Towards a Knowledge Graph for a Research Group with Focus on Qualitative Analysis of Scholarly Papers

Vera G. Meister[1][0000-0002-2780-0222]

[1] Brandenburg University of Applied Sciences, Brandenburg a. d. H., Germany
vera.meister@th-brandenburg.de

Abstract. Support of scientific workflows by semantic technology gains increasing interest in recent years. Huge efforts are put on providing structured, standard-based meta data and on machine based qualitative analysis of unstructured content of scholarly papers. This helps researchers to stay oriented in an ever growing and gaining complexity field. Semantic technologies have also the potential to support the in-depth involvement in scholarly papers, like practiced in research seminars. The paper reports on the preliminary results of an undertaking to support the collaborative documentation and reuse of qualitative analysis of scholarly papers in an information systems research group. A vocabulary is developed and openly provided. The system is implemented on the base of OntoWiki and can be accessed openly.

Keywords: Qualitative literature analysis, Scientific workflows, Research Group Knowledge Base, Collaborative annotation.

1 Introduction

Research groups form the smallest, often informal social entity in the scientific system. Their performance and their cohesion are mainly based on shared scientific interests and a common, high level of expertise in the research field. Even if this research field is narrowly specified, it remains a great challenge to keep in view the state of knowledge. Beyond the awareness of other research groups and influential researchers in the field, a qualitative expert analysis should focus on research questions, on methods applied to them, as well as on research findings and their critical disputation. Undoubtedly, regular scientific seminars are a traditional and effective instrument for this, since they create a collective realm of experience and discussion.

The small, informal research group Business Modeling and Knowledge Engineering (BMaKE) at the Brandenburg University of Applied Sciences has established such a seminar recently. This group is anchored in the program of information systems. While the selection of the papers to be discussed and the structure to be used in expert analysis were quickly agreed, the form of the knowledge base to be created for storing the analysis results led immediately to the following research question:

- How to build a sustainable infrastructure for storing the knowledge, collectively worked out in seminars, in a systematic, structured and easy to re-use way?

The collaboration environments and systems successfully used so far in project work and teaching (Google Drive, GitHub, Confluence, Slack) are quite suitable for the exchange of data and information. They fall short in providing a systematic knowledge storage which can be queried flexibly, since they don't implement the necessary knowledge graph structure.

At this point, the research question has not yet been definitively answered. The paper aims at presenting the initiated approach and at discussing the experiences so far. Therefore, the remainder of the paper is organized as follows: Section 2 provides an overview of relevant work on semantic analysis and structuring of scholarly papers content. The elaborated vocabulary to support the knowledge base is presented in Section 3, whereas Section 4 introduces the preliminary system design for the targeted knowledge base. Section 5 reflects the first implementation experiences. The paper closes with a short conclusion and an outlook on further work in Section 6.

2 Related Work

There are different lines of research dealing with semantic analysis and the deployment of structured data on scholarly papers and other relevant objects of scientific environments and workflows, like conferences, proceedings, people, and projects. Table 1 gives a brief overview mentioning exemplary work in the field as well as main research objectives and findings for each of these lines.

Table 1. Lines of research in scholarly papers analysis and structuring

Line of research	Exemplary work	Research Objectives	Main findings
Meta data extraction	Adding semantics to digital libraries [1]	Provide meta data in a standard-based, reusable and structured way	Linked open data publications framework
Collaborative annotation	OpenResearch collaborative management [2]	Enrich structured data about scholarly papers and/or related events	Data model, System architecture based on SMW, LOD services
Production of natively structured data	RASH framework enabling HTML+RDF submissions [3]	Establish standards, formats and frameworks for natively providing structured data	Specification for writing research articles in simplified HTML (RASH)
Text analysis, data mining and machine learning	Knowledge extraction from scientific publications [4]	Elicitation of inner semantics hidden in texts, figures and other unstructured data	Dr. Inventor Text Mining Framework for automated analysis of scientific publications

The results of meta data extraction projects like presented in Table 1 can be used as basic input for the research group knowledge base. The undertaking itself is a kind of collaborative annotation, but with a more specific focus. The increasing production of natively structured data will also support a basic input – as it looks today. However, it is not impossible that this form of publication will also support very specific, qualitative analysis questions in the future. The methods of text analysis and machine learn-

ing are the closest to the qualitative analysis of scholarly papers. Though, since a qualitative analysis is very field-specific, a high-quality training set is required. Perhaps the knowledge base presented here can serve as a training set for automatic qualitative analysis for scholarly papers in the field of Business Modeling and Knowledge Engineering from the Information Systems' perspective.

3 Vocabulary for Qualitative Analysis of Scholarly Papers

Like stated above, the main objective of the required knowledge base is to support the research group's collective analysis of scientific publications in the field of information systems. It is therefore obvious to structure scholarly papers according to their main qualitative features: (i) research objectives, (ii) research methods, (iii) research findings, (iv) future work, and (v) critical issues (comp. e.g. [5]). To allow semantically rich queries to the knowledge base, these features shall be further structured, whenever possible. Candidates for doing this are the research methods and the research objectives. The main research methods in information systems are described in [6]. For structuring the research objectives, a flexible, pairwise combination of research activities and research artifacts can be applied. Both can be modeled as clear enumerations when a limited research field is considered (see Table 2).

Table 2. Field-specific enumerations for qualitative analysis of scholarly papers

Research objective		Research method
Research activity	Research artifact	
analyze	Application	Action Research
collect	Blueprint	Argumentative Deductive Analysis
conceptualize	Business Process	Case Study
construct	Development Framework	Conceptual Deductive Analysis
define	Documentation	Design Science (Hevner)
design	Infrastructure	Ethnography
develop	IT System	Field Experiment
elicit	Linked Data	Formal Deductive Analysis
enhance	Method	Grounded Theory
evaluate	Modeling Language	Laboratory Experiment
extend	NLP Artifact	Literature Analysis
extract	Ontology	Prototyping
implement	Policy	Qualitative Research
integrate	Requirements	Quantitative Research
prove	Standard	Reference Modeling
provide	Term Definition	Simulation
structure	Workflow	

Two independent approaches were pursued in the search for reusable classes, relations and attributes for the required knowledge schema. As a vocabulary with increasing importance for websites first *Schema.org* was examined. It was found that the rather formal, accompanying information on scholarly papers necessary for the use case can be modeled adequately with elements of this vocabulary. The mentioned

above qualitative features of papers may reuse the relation *schema:about*, but no fitting elements were found themselves. For filling the gaps, the SPAR Ontologies [7], in particular the Discourse Elements Ontology (DEO), were considered in more detail. The arguments for not reusing DEO are the following:

1. There are substantial differences between rhetorical elements used by an author or detected by automatic text analysis, as assumed in DEO, and qualitative features of a paper detected by expert analysis, as intended here. E.g. critical issues are an individual estimation of a human reader and therefore are not provided in DEO.
2. As already stated, some of the analyzed features are to be specified as enumerations. DEO don't implement such constraints on data types.

Therefore, these entities were modeled as new specific classes which nevertheless are semantically and structurally integrated in the *Schema.org* frame. Fig. 1 shows the high-level schema of the vocabulary. Red nodes are taken from *Schema.org*, the other ones are specifically modeled. Green nodes are of type Enumeration, whereas the white nodes stand for abstract concepts implementing a list structure. The vocabulary is documented on GitHub[1] and referenced in LOV[2].

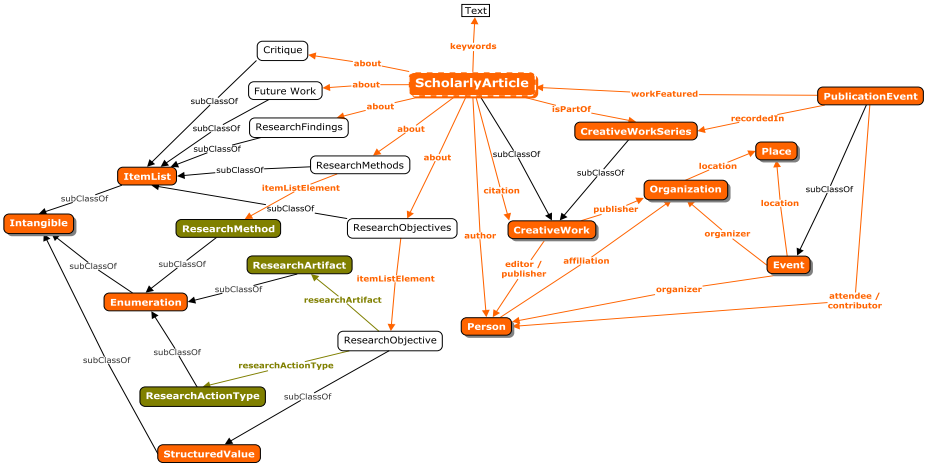

Fig. 1. High-level schema of the scholarly papers vocabulary

4 Preliminary System Design of the Knowledge Graph

The target system can be described as a knowledge graph, as defined in [8] and further specified in [9]. Fig. 2 shows an abstract model of this knowledge graph where the characteristic elements, particularly the exploited knowledge sources and the provided knowledge services, are adapted to the use case under consideration. The shaded items in the model represent already implemented, at least partly, elements. The boxes with dots reflect the further extensibility of the system.

---

1 https://github.com/bmake/scholarlygraph/

2 http://lov.okfn.org/dataset/lov/vocabs/spvqa

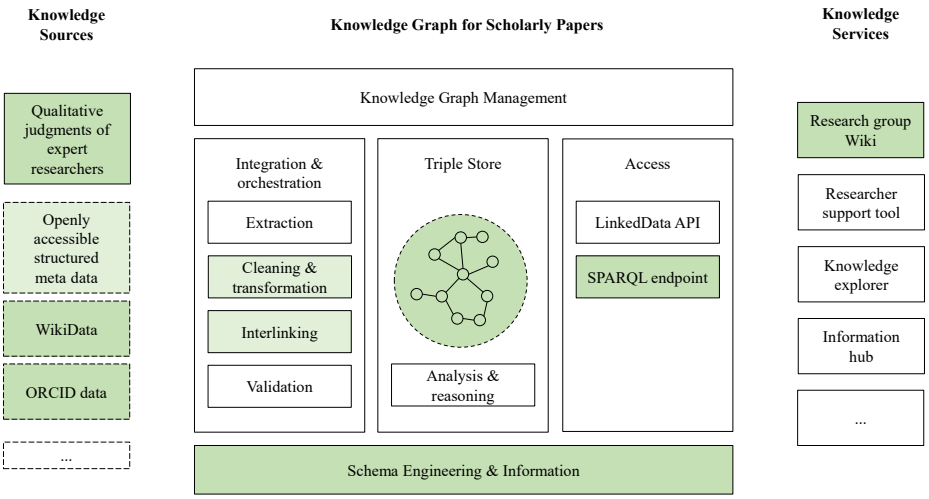

Fig. 2. Abstract model of the knowledge graph for scholarly papers

Now, the system is implemented as an out of the box OntoWiki [10] comprising a standard wiki interface, a triple store and a SPARQL endpoint. It is populated manually by researchers during their qualitative analysis of seminar papers. Even external sources of structured meta data are for the time of writing queried and interlinked manually. Editing is supported either by Turtle templates for creating importable data dumps or can be performed directly in the wiki. This preliminary workflow is additionally used to evaluate processes and sources for automatic data input. Vocabulary (schema) information is provided by the documentation mentioned above.

5 First Implementation Experience

The preliminary implementation as described in the previous section can be considered as a research prototype[3]. Since the system aims at the structured documentation and flexible reuse of seminar output of the BMaKE research group, the knowledge base is growing slowly, but continuously. At the time of writing, 35 scholarly papers from 11 publications correlated with 9 publication events (conferences) are analyzed. They are interlinked with more than 100 authors, nearly 50 organizations and places. Each month 5 to 10 new papers will be analyzed and added to the knowledge base.

The immediate support of the research group's work allows an in-process evaluation of the support quality and a deeper elicitation of needs and requirements. The first experiences in using the system in the context of scientific seminars shows the following informal results:

1. Pure consumers of the system assessed it as very helpful in gathering deeper knowledge in the research field.

[3] https://bmakewiki.th-brandenburg.de

2. Active editors reported very clearly the necessity of implementing automated bulk import for the formal metadata of scholarly papers.
3. Overall, it becomes obvious, that the used out of the box system does not support natively a range of required forms and visualizations. Even the support of external linked data is weak. Hence, the system shall be modified by custom application development, preferably by means of the OntoWiki framework [10].

6 Conclusion and Further Work

According to preliminary assessment, a knowledge graph can be considered as a sustainable infrastructure for storing and reusing the results of qualitative analyses of scholarly papers. Even the preliminary implementation presented in this paper was evaluated as an effective (even if up to now not efficient) measure to support the work of a research group. There are three main lines of further development of the system: (i) Formal meta data which are not object of qualitative analysis must be integrated in an automatic way reusing structured data provided by open sources. (ii) A well-usable template-based form should be developed for capturing the results of the qualitative analysis. (iii) The use cases for the support of the research work must be elicited systematically and on this basis the research group wiki should be adapted. These development steps shall than be followed by a formal, structured evaluation.

References

1. Aslam, M.A. e. a.: A Generic Framework for Adding Semantics to Digital Libraries. In: Ciuciu, I. e. a. (eds.) OTM 2016. LNCS, vol. 10034, pp. 277-281. Springer, Cham (2017).
2. Vahdati, S., e. a.: OpenResearch - Collaborative Management of Scholarly Communication Metadata. In: Blomqvist E., Ciancarini P., Poggi F., Vitali F. (eds.) EKAW 2016. LNCS, vol. 10024, pp. 778-793. Springer, Cham (2016).
3. Di Iorio, A, e. a.: The RASH Framework. In: ISWC 2015, Poster & Demo Session, http://ceur-ws.org/Vol-1486/paper_72.pdf, last accessed 2017/08/05.
4. Ronzano, F., Saggion, H.: Knowledge Extraction and Modeling from Scientific Publications. In: González-Beltrán A., Osborne F., Peroni S. (eds.) SAVE-SD 2016. LNCS, vol. 9792, pp. 11-25. Springer, Cham (2016).
5. Rossig, W., Prätsch, J.: Wissenschaftliche Arbeiten. 7th edn. BerlinDruck, Achim (2008).
6. Wilde, T., Hess, T.: Forschungsmethoden der Wirtschaftsinformatik – Eine empirische Untersuchung. In: WIRTSCHAFTSINFORMATIK 49(4), 280–287 (2007).
7. Peroni, S.: The Semantic Publishing and Referencing Ontologies. In: Semantic Web Technologies and Legal Scholarly Publishing, pp. 121-193. Springer, Cham (2014).
8. Paulheim, H.: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. In: Semantic Web Journal (Preprint), 1–20 (2016).
9. Meister, V., Jetschni, J., Kreideweiß, S.: Konzept und Prototyp einer dezentralen Wissensinfrastruktur zu Hochschuldaten für Mensch und Maschine. In: INFORMATIK 2017. LNI. GI e. V., Bonn (2017) *in print*.
10. Frischmuth, P., Arndt, N., Martin, M.: OntoWiki 1.0. In: SEMANTiCS 2016, Poster & Demo Session, http://ceur-ws.org/Vol-1695/paper11.pdf, last accessed 2017/08/05.