Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles

Said Fathalla[1,3(✉)], Sahar Vahdati[1], Sören Auer[4,5], and Christoph Lange[1,2]

[1] Smart Data Analytics (SDA), University of Bonn, Bonn, Germany
{fathalla,vahdati,langec}@cs.uni-bonn.de
[2] Fraunhofer IAIS, Sankt Augustin, Germany
[3] Faculty of Science, Alexandria University, Alexandria, Egypt
[4] Computer Science, Leibniz University of Hannover, Hanover, Germany
[5] TIB Leibniz Information Center for Science and Technology, Hannover, Germany
soeren.auer@tib.eu

Abstract. Despite significant advances in technology, the way how research is done and especially communicated has not changed much. We have the vision that ultimately researchers will work on a common knowledge base comprising comprehensive descriptions of their research, thus making research contributions transparent and comparable. The current approach for structuring, systematizing and comparing research results is via survey or review articles. In this article, we describe how surveys for research fields can be represented in a semantic way, resulting in a knowledge graph that describes the individual research problems, approaches, implementations and evaluations in a structured and comparable way. We present a comprehensive ontology for capturing the content of survey articles. We discuss possible applications and present an evaluation of our approach with the retrospective, exemplary semantification of a survey. We demonstrate the utility of the resulting knowledge graph by using it to answer queries about the different research contributions covered by the survey and evaluate how well the query answers serve readers' information needs, in comparison to having them extract the same information from reading a survey paper.

Keywords: Semantic metadata enrichment · Quality assessment · Recommendation services · Scholarly communication · Semantic publishing

1 Introduction

Despite significant advances in technology in the last decades, the way how research is done and especially communicated has not changed much. Researchers still encode their findings in sequential text accompanied by illustrations and wrap these into articles, which are mostly published in printed form or as semi-structured PDF documents online. We have the vision that ultimately researchers will rather work on a common knowledge base comprising

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 315–327, 2017.
DOI: 10.1007/978-3-319-67008-9_25

comprehensive descriptions of their research, thus making research contributions transparent and directly comparable. The current approach for structuring, systematizing and comparing research results is via survey or review articles. Such articles usually select a number of articles describing comparable research and (a) develop a common organization scheme with feature classifications, (b) provide a conceptualization of the research domain with mappings to the terminologies used in the individual articles, (c) compare and possibly benchmark the research approaches, implementations and evaluations described in the articles and (d) identify directions for future research. As a result, survey and review articles significantly contribute to structuring a research domain and make its progress more transparent and accessible. However, such articles still share the same deficiencies as their original research counterparts – the content is not represented according to a formal knowledge representation and not machine comprehensible, which prevents systematic identification of conceptualization problems as well as the building of intelligent search, exploration and browsing applications on top.

In this article, we describe how surveys for research fields can be represented in a semantic way resulting in a knowledge graph that describes the individual research problems, approaches, implementations and evaluations in a structured and comparable way. We present a comprehensive ontology for capturing the content of survey articles. The ontology is structured around four core concepts:

- *research problem* – describing a challenge in a particular field, possibly hierarchically decomposed into sub-problems,
- *approach* – describing attributes and features of particular research approaches,
- *implementation* – describing the implementation of an approach in a concrete technical environment,
- *evaluation* – describing the benchmarking of an implementation in a certain formally defined evaluation scenario.

As a result of structuring and representing research advances according to such a semantic scheme, they will become more comparable and accessible. For example, research addressing a certain problem can be automatically retrieved, approaches can be compared according to their features or w.r.t. evaluation results in a certain defined setting. In particular, we discuss possible applications and present an evaluation of our approach with the retrospective, exemplary semantification of a survey resulting in a knowledge graph comprising a comprehensive description of the respective research.

The ultimate aim of this work is to enable the provision of better and more intelligent services for the discovery of scientific work.

We illustrate our methodology with the example of the following three survey articles:

- Bringing Relational Databases into the Semantic Web: A Survey [12].
- A Survey of Current Link Discovery Frameworks [6].
- Querying over Federated SPARQL Endpoints —A State of the Art Survey [9].

The remainder of the article is structured as follows: We present an overview on related work in Sect. 2. The conceptualization of a knowledge graph of research advances is described in Sect. 3. We present a methodology for semantifying survey articles in Sect. 4. An evaluation describing typical usage scenarios and queries for exploring the knowledge graph in these scenarios is presented in Sect. 5. We conclude with an outlook on future work in Sect. 6.

2 Related Work

In the last decade, there has been a mass growth in scholarly communications due to the impact of the ubiquitous availability of the Internet, email, and web-based services on scholarly communication. The preparation of manuscripts as well as the organization of conferences, from submission to peer review to publication, have become considerably easier and efficient. Research is based on digital assets, such as datasets, services, and produces its output in digital form.

Capadisli's "linked research" approach starts with HTML and enriches it semantically, encapsulating publication meta-data and content [3]. Researchers are encouraged and enabled to announce their research so that they will be both authors and publishers. Research Articles in Simplified HTML (or RASH) is another Web-first format for writing HTML-based scholarly papers [8]. RASH enables a formal representation of the structure an article, which is linked to semantically related articles, thus supporting integration of data between papers. Both approaches scale up to a semantic representation of the full details of a research investigation, but hardly any author has made this effort manually.

Several efforts on developing ontologies as well as publishing reusable, machine-comprehensible (meta)data (i.e. *linked open data*) related to scholarly entities such as publications, scientific events, authors etc., have been carried out so far [10]. For example, the Springer LOD dataset[1] contains metadata about conference proceedings from the Lecture Notes in Computer Science series and aims at answering trust-related questions of different stakeholders. Bryl et al. mention questions such as "Should I submit a paper to this conference?", and point out that the data that is required for answering such questions is not easily available but, e.g., hidden in conference management systems [2]. The Semantic Web Dog Food (SWDF) dataset[2] and its successor *ScholarlyData*[3] are among the pioneers of datasets of comprehensive scholarly communication metadata. All these works support scholarly communication by giving end users easy access just to metadata about research-related entities, not to the research findings. However, none of them provides services to ease the process of gaining an overview of a field, which is what we introduce in this work.

[1] http://lod.springer.com/.
[2] http://data.semanticweb.org/.
[3] http://www.scholarlydata.org/dumps/.

3 Conceptualization

In different research disciplines there is, depending on the culture of that domain, a need for studying the literature on a specific topic to write a survey or review article, which facilitates comprehension of the topic. Experts in the field commonly create such reviews ready for the community. The readers of such review articles are often peer-researchers in the field, in particular also young researchers aiming to get an overview. Due to the representation of review articles as unstructured text, it is impossible to automatically extract and analyze information from them. In the remainder of this section, we introduce the concepts, terms and vocabularies that we defined for representing the content of review articles.

3.1 SemSur Ontology

SemSur, the Semantic Survey Ontology, is a core ontology for describing individual research problems, approaches, implementations and evaluations in a structured, comparable way. We describe its structure and contents, which captures detailed terminological knowledge about survey articles, e.g., evaluation method, hypothesis, benchmark, and experiment. SemSur is represented in the Web Ontology Language (OWL) and developed using Protégé 5.2.0 [5]. We defined new vocabularies in the OpenResearch namespace[4]. Table 1 shows the ontology statistics.

Table 1. Overview of SemSur ontology statistics.

Metrics	Count	Metrics	Count
Classes	197	Object properties	149
Data properties	78	Instances	220
Subclass relationships	234	Transitive properties	2
Inverse properties	14	Symmetric properties	2

3.2 Reuse of Ontological Knowledge

Technologies for efficient and effective reuse of ontological knowledge are one of the key success factors for developing ontology-based systems [11]. Therefore, the first step in building our knowledge graph is reusing vocabularies from related existing ontologies on the Web, since reuse increases the value of semantic data. We have selected the most closely related ontologies listed in the Linked Open Vocabularies (LOV) directory[5]. Existing related vocabularies are shown in Table 2.

[4] http://openresearch.org; prefix (`or`).
[5] Linked Open Vocabularies: http://lov.okfn.org/dataset/lov/vocabs.

Table 2. Prefixes and namespace URIs of reused vocabularies.

Prefix	Vocabulary	URI
dcterms	Dublin Core Metadata Initiative (DCMI)	http://purl.org/dc/terms/
swrc	Semantic Web for Research Communities	http://swrc.ontoware.org/ontology#
foaf	Friend of a Friend ontology	http://xmlns.com/foaf/0.1/
mls	Machine Learning Schema	https://www.w3.org/ns/mls#
deo	The Discourse Elements Ontology	http://purl.org/spar/deo/
lsc	Linked Science Core Vocabulary	http://linkedscience.org/lsc/ns#
doap	Description of a Project	http://usefulinc.com/ns/doap#

For modeling the top level metadata of a scientific article as a whole, we reuse the DC, SWRC and FOAF ontologies. The Dublin Core Metadata Initiative (DCMI)[6] provides a standard vocabulary for describing resources. *SWRC* (Semantic Web for Research Communities) describes research communities and relevant related concepts such as persons, organizations, bibliographic metadata and relationships between them. The FOAF ontology describes persons and their activities. For modeling the inner structure of a scientific article independently of the field of research we use DEO (Discourse Elements Ontology) and LSC (Linked Science Core). *DEO* is an ontology for describing the major elements within journal articles such as Abstract, Introduction, Reference List and Figures. *LSC* is designed for describing scientific resources including Publication, Researcher, Method, Hypothesis, and Conclusion. To model concepts of specific fields of research we use MLS and DOAP and may in future use additional ontologies. MLS is a standard schema published by the W3C Machine Learning Schema community group for machine learning algorithms, data mining, datasets, and experiments. DOAP (Description of a Project) is a vocabulary that describes software projects and related concepts.

Figure 1 gives an overview of a SemSur knowledge graph describing individual research problems, approaches, implementations and evaluations. For better readability of the visualization some classes are omitted. Namespace prefixes are used according to prefix.cc[7].

3.3 SemSur Classes

The SemSur ontology imports classes from the ontologies introduced in Sect. 3.2 in addition to its owns classes. Some of these classes need more specialization so we created respective subclasses. For instance, we added three subclasses `Mathematical Model`, `ArchitecturalModel` and `PipelineModel` for the `Model` class inherited from the MLS ontology. Another concern is the integration of imported ontologies. In other words, classes imported from an ontology

[6] Dublin Core Metadata Initiative: http://dublincore.org/.
[7] Namespace look-up tool for RDF developers: http://prefix.cc/.

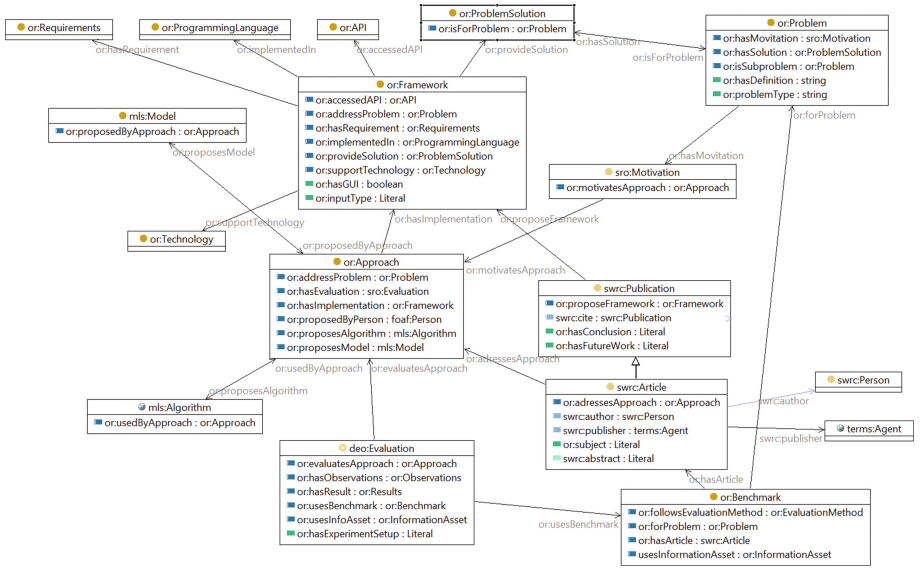

Fig. 1. Overview of a SemSur knowledge graph

should have a suitable relation with related classes found in other ontologies. For instance, the `Article` class (from SWRC) should have a relation with the `Conclusion` class (from LSC) with the relation *produces* (from LSC). Some of the reused classes are shown in Table 3.

3.4 SemSur Relations

SemSur provides a taxonomic class hierarchy. For instance, `Article` is a subclass of `Publication` as shown in Fig. 1. There are some transitive relations such `usesFramework` and `isSubproblem`. For instance, if X `isSubproblem` of Y and Y `isSubproblem` of Z then it could be inferred that X `isSubproblem` of Z. Also, there are some symmetric relations such as `hasRelatedProblem` and some

Table 3. Classes and relations reused by SemSur.

Ontology	Reused classes	Reused relations
SWRC	Article	year
foaf	Person	name
mls	Model, Algorithm, Information Entity	hasInput, hasOutput
lsc	Conclusion	Produces, timeAccepted
deo	FutureWork, Evaluation, Motivation	

inverse relations such as `proposesModel` and `proposedByApproach`. In addition, we borrow some relations from different ontologies as shown in Table 3.

3.5 SemSur Instances

Creating instances of classes is the last step of common knowledge engineering methodologies [1]. The required steps for creating a knowledge graph are: (1) identify the classes, (2) create instances of these classes, and (3) add values for the associated properties [7]. For example, creating the instance *ANAPSID-framework*, which is a specific adaptive query processing engine for SPARQL endpoints, requires (1) identify the `Framework` class, (2) create the instance, and (3) add values of properties such as `hasGUI`, `platform` and `implementedIn`. The complete instance is shown in Fig. 2.

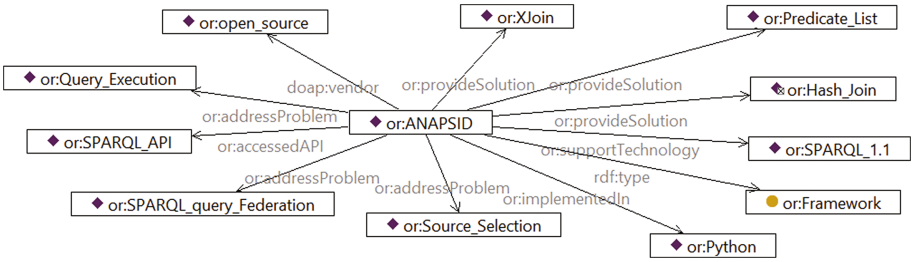

Fig. 2. SemSur ontology instance describing the ANAPSID framework.

SemSur contains a total of 220 instances for 14 classes. Our vision is that researchers who own a piece of research result or know about it create such instances as wiki pages where other researchers contribute to complete it. Providing semantic forms enable researchers of other domains easily create instances of research results from their community. Overall, we have 29 instances for classes, 95 instances for person, and 29 frameworks. 13 problems are instantiates with solutions and 7 without.

4 Methodology

The methodology of populating the SemSur knowledge graph is divided into two main phases: (1) select a narrow research field with many comparable approaches, problem and implementations, e.g. question answering, link discovery, SPARQL query federation or relationship extraction (2) build the knowledge graph comprising comprehensive descriptions of a specific research field and instantiating individual research articles in that field. The overall workflow of this study (see Fig. 3) comprises four steps: (1) Article selection, (2a) Formalization, (2b) Ontology development, and (3) querying the ontology to demonstrate its potential usage.

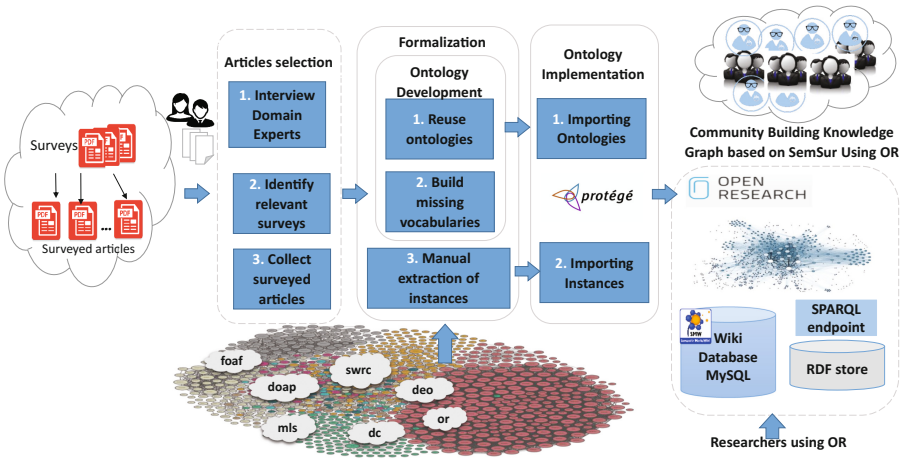

Fig. 3. Overview of the workflow for the proposed research knowledge graph population methodology.

Article selection: To demonstrate the feasibility of our approach, three survey articles for three different topics (mentioned in Sect. 1) have been selected by domain experts. These articles are used as main references by researchers in the domain to obtain an overview about frameworks, models, evaluation methods and research methodologies. From modeling these three survey articles, we collected all the 29 individual research articles covered by these surveys. These articles are addressed in the survey papers as references and we used them to create instances of the SemSur ontology.

Formalization: This step contains ontology development and manual extraction of the instances. As mentioned in Sect. 3.2, after conceptualization of the domain and interviewing experts, the SemSur ontology is instantiated. In the development of SemSur, we reused already existing and relevant ontologies with their proper suit and for the missing terms we developed our own. In parallel, we studied the articles and interviewed domain experts to extract instances describing the content of the 29 articles (done by the two first authors) based on ontology classes. This helped us to develop the ontology and also ease instance creation in the knowledge management system.

Ontology implementation: SemSur ontology was written in OWL using the Protégé, the open source ontology editor and knowledge management system. The extracted instances have been imported into Protégé, an example is shown in Fig. 2.

Querying the ontology: Querying SemSur is performed using the *SnapSPARQL* query framework [4]. A list of 30 pre-defined queries (10 for each of the three surveys) has been created for evaluating the approach, which will be presented in Sect. 5. The questions were developed with the help of domain experts;

these are common questions or requirements on which new researchers in the domain often need to spend months of research to obtain such an overview.

To provide collaboratively editable and queryable version of instances, we use the semantic wiki-based platform [13] *OpenResearch.org (OR)*. We now enhanced its ontology by importing SemSur and asking the community to cover metadata about research results, e.g., developed tools, frameworks, methodologies, etc. Using the Semantic MediaWiki extension, we are able to represent the SemSur knowledge graph, to provide an environment for its curation and for creating overviews of the respective research domains (e.g., using the evaluation queries). A sample wiki page[8] of an instance is added to OpenResearch. In the right hand side, the information box is shown in which semantic representation of the instance and its properties based on SemSur is presented.

5 Evaluation

In this section we describe the method and the results of evaluating our approach.

5.1 Evaluation Method

We first succinctly introduce the evaluation setup and then discuss the result. The evaluation started with the phase of letting researchers first read the given overview questions and letting them try in their own way to find the respective answer. We followed these steps:

- A set of 10 predefined natural language queries has been prepared for evaluation Table 4. Then, asking participants to try to answer these queries using their own tools and services. The queries were chosen in increasing order of complexity.
- We implemented SPARQL queries corresponding to each of these queries to enable non-expert participants, not familiar with SPARQL, to query the knowledge graph.
- We asked researchers to review the answers of the pre-defined queries that we formulated based on the SemSur ontology. We asked them to tell us whether they consider the provided answers and the way queries are formulated comprehensive and reasonable.
- We finally asked the same researchers to fill in a satisfaction questionnaire with 18 questions[9].

As an example, the SPARQL implementation of Q5 is listed below. Figure 4 shows the results of this query using OR.

[8] http://openresearch.org/ANAPSID:_An_Adaptive_Query_Processing_Engine_for_SPARQL_Endpoints.
[9] https://goo.gl/eZC4UL.

```
SELECT DISTINCT ?Framework ?Problem ?subProblem ?solution ?platform ?hasGUI ?t
WHERE {
  ?Framework    or:addressProblem     ?subProblem .
  ?subProblem   or:isSubproblem       ?Problem .
  ?Framework    or:provideSolution    ?solution .
  ?solution     or:hasSolution        ?subProblem .
  ?Framework    or:hasGUI             ?hasGUI .
  ?Framework    or:supportTechnology  ?t .
  ?t            foaf:name             "SPARQL␣1.1"
OPTIONAL {
  ?Framework    doap:platform         ?platform
} }
```

Table 4 shows the 10 sample SemSur knowledge graph evaluation queries for the three surveys. Note, that these are prototypical queries, which can be easily adapted to obtain similar information in other fields (we did the same for the other two surveys).

Acronym	Name	GUI	OS	Platform	Supports Technology	Implemented in	Addresses Problem
ANAPSID	An Adaptive Query Processing Engine for SPARQL Endpoints	True	Linux	ANAPSID	SPARQL 1.1	Python	SPARQL Query Federation Query_Execution Source_Selection
FedX	Optimization Techniques for Federated Query Processing on Linked Data	True	Windows	Sesame	SPARQL 1.0	Java	SPARQL Query Federation Query_Execution Source_Selection

Fig. 4. Sample overview query run on OR to show list of frameworks in SPARQL Federated Query

In the end, we asked participants to fill a questionnaire with 18 questions. The result of this evaluation is discussed in the following section.

5.2 Evaluation Results

To obtain answers of queries, 5 out of the 9 researchers immediately started with well-known standard Web search engines to explore the given topic. They tried to use several variations of keywords from the questions, e.g., "Federated Query Engines", "SPARQL Federation", etc. They also used digital libraries and scientific metadata services, e.g., ACM DL or Microsoft Academic Search, following the same approach and sometimes using advanced search options and filters. However, the retrieved results were either out of scope for the question but more related to the search keywords. All subjects unanimously agreed that the current way would not help them unless they explored more and read some survey articles on topic.

Table 4. 10 SemSur knowledge graph evaluation queries

Query #	Text
Q1	What are the possible strategies of "query execution" for DQE?
Q2	What are the programming languages used for implementing FQE over SPARQL endpoints?
Q3	Which evaluation metrics, information assets, results and benchmarks are used to evaluate LD frameworks?
Q4	What are the research problems related to database-ontology mapping?
Q5	What frameworks support SPARQL 1.1 or SPARQL 1.0 federation extension along with the platform, addressed problems, OS in which they can run, programming language used and have a GUI or not?
Q6	What are the frameworks that address the SPARQL query federation problem along with the articles where they are described, the publication year and authors names?
Q7	Which are the frameworks that solve the problem of query execution over a federation of SPARQL endpoints and support SPARQL 1.0?
Q8	What are the scientific articles that tackle the problem of generating RDF data from existing large quantities of data residing in relational databases?
Q9	What experiment setups should be considered for evaluating a DQE against SPARQL endpoints?
Q10	What are the motivations, the approaches and frameworks for current LD frameworks?

Overall, 8 researchers found it difficult to collect information and reach a conclusive overview of the research topics or related work using current methods. Six of the participants pointed out that for some of the overview questions, search engines were as good as the proposed system particularly when the framework name is part of the search keyword. They all agreed that for complicated questions our SemSur approach outperformed any existing approach/tool. Seven participants agreed that our system would be helpful for both new and experienced researchers. Two-thirds of them strongly agreed that the time and effort they spent to find such information using our system in comparison to other traditional ways is relatively low. Finally, 100% of the participants would like to use SemSur approach in their further research for studying the literature of a research topic or writing a survey article. Since the results of queries were shown to the participants in table view, the main feedback from all participants about possible improvements was to provide a better way of data representation.

6 Conclusions

In this article we presented SemSur, a Semantic Survey Ontology, and an approach for creating a comprehensive knowledge graph representing research findings.

We see this work as an initial step of a long-term research agenda to create a paradigm shift from document-based to knowledge-based scholarly communication. Our vision is to have this work deployed in an extended version of the existing OpenResearch.org platform.

We have created instances of three selected surveys on different fields of research using the SemSur ontology. We evaluated our approach involving nine researchers. As we see in the evaluation results, SemSur enables successful retrieval of relevant and accurate results without users having to spend much time and effort compared to traditional ways. This ontology can have a significant influence on the scientific community especially for researchers who want to create a survey article or write literature on a certain topic. The results of our evaluation show that researchers agree that the traditional way of gathering an overview on a particular research topic is cumbersome and time-consuming. Much effort is needed and important information might be easily overlooked. Collaborative integration of research metadata provided by the community supports researchers in this regard. Interviewed domain experts mentioned that it might be necessary to read and understand 30 to 100 scientific articles to get a proper level of understanding or an overview of a topic or sub-topics. A collaboration of researchers as owners of each particular research work to provide a structured and semantic representation of their research achievements, can have a huge impact in making their research more accessible. A similar effort is spent on preparing survey and overview articles.

Integrating our methodology with the procedure of publishing survey articles can help to create a paradigm shift. We plan to further extend the ontology to cover other research methodologies and fields. For a more robust implementation of the proposed approach, we are planning to use and significantly expand the OpenResearch.org platform and a user-friendly SPARQL auto-generation services for accessing metadata analysis for non-expert users. More comprehensive evaluation of the services will be done after the implementation of the curation, exploration and discovery services. In addition, our intention is to develop and foster a living community around OpenResearch.org and SemSur, to extend the ontology and to ingest metadata to cover other research fields.

Acknowledgments. This work has been supported by the H2020 project no. 645833 (OpenBudgets.eu). The authors would like to thank Prof. Maria-Esther Vidal and Afshin Sadeghi for their support. We also appreciate the help of all participants of the evaluation. This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

References

1. Antoniou, G., Van Harmelen, F.: A Semantic Web Primer. MIT Press, Cambridge (2004)
2. Bryl, V., et al.: What's in the proceedings? combining publisher's and researcher's perspectives. In: Proceedings of the 4th Workshop on Semantic Publishing (SePublica) (2014)
3. Capadisli, S., Riedl, R., Auer, S.: Enabling accessible knowledge. In: Conference for E-Democracy and Open Governement, p. 257 (2015)
4. Horridge, M., Musen, M.: Snap-SPARQL: a java framework for working with SPARQL and OWL. In: Tamma, V., Dragoni, M., Gonçalves, R., Ławrynowicz, A. (eds.) OWLED 2015. LNCS, vol. 9557, pp. 154–165. Springer, Cham (2016). doi:10.1007/978-3-319-33245-1_16
5. Musen, M.A.: The Protég'e project: a look back and a look forward. In: AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, vol. 1(4) (2015)
6. Nentwig, M., et al.: A survey of current link discovery frameworks. Semant. Web **8**(3), 419–436 (2017)
7. Noy, N.F., McGuinness, D.L., et al.: Ontology development 101: A guide to creating your first ontology (2001)
8. Peroni, S., et al.: Research articles in simplified HTML: a web-first format for HTML-based scholarly articles. Technical report, PeerJ Preprints (2016)
9. Rakhmawati, N.A., et al.: Querying over federated SPARQL endpoints - a state of the art survey. In: CoRR abs/1306.1723 (2013)
10. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing **22**(2), 85–94 (2009)
11. Simperl, E.: Reusing ontologies on the semantic web: a feasibility study. Data Knowl. Eng. **68**(10), 905–925 (2009)
12. Spanos, D.-E., Stavrou, P., Mitrou, N.: Bringing relational databases into the semantic web: a survey. Semant. Web **3**(2), 169–209 (2012)
13. Vahdati, S., Arndt, N., Auer, S., Lange, C.: OpenResearch: collaborative management of scholarly communication metadata. In: Blomqvist, E., Ciancarini, P., Poggi, F., Vitali, F. (eds.) EKAW 2016. LNCS (LNAI), vol. 10024, pp. 778–793. Springer, Cham (2016). doi:10.1007/978-3-319-49004-5_50