIN for automated debugging of infinite executions of Java." [4] Since in this paper's abstract, the result is not expressed in a single sentence, it is paraphrased from several sentences across the abstract as: *a new software approach for automatic debugging that combines model checking and runtime monitoring.*

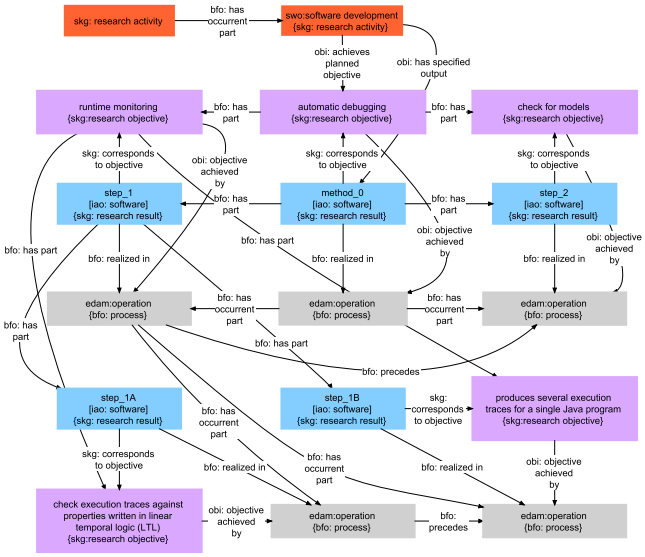

Figure 6: Illustration of the modeling of a sequential order of steps in a software as the result of a scientific investigation (for reasons of clarity of representation, only one direction of relations is shown)

We refer the reader to Fig. 6. For the result, the model primarily relies on three concepts, viz. RESEARCHRESULT of type *iao:software*, RESEARCHOBJECTIVE, and *bfo:process*. A summary of the figure is that the software was developed to achieve an overarching objective (i.e. "automatic debugging" as RESEARCHOBJECTIVE) which involved sub-objectives modeled as two main steps (see instantiated boxes with labels "step1" and "step2", both of type *iao:software*) where the first step itself was split further. Every step is associated with an instance of RESEARCHOBJECTIVE (see the steps in blue boxes linked by the **skg: corresponds to objective** property with the purple boxes). And finally that every objective culminates in a process, i.e. the execution of the method that is coded in the software (see purple boxes for RESEARCHOBJECTIVE linked by property **obi:objective achieved by** to *bfo:process* in the gray boxes).

At a high level of granularity, the software (i.e. blue box with value "method0" as instance of *iao:software*) is modeled as a RESEARCHRESULT associated with a RESEARCHACTIVITY that is qualified as type *swo: software development*. At finer granularity of specification, the software result is elaborated with parts of the software having dedicated functions with each part of the result/software having its own RESEARCHOBJECTIVE specified.

This model presents a case for modeling algorithms or any other plan specification in the RCM. Since, equivalently, we understand a software to be a specific type of plan specification which has an objective and that is realized in an operation process. Essentially, for all plan specifications, we want to model a sequential ordering of a process with its parts, which can be done in the RCM as demonstrated.

With the four examples presented above, we have illustrated five design considerations that were made in developing the RCM and its specification via KGCs: 1) nested modeling to represent data at different levels of granularity; 2) defining new classes to represent TBox expressions for modeling e.g. hypotheses; 3) semi-structured modeling with natural language expressions; 4) modeling of results, both numeric and non-numeric; and 5) modeling of sequential order for plan specifications.

4.3 Agriculture

Case 4: Whether the RCM can model an aspect of the *contribution* that intermingles various concepts. In this domain, we select examples which require a more comprehensive reach into the RCM entailing the utilization of an interaction between various concepts to model the *contribution* of an investigation. For the purpose, we use the research paper: "Soil structural responses to alterations in soil microbiota induced by the dilution method and mycorrhizal fungal inoculation" [26], a publication in the Agriculture domain. From its abstract, we select the sentence: *After seven months, principal components analysis (PCA) separated bacterial community composition primarily on planting regime.* Further, from the abstract, we consider more information about "principal components analysis (PCA)" in terms of the following paraphrase: *before PCA some material processing of sterile soil was conducted, followed by different planting regimes.* Next, since we aim to consider a model with several concepts, we select a second sentence: *A consistent finding in planted and unplanted soils was that aggregate stability was positively correlated with small pore sizes.* This sentence informs us that there is a research result which is based on an assay, i.e. a test procedure.

We present the model in Figure 7. First of all, the model would involve the repeated instantiation of concepts. Consider in the figure, the instances of RESEARCHACTIVITY (see the red boxes connected by **bfo:has occurrent part**) and of RESEARCHRESULT (see the blue boxes connected by **bfo:has part**). For the first example, it entails modeling a sequence of activities for depicting "material processing

that took 7 months followed by PCA." Consider this depicted by the two instances of ResearchActivity, i.e. *obi:material processing* connected by **bfo:precedes** with *ncit:principal component analysis*. In the earlier example we depicted the modeling of a numeric value as a ResearchResult. In this example, however, we model the timeframe as a numeric value for the duration of a ResearchActivity. Thus the RCM design flexibly allows the modeling of such contextual relations. Finally, in this example, we use instances of two new concepts by instantiating their corresponding KGCs, viz. ResearchMaterial (see the green box) and ResearchMethod (see the yellow box). With this small-world model, we demonstrate the feasibility of the RCM to model different information units via various KGCs.

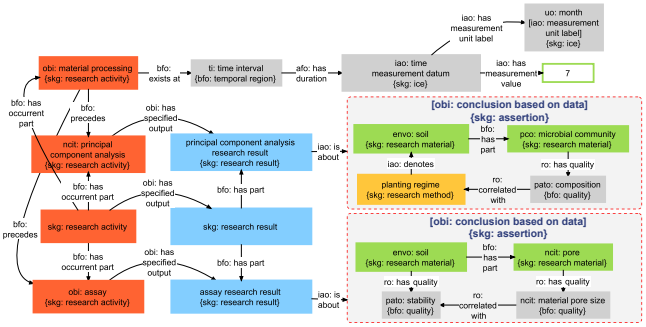

Figure 7: Illustration of the RCM for the small-world *contribution* example in Agriculture that models various kinds of information (ResearchActivity, ResearchResult, ResearchMaterial, ResearchMethod)

5 RELATED WORK

We review the relevant prior work in the context of semantic scientific publishing.

Semantic Authoring. All major semantic authoring systems (e.g. the semantic LaTeX extensions sTeX [23], SALT [16]) have so far neglected the specific use case of research contributions. This can be partially explained by the fact that these frameworks have had different development foci – mathematics for sTeX, rhetorical structures for SALT. Based upon our review of existing work, a semantic authoring support for research contributions as we propose does not formally exist at present.

Creating Knowledge Graphs from Scholarly Publications. Generally, Knowledge Graphs (KG) of unstructured text are generated at scale over a large mass of scholarly literature by the identification and extraction of concepts and their relations guided by ontological definitions and structure using Natural Language Processing techniques. The iASiS knowledge graph [34] is generated over biomedical scholarly publications on Lung Cancer and Dementia coupled with the standarized biomedical ontologies such as UMLS [9]. Further, Research Spotlight (RS) [29] is an interdisciplinary system that leverages the Scholarly Ontology (SO) [28] and deep syntactic analysis to extract information from scholarly articles and publish information as linked data. Theirs is a research activity centered model.

Thus, the themes of knowledge graphs vary between domain-specific subjects such as diseases in biomedicine, or domain-independent subjects such as research activity. While scientific articles are stored in silos isolated from each other, knowledge graphs demonstrate how this is overcome by semantically combining different units of information. Whereas manual data entry based upon the Knowledge Graph Cells (KGC)s presented in this paper generates knowledge graphs in the backend, these KGCs can be separately leveraged together with a set of ontologies coupled with NLP to generate its research-contribution-based knowledge graphs at scale.

Using Templatized Concepts. Various ways of organizing RDF statements in a knowledge graph have been proposed for the purpose of efficiently tracking provenance data and enforcing data consistency. The concept of *RDF molecules*, initially proposed by [11] and further elaborated by [12, 27], is conceptually related to our concept of KGCs, as the set of quad templates of a KGC is similar to the specification of an RDF molecule. Our concept of KGCs differs from RDF molecules in that we combine the molecule idea with the idea of using named graphs for organizing triples in a tuple store. Moreover, we extend its functionality: we do not use KGCs for decomposing a given graph, but instead use them as building blocks for organizing incoming data as well as for structuring and organizing input forms for the interface, for restricting the expressivity of RDF/OWL and for establishing a uniform data model.

Reasonable Ontology Templates (OTTR)[10] is a language and generic macro mechanism for specifying and instantiating RDF graph and OWL ontology modeling patterns [30, 31]. While it does not represent a pattern itself, OTTR is designed to support interaction with OWL or RDF knowledge graphs at the level of modeling patterns. As such, it may be used, at least partly, for specifying a KGC as a parameterized RDF graph. Unfortunately, however, OTTR currently does not support quads and thus named graphs. On the other hand, the Shapes Constraint Language (SHACL)[11] is a language that provides a grammar for describing KGCs.

Comparisons of Research Contributions. The current state of available resources for comparing research contributions are manually curated portals [1–3]. The SemSur ontology [13] has a closely related objective to ours–that of modeling research contributions in a semantic and machine-interpretable format to make them more transparent and comparable. Its focus, however, lies on modeling contents from the Computer Science domain, while with our RCM we provide a more general and domain-independent model that can be considerably adapted to the needs and requirements of particular research fields by introducing domain-specific KGCs. Moreover, RCM's nested structure allows the modeling of information on various levels of granularity.

6 CONCLUSION

A continuously increasing number of research communities start to agree that we need to build the Internet of FAIR Data and Services (IFDS) [8] that scales with Big Data, provides rich machine-actionable metadata with human-readable interface outputs and search capabilities, and assigns all relevant digital objects a Unique Persistent and Resolvable Identifier (UPRI). We designed the RCM

[10] https://ottr.xyz/
[11] https://www.w3.org/TR/shacl/

to be FAIR-compliant in a very strict interpretation of the FAIR principles. This is achieved by virtue of the translation of the natural language based recordings of research contributions into formalized semantic knowledge graphs. The RCM uses a suite of KGCs as building blocks that, when applied to modeling contents of scholarly publications, makes these contents findable, accessible, interoperable, and reusable for humans and machines alike. This entails, on the one hand, mapping human-readable labels and definitions to their respective UPRIs and making the UPRIs findable through these labels and translating the semantic graph into a human-readable form as HTML pages. On the other hand, it requires reusing ontology terms of well established ontologies for documenting the contents of a publication wherever possible. With the RCM, we provide a data model and modeling pattern that is general enough that it can be applied for modeling various contents using the same set of templates. If implemented in an SKG application such as the ORKG using the KGCs, the RCM would ensure content FAIR-compliance. With its potential to provide an application framework with a native graph data structure that is well integrated with RDF/OWL and that allows handling graph data and manipulating and displaying SPARQL results–a framework that is still lacking[15]–the concept of KGCs represents a promising idea. Its actual implementation will require the development of a suitable middleware and frontend. We fully support the adoption of KGCs by other interested researchers to demonstrate its practicality.

REFERENCES

[1] [n.d.]. NLP-progress. https://nlpprogress.com/. Accessed: 2020-01-14.
[2] [n.d.]. OpenAI. https://openai.com/. Accessed: 2020-01-14.
[3] [n.d.]. Papers with Code. https://paperswithcode.com/. Accessed: 2020-01-14.
[4] Damián Adalid, Alberto Salmerón, María del Mar Gallardo, and Pedro Merino. 2014. Using SPIN for automated debugging of infinite executions of Java programs. *Journal of Systems and Software* 90 (2014), 61–75.
[5] Robert Arp, Barry Smith, and Andrew D Spear. 2015. *Building ontologies with basic formal ontology.* Mit Press.
[6] Sören Auer. 2018. Towards an Open Research Knowledge Graph. https://doi.org/10.5281/zenodo.1157185
[7] Sören Auer and Sanjeet Mann. 2019. Towards an Open Research Knowledge Graph. *The Serials Librarian* 76, 1-4 (2019), 35–41. https://doi.org/10.1080/0361526X.2019.1540272 arXiv:https://doi.org/10.1080/0361526X.2019.1540272
[8] Paul Ayris, Jean-Yves Berthou, Rachel Bruce, Stefanie Lindstaedt, Anna Monreale, Barend Mons, Yasuhiro Murayama, Caj Södergård, Klaus Tochtermann, and Ross Wilkinson. 2016. *Realising the European open science cloud.* European Union, Luxembourg. https://doi.org/10.2777/940154
[9] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.
[10] A Brazma. 2001. On the importance of standardisation in life sciences. *Bioinformatics (Oxford, England)* 17, 2 (2001), 113.
[11] Li Ding, Yun Peng, Paulo Pinheiro da Silva, Deborah L McGuinness, et al. 2005. Tracking rdf graph provenance using rdf molecules. *TR-CS-05-06* (2005).
[12] Kemele M Endris, Mikhail Galkin, Ioanna Lytra, Mohamed Nadjib Mami, Maria-Esther Vidal, and Sören Auer. 2017. MULDER: querying the linked data web by bridging RDF molecule templates. In *International Conference on Database and Expert Systems Applications*. Springer, 3–18.
[13] 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 (2018), 151–162.
[14] Johannes Frey, Kay Müller, Sebastian Hellmann, Erhard Rahm, and Maria-Esther Vidal. 2019. Evaluation of Metadata Representations in RDF stores. *Semantic Web* Preprint (2019), 1–25.
[15] Fabien Gandon, Franck Michel, Olivier Corby, Michel Buffa, Andrea Tettamanzi, Catherine Faron Zucker, Elena Cabrio, and Serena Villata. 2019. Graph Data on the Web: extend the pivot, don't reinvent the wheel. *arXiv preprint arXiv:1903.04181* (2019).
[16] Tudor Groza, Siegfried Handschuh, Knud Möller, and Stefan Decker. 2007. SALT-Semantically Annotated LaTeXfor Scientific Publications. In *European Semantic Web Conference*. Springer, 518–532.
[17] Laura E Hughes, Boyd CP Ghosh, and James B Rowe. 2013. Reorganisation of brain networks in frontotemporal dementia and progressive supranuclear palsy. *NeuroImage: Clinical* 2 (2013), 459–468.
[18] Syed Islam, Jens Krinke, David Binkley, and Mark Harman. 2014. Coherent clusters in source code. *Journal of Systems and Software* 88 (2014), 1–24.
[19] Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T Evelo, et al. 2019. FAIR Principles: Interpretations and Implementation Considerations.
[20] Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge *(K-CAP '19)*. Association for Computing Machinery, New York, NY, USA, 243–246. https://doi.org/10.1145/3360901.3364435
[21] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing* 23, 3 (2010), 258–263.
[22] Mika Kivimäki, Martin J Shipley, Charlotte L Allan, Claire E Sexton, Markus Jokela, Marianna Virtanen, Henning Tiemeier, Klaus P Ebmeier, and Archana Singh-Manoux. 2012. Vascular risk status as a predictor of later-life depressive symptoms: a cohort study. *Biological psychiatry* 72, 4 (2012), 324–330.
[23] Michael Kohlhase. 2008. Using as a semantic markup format. *Mathematics in Computer Science* 2, 2 (2008), 279–304.
[24] Esther Landhuis. 2016. Scientific literature: information overload. *Nature* 535, 7612 (2016), 457–458.
[25] Yair Levy and Timothy J Ellis. 2006. A systems approach to conduct an effective literature review in support of information systems research. *Informing Science* 9 (2006).
[26] Sarah L Martin, Sacha J Mooney, Matthew J Dickinson, and Helen M West. 2012. Soil structural responses to alterations in soil microbiota induced by the dilution method and mycorrhizal fungal inoculation. *Pedobiologia* 55, 5 (2012), 271–281.
[27] Andrew Newman, Jane Hunter, Yuan-Fang Li, Chris Bouton, and Melissa Davis. 2008. A scale-out RDF molecule store for distributed processing of biomedical data. (2008).
[28] Vayianos Pertsas and Panos Constantopoulos. 2017. Scholarly Ontology: modelling scholarly practices. *International Journal on Digital Libraries* 18, 3 (2017), 173–190.
[29] Vayianos Pertsas and Panos Constantopoulos. 2018. Ontology-driven information extraction from research publications. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 241–253.
[30] Martin G Skjæveland, Henrik Forssell, Johan W Klüwer, Daniel P Lupp, Evgenij Thorstensen, and Arild Waaler. 2017. Pattern-Based Ontology Design and Instantiation with Reasonable Ontology Templates.. In *WOP@ ISWC*.
[31] Martin G Skjæveland, Daniel P Lupp, Leif Harald Karlsen, and Henrik Forssell. 2018. Practical ontology pattern instantiation, discovery, and maintenance with reasonable ontology templates. In *International Semantic Web Conference*. Springer, 477–494.
[32] Barry Smith, Mauricio Almeida, Jonathan Bona, Mathias Brochhausen, Werner Ceusters, Melanie Courtot, Randall Dipert, Albert Goldfain, Pierre Grenon, Janna Hastings, William Hogan, Leonard Jacuzzo, Ingvar Johansson, Chris Mungall, Darren Natale, Fabian Neuhaus, Anthony Petosa, Robert Rovetto, Alan Ruttenberg, Mark Ressler, and Stefan Schulz. 2015. Basic Formal Ontology 2.0. (2015). https://github.com/BFO-ontology/BFO/blob/master/docs/bfo2-reference/BFO2-Reference.pdf
[33] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* 25, 11 (2007), 1251.
[34] Maria-Esther Vidal, Kemele M. Endris, Samaneh Jazashoori, Ahmad Sakor, and Ariam Rivas. 2019. Transforming Heterogeneous Data into Knowledge for Personalized Treatments—A Use Case. *Datenbank-Spektrum* 19 (2019), 95–106.
[35] Lars Vogt. 2010. Spatio-structural granularity of biological material entities. *BMC bioinformatics* 11, 1 (2010), 289.
[36] Lars Vogt. 2013. eScience and the need for data standards in the life sciences: in pursuit of objectivity rather than truth. *Systematics and biodiversity* 11, 3 (2013), 257–270.
[37] Lars Vogt. 2019. Levels and building blocks—toward a domain granularity framework for the life sciences. *Journal of biomedical semantics* 10, 1 (2019), 4.
[38] Lars Vogt, Roman Baum, Philipp Bhatty, Christian Köhler, Sandra Meid, Björn Quast, and Peter Grobe. 2019. SOCCOMAS: a FAIR web content management system that uses knowledge graphs and that is based on semantic programming. *Database* 2019 (2019).
[39] Xiaoshu Wang, Robert Gorlitsky, and Jonas S Almeida. 2005. From XML to RDF: how semantic web technologies will change the design of'omic'standards. *Nature biotechnology* 23, 9 (2005), 1099.
[40] 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 (2016).