Table 1: Overview of answers to the key aspects covered by the evaluation questionnaire and other metrics recorded during the interviews.

Participant Nr	Navigation	Terminology	Auto Complete	Guidance Needed	Suggest To Others	UI likeness	Time
	5 = Very intuitive	5 = Easy to understand	5 = Very helpful	5 = All the time	9 = Very likely	9 = Very much	in mins
1	4	4	5	3	2	6	16
2	2	3	5	4	8	7	19
3	4	5	5	3	9	7	15
4	3	3	5	3	6	7	13
5	4	3	5	3	6	8	14
6	4	3	5	3	8	9	13
7	3	4	5	3	7	6	19
8	3	2	4	3	8	6	13
9	4	5	3	3	7	5	14
10	4	5	5	1	8	8	22
11	4	5	5	1	8	8	20
12	-	-	-	-	-	-	21
Average	**4**	**4**	**5**	**3**	**7**	**7**	**17**

classification and linking. As a first step, we trained a neural network based machine learning model for named entity recognition using in-house developed annotations on the Elsevier Labs corpus of Science, Technology, and Medicine[4] (STM) for the following generic concepts: process, method, material and data. We use the Beltagy et al. [4] Named Entity Recognition task-specific neural architecture atop pretrained SciBERT embeddings with a CRF-based sequence tag decoder [18].

Linking scholarly knowledge to other knowledge graphs including those from the open domain as well as domain specific graphs such as ULMS [5] is another important feature. Most importantly, such linking enables semi-automated enrichment of research contributions.

3 EVALUATIONS

The ORKG infrastructure, its services, features, performance and usability are continually evaluated to inform the next iteration and future developments. Among other preliminary evaluations and results, we present here the first front end user evaluation.

Following a qualitative approach, the evaluation of the first iteration of front end development aimed to determine user performance, identify major (positive and negative) aspects, and user acceptance/perception of the system. The evaluation process had two components: (1) instructed interaction sessions and (2) a short evaluation questionnaire. This evaluation resulted in data relevant to our first research question.

Supported by two instructors, we conducted instructed interaction sessions with 12 authors of articles presented at the DILS2018[5] conference. At the start of each session, the instructor briefly explained the underlying principles of the infrastructure. Then, participants engaged with the system without further guidance from the instructor. However, at any time they could ask the instructor for assistance. For each participant, we recorded the time required to complete the task (to determine the mean duration of a session), the instructor's notes and the participant's comments.

[4]https://github.com/elsevierlabs/OA-STM-Corpus
[5]https://www.springer.com/us/book/9783030060152

Table 2: Time (in seconds) needed to perform State-of-the-Art comparisons with 2-8 research contributions using the baseline and ORKG approaches.

	Number of compared research contributions						
	2	**3**	**4**	**5**	**6**	**7**	**8**
Baseline	0.00026	0.1714	0.763	4.99	112.74	1772.8	14421
ORKG	0.0035	0.0013	0.01158	0.02	0.0206	0.0189	0.0204

In addition to the instructed interaction sessions, participants were invited to complete a short evaluation questionnaire. The questionnaire is available online[6]. Treated as a qualitative instrument, its aim was to collect further insights into user experience. The paper-based questionnaire consisted of 11 questions. These were designed to capture participant thoughts regarding the positive and negative aspects of the system following their instructed interaction session. Participants completed their questionnaire after the instructed interaction session. All 12 participants answered the questionnaire. The interaction notes, participant comments and the time recordings were collected together with questionnaire responses and analysed in light of our research questions.

A dataset summarizing the research contributions collected in the experiment is available online[7]. The data is grouped into four main categories. *Research Problem* describes the main question or issue addressed by the research contribution. *Approach* describes the solution taken by the authors. *Implementation & Evaluation* were the most comprehensively described aspects, arguably because it was easier for participants to describe technical details compared to describing the problem or the approach.

In summary, 75% of the participants found the front end developed in the first iteration fairly intuitive and easy to use. Among the participants, 80% needed guidance only at the beginning while 10% did not need guidance. The time required to complete the task was 17 minutes on average, with a minimum of 13 minutes and a maximum of 22 minutes.

Further details of the questionnaire, including participant ratings on main issues, are summarized in Table 1. While the cohort of participants was too small for statistically significant conclusions, these results provided a number of important suggestions that informed the second iteration of front end development, which had a first evaluation at TPDL2019[8].

We have performed preliminary evaluations also of other components of the ORKG infrastructure. The experimental setup for these evaluations was an Ubuntu 18.04 machine with Intel Xeon CPUs 12×3.60 GHz and 64 GB memory.

With respect to the literature comparison feature, we compared our approach in ORKG with a baseline approach that uses brute force to find the most similar predicates and thus checks every possible predicate combination. Table 2 shows the time needed to perform the comparison for the baseline approach and for the approach we implemented and presented above. As the results suggest, our approach clearly outperforms the baseline and the performance gain can be attributed to more efficient retrieval. The experiment is

[6]https://doi.org/10.5281/zenodo.2549918
[7]https://doi.org/10.5281/zenodo.3340954
[8]http://www.tpdl.eu/tpdl2019/

limited to 8 contributions because the baseline approach does not scale to larger sets.

We also tested the vertical scalability in terms of response time. For this, we created a synthetic dataset of papers. Each paper includes one research contribution described by three statements. The generated dataset contains 10 million papers or 100 million nodes. We tested the system with variable numbers of papers and the average response time to fetch a single paper with its related research contribution is 60 ms. This suggests that the infrastructure can handle large amounts of scholarly knowledge.

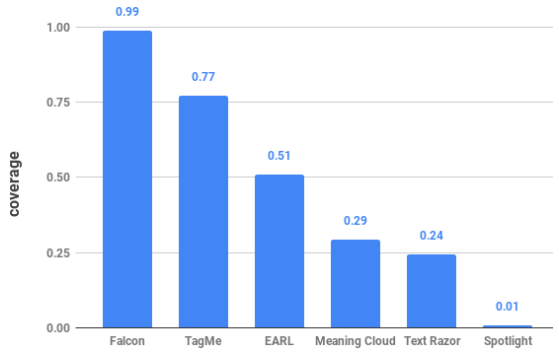

Figure 1: Coverage values of different NED systems over the annotated entities of the STM corpus.

We evaluated the performance of a number of existing NED tools on scholarly knowledge, specifically Falcon [23], DBpedia Spotlight [20], TagME [15], EARL [14], TextRazor[9] and MeaningCloud[10]. These tools were used to link to entities from Wikidata and DBpedia. We used the annotated entities from the STM corpus as the experimental data. However, since there is no gold standard for the dataset, we only computed the coverage metric $\zeta = \frac{\#\ of\ linked\ entities}{\#\ of\ all\ entities}$. Figure 1 summarizes the coverage percentage for the evaluated tools. The results suggest that Falcon is most promising.

4 CONCLUSION

This article described the first steps of a larger research and development agenda that aims to enhance document-based scholarly communication with semantic representations of communicated scholarly knowledge. We presented the architecture of the proposed infrastructure and some of its key features. We reported the results of a first user evaluation. By integrating crowdsourcing and automated techniques in natural language processing, initial steps were also taken and evaluated that advance multi-modal scholarly knowledge acquisition using the ORKG.

ACKNOWLEDGMENTS

This work was co-funded by the European Research Council for the project ScienceGRAPH (Grant agreement ID: 819536) and the TIB Leibniz Information Centre for Science and Technology. The authors would like to thank the participants of the ORKG workshop series, for their contributions to ORKG developments. We also like to thank our colleagues Kemele M. Endris, Viktor Kovtun, Arthur Brack and Anett Hoppe for their contributions.

[9]https://www.textrazor.com/docs/rest
[10]https://www.meaningcloud.com/developer

REFERENCES

[1] Hugo F. Alrøe and Egon Noe. 2014. Second-Order Science of Interdisciplinary Research: A Polyocular Framework for Wicked Problems. *Constructivist Foundations* (2014), 65–76.
[2] Amir Aryani and Jingbo Wang. 2017. Research Graph: Building a Distributed Graph of Scholarly Works using Research Data Switchboard. In *Open Repositories CONFERENCE*.
[3] Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. 2018. Towards a Knowledge Graph for Science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*.
[4] Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. SciBERT: Pretrained Contextualized Embeddings for Scientific Text. *CoRR* abs/1903.10676 (2019).
[5] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004).
[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.
[7] Jeroen Bosman, Ian Bruno, Chris Chapman, Bastian Greshake Tzovaras, Nate Jacobs, Kramer, and et al. 2017. The Scholarly Commons - principles and practices to guide research communication. https://doi.org/10.31219/osf.io/6c2xt
[8] Brodaric Boyan, Reitsma Femke, and Qiang Yi. 2008. SKIing with DOLCE: toward an e-Science Knowledge Infrastructure. *Frontiers in Artificial Intelligence and Applications* 183, Formal Ontology in Information Systems (2008), 208–219.
[9] Adrian Burton, Hylke Koers, Paolo Manghi, Markus Stocker, Martin Fenner, Amir Aryani, Sandro La Bruzzo, Michael Diepenbroek, and Uwe Schindler. 2017. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Magazine* Volume 23, 1/2 (2017).
[10] Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. 2018. The GeoLink knowledge graph. *Big Earth Data* 2, 2 (2018), 131–143.
[11] Herbert Van de Sompel and Carl Lagoze. 2009. All aboard: toward a machine-friendly scholarly communication system. In *The Fourth Paradigm*.
[12] Anita De Waard, Leen Breure, Joost G Kircz, and Herre Van Oostendorp. 2006. Modeling rhetoric in scientific publications. In *International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006*.
[13] P. Donohoe, J. Sherman, and A. Mistry. 2015. The Long Road to JATS. In *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2015 [Internet]*. Bethesda (MD): National Center for Biotechnology Information (US).
[14] Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri, and Jens Lehmann. 2018. EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs. In *The Semantic Web – ISWC 2018*. Springer International Publishing, Cham, 108–126.
[15] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities).. In *CIKM*, Jimmy Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson, and Aijun An (Eds.). ACM, 1625–1628.
[16] Karen L. Hanson, Tim DiLauro, and Mark Donoghue. 2015. The RMap Project: Capturing and Preserving Associations Amongst Multi-Part Distributed Publications. In *Proceedings of the 15th ACM/IEEE-CE JCDL*. ACM, 281–282.
[17] Alexander Hars. 2001. Designing Scientific Knowledge Infrastructures: The Contribution of Epistemology. *Information Systems Frontiers* 3, 1 (2001), 63–73.
[18] Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *CoRR* abs/1603.01354 (2016).
[19] Vera G. Meister. 2017. Towards a Knowledge Graph for a Research Group with Focus on Qualitative Analysis of Scholarly Papers. In *Proceedings of the First Workshop on Enabling Open Semantic Science co-located with 16th International Semantic Web Conference (ISWC 2017)*. 71–76.
[20] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL* 2 (2014), 231–244.
[21] Silvio Peroni. 2014. The Semantic Publishing and Referencing Ontologies. In *Semantic Web Technologies and Legal Scholarly Publishing*. Law, Governance and Technology, Vol. 15. Springer, Cham, 121–193.
[22] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Piscataway, NJ, 133–142.
[23] Ahmad Sakor, Isaiah Onando Mulang, Kuldeep Singh, Saeedeh Shekarpour, Maria Esther Vidal, Jens Lehmann, and Sören Auer. 2019. Old is gold: linguistic driven approach for entity and relation linking of short text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2336–2346.
[24] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.