Semantic Representation of Scientific Publications

Poster · August 2019

DOI: 10.13140/RG.2.2.29953.40806

5 authors, including:

Sahar Vahdati
University of Bonn
47 PUBLICATIONS **187** CITATIONS
SEE PROFILE

Said Fathalla
University of Bonn
34 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Sören Auer
Leibniz Universität Hannover
452 PUBLICATIONS **13,181** CITATIONS
SEE PROFILE

Christoph Lange
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
261 PUBLICATIONS **1,651** CITATIONS
SEE PROFILE

Semantic Representation of Scientific Publications

Sahar Vahdati[1], Said Fathalla[1,3], Sören Auer[4,5],
Christoph Lange[1,2], and Maria-Esther Vidal[4,5]

[1] Smart Data Analytics (SDA), University of Bonn, Germany
{vahdati,fathalla,langec}@cs.uni-bonn.de
[2] Fraunhofer IAIS, Germany
[3] Faculty of Science, Alexandria University, Egypt
[4] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
[5] L3S Research Centre, Leibniz University of Hannover, Germany
maria.vidal,soeren.auer@tib.eu

Abstract In this work, we tackle the problem of generating comprehensive overviews of research findings in a structured and comparable way. To bring structure to such information and thus to enable researchers to, e.g., explore domain overviews, we present an approach for automatic unveiling of realm overviews for research artifacts (Aurora), an approach to generate overviews of research domains and their relevant artifacts. Aurora is a semi-automatic crowd-sourcing workflow that captures such information into the OpenResearch.org semantic wiki. Our evaluation confirms that Aurora, when compared to the current manual approach, reduces the effort for researchers to compile and read survey papers.

Keywords: Metadata Extraction, Semantic Publishing, Knowledge Graphs

1 Introduction

The goal of preparing survey articles is to summarize the benefits and limitations of the existing research work in a specific research topic by producing, for example, tables of comparisons for state-of-the-art tools [3]. However, creating such document-based overviews requires significant effort from the researchers to not only describe the main characteristics of the examined approaches but also to compare and position them. We tackle the problem of semantically describing scientific papers using overviews that represent the *main characteristic* of a paper. We present Aurora, a semi-automatic crowd-sourcing workflow that captures information about research contributions in the *OpenResearch.org* semantic wiki; it facilitates the description of the scientific papers with the SemSur ontology [1] into a knowledge graph. Further, Aurora enables the generation of surveys comprising comprehensive analytics which cover different research domain overviews by querying the knowledge graph. Fathalla et al. [1, 2] define an ontology that contains core concepts for describing scholarly documents, whereas Aurora facilitates the usage of ontologies. SemSur ontology is employed as the reference ontology for modeling research findings in scientific articles [2]. Knowledge encoded in document-oriented scientific publications reveals the *main characteristic* of the approach defined in a paper; they correspond to the results of

multiple steps of the research methodology, e.g., problem statement, proposed approach, and evaluation methods. Exploiting semantic representations of scientific papers involves: Figure 1 depicts the pipeline of Aurora based on the aforementioned required steps. Aurora receives a set of scholarly artifacts and annotations provided by the crowd, and outputs a knowledge graph representing the main parts of the artifacts. Traversing and exploring the generated knowledge graph facilitates to analysis and comparison of the existing approaches in a particular domain. In order to demonstrate the descriptive capability of the Aurora knowledge graph, several complex questions are evaluated. The observed results suggest that representing research artifacts in terms of domain overviews facilitates the exploration of related approaches. Our contributions are summarized as follows: 1) An ontology for creating a scholarly knowledge graph representing research contributions in a semantic way, 2) Aurora, a framework implemented on top of the crowd-sourcing platform *OpenResearch.org* which relies on extraction and curation methods for a scholarly knowledge graph, and 3) Providing domain overviews and suggestions such as which publication to read, which tools to use, where to publish similar results on sub-networks of the knowledge graph.

2 The proposed Solution

This work focuses on facilitating scholarly communication by semantically representing elements of regular research articles and other types of scholarly artifacts, e.g., tools and frameworks. As mentioned initially, writing survey articles is a traditional but still the most widespread way of creating a systematic overview about scholarly artifacts. Survey articles summarize and organize the state of the art and emphasize the classification of the existing developments, frameworks, or evaluation methods. Despite legal and technical improvements to the access to scientific publications and despite a number of services developed for supporting scholarly communication, such as research repositories, search engines and digital libraries, it is often a time-consuming task to find information such as: papers containing certain information about a topic of interest, a comparison of state of the art tools, and Overview of approaches addressing a particular research problem. To support scholars who want to find and learn from such overviews about a domain or those researchers attempting to create such overviews, we aim at representing such information in a research knowledge graph underlying our *OpenResearch.org* (OR) platform.

3 Metaresaerch Queries

In this section, we describe the design and expectation procedure of a set of metaresaerch queries on top of the Aurora knowledge graph. The goal of this representation is to evidence the descriptive capability of the Aurora knowledge graph in terms of complex queries. The queries are relatively hard to directly answer using existing scholarly search engines ad other services e.g., Google

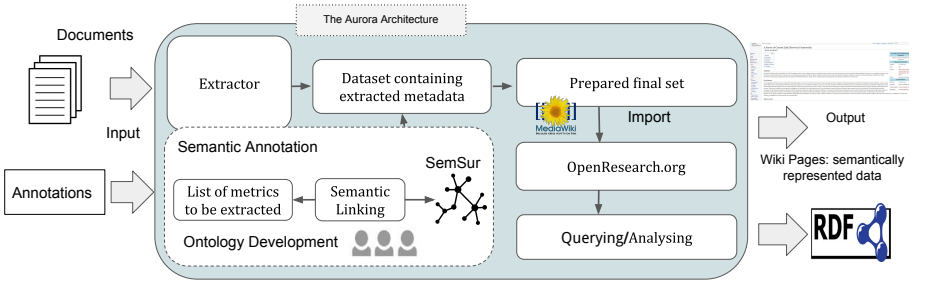

Figure 1: Aurora Pipeline.

Schoolar, DBLP, Semantic Scholar. The answer for such queries requires content of scholarly publications, particularly survey papers which include summaries and comparisons. For instance, the table on the left side in Figure 2 is one of these tables, extracted from the survey paper "Querying over Federated SPARQL Endpoints–A State of the Art Survey" [4]. Creating such a table takes considerable time and effort [2] for a researcher providing a survey paper for the community. This evaluation aims at illustrating how Aurora enables the execution of complex queries that requires information from different sources about various artifacts. To answer such queries, normally researchers are required to do a comprehensive literature review on a specific topic by being able to conclude such comparison tables. Figure 2 shows a comparison between a table in a survey paper and a corresponding table generated by our approach.

Table II: The Existing Frameworks Supports Federation over SPARQL Endpoints without reformulating query to SPARQL 1.1.

Framework	Catalogue	Platform	Source Selection	Cache	Query Execution	Source Tracking	GUI
DARQ	Service Description	Jena	Statistic of Predicate	✓	Bind Join or Nested Loop Join	Static	✗
ADERIS	Predicate List during setup phase	✗	Predicate List	✗	Nested Loop Join	Static	✓
FedX	✗	Sesame	ASK	✓	Bind Join paralleliza-tion	Dynamic	✓
Splendid	VoID	Sesame	Statistic + ASK	✗	Bind Join or Hash Join	Static	✗
GDS	Service Description	Jena	Statistic of Predicate	✓	Bind Join or Semi Join	Dynamic	✗
Avalanche	Search Engine	Avalanche	Statistic of predicates and ontologies	✓	Bind join	Dynamic	✗
Distributed SPARQL	✗	Sesame	✗	✗	Bind join	✗	✗

Table generated by human in a survey paper

Tool/Ontology	Data Catalogue	Platform	GUI
ANAPSID	Predicate list and endpoint status	ANAPSID	Yes
SemWIQ	RDF stats + VoID	Jena	Yes
Avalanche	Search Engine	Avalanche	No
FedX	-	Sesame	Yes
DARQ	Service Description	Jena	No
GDS	Service Description	Jena	No
WoDQA	VoID stores	Jena	Yes
SPLENDID	VoID	Sesame	No
ADERIS	Predicate List during setup phase	-	Yes
Distributed SPARQL	-	Sesame	No

Table *automatically* generated by OpenResearch.org

Figure 2: A table included in a survey paper (left-side) is compared to the table generated (right-side) by querying the knowledge graph.

Aurora is implemented on top of the OpenResearch.org platform which is built upon Semantic MediaWiki [5]. In order to obtain sample queries to be implemented by Aurora, five researchers brainstormed (authors of this paper– three senior and two junior) from the domain of Linked data and Knowledge Engineering and Data Management. A set of ten predefined natural language queries has been finalized to be implemented as *ASK* queries. To demonstrate

the efficiency of our proposed approach, we used a set of four survey papers and extracted information of 36 papers surveyed, as a seed in order to create a knowledge graph representing research findings in these papers referenced by the four survey papers. Assume that a researcher wants to answer these queries using the current scholarly search engines, how much time and effort does it take to answer these queries? These ten queries are currently implemented and made available on *OpenResearch.org*[6]. These queries cover essential findings presented in the four survey articles; they are chosen in increasing order of complexity.

For getting an overview of a certain topic, we prepared a query: *Query 1*: get an overview of the Tools addressing the research problem "SPARQL Query Federation". The results of such query are available at `http://openresearch.org/wiki/Papers_query1`.

4 Discussion

We presented Aurora for representing research findings in computer science in a semantic way and crowd-sourcing the creation of these semantic representations using a semantic wiki. We evaluated the approach with a number of competency questions, which can now be answered using Aurora and simple queries, instead of long-term survey compilation. We believe that Aurora is a one of the initial steps towards transforming the current document-based information flows in scholarly communication into knowledge-based ones. Aurora follows the problem-approach-implementation-evaluation methodological pattern, which is widespread in computer science and other engineering fields.

References

1. Fathalla, S., Vahdati, S., Auer, S., Lange, C. SemSur: A Core Ontology for the Semantic Representation of Research Findings. In: *Proceedings of the 14th International Conference on Semantic Systems*. ACM. 2018. In press.
2. Fathalla, S., Vahdati, S., Auer, S., Lange, C. Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, p. 315.
3. Kitchenham, B. Procedures for performing systematic reviews. In: Keele, UK, Keele University 33(2004) (2004), pp. 1–26.
4. Rakhmawati, N. A., Umbrich, J., Karnstedt, M., Hasnain, A., Hausenblas, M. Querying over Federated SPARQL Endpoints - A State of the Art Survey. In: CoRR abs/1306.1723 (2013).
5. Vahdati, S., Arndt, N., Auer, S., Lange, C. OpenResearch: Collaborative Management of Scholarly Communication Metadata. In: *EKAW*. 2016.

[6] `https://www.openresearch.org/wiki/Papers`