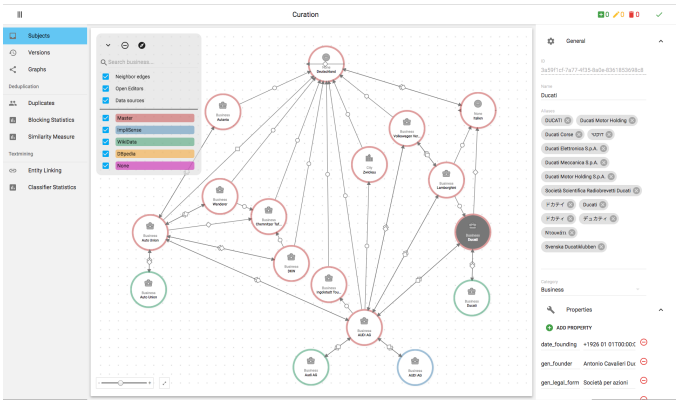

Figure 3: Graph view in the Curation Interface

4 DEMONSTRATION OVERVIEW

First we present the curation of the knowledge base including the monitoring of the intermediate data processing steps. Afterwards we demonstrate the graph exploration using ELEX. During the demonstration we will guide the participants through the following steps:

(1) **Data inspection**: We start by giving an overview of the business entities in the knowledge base. We search for a specific business entity and present various possibilities to present the available information. We then show the subgraph around the selected business entity and demonstrate the difference between the conventional tabular view and the graph view of the entity's attributes and relations.
(2) **Data curation**: We demonstrate the curation capabilities of the graph view, by changing the attributes of a selected entity. In addition, we add new business entities and delete existing ones. We then present the status view of the changes made. The operations displayed will be applied to the knowledge base in the next step of the demonstration.
(3) **History overview**: During this part of the demonstration we show the various data processing steps that have been applied to the knowledge base. As a starting point, we show the difference between the current state of the knowledge base and the knowledge base before manual curation. We then commit the changes made in the previous step to the knowledge base.
(4) **Monitoring of intermediate steps**: We present the tools for monitoring the intermediate steps of the Deduplication and Text Mining components. We first show the duplicates of a previously performed duplication run and inspect some of them. Then we show the evaluation of different blocking keys as well as the evaluation of different threshold values for the duplicate detection. For the Text Mining component we will first demonstrate the entity linking overview. This view displays both articles and their links to the corresponding knowledge base entires. We then show the evaluation tool for the different classification models used within the Text Mining component. Participants are presented with several evaluations for different classification models and their parameters.

After demonstrating the Curation Interface, we continue the demonstration by introducing ELEX. We first show that the changes made in the Curation Interface will already be accessible from ELEX. We continue the demonstration by presenting the following key features of ELEX: high-performance display and exploration capabilities of the knowledge graph, searching for single entities and filtering the knowledge graph, inspecting single entities, displaying the graph in different layouts and giving feedback to point out errors for single nodes and edges, e.g., an error in a data field or an incorrect relation.

5 CONCLUSION

We present CurEx, a modular system to integrate structured and unstructured data sources into a domain-specific knowledge base and create explorable knowledge graphs for specific domains. The system is based on scalable technologies and is therefore able to process large amounts of data, making it suitable for real-world scenarios. It consists of two major components specifically designed for integrating information from structured and unstructured data sources. Since all subcomponents are implemented as Spark jobs, they are easily replaced or extended. It provides two distinct user interfaces, each addressing the individual needs of a specific user group. As such, a data engineer can control and manage the integration process using the Curation Interface, whereas a normal end-user uses ELEX to explore the knowledge graph and submit feedback.

As future directions we focus on the improvement of the individual subcomponents. For example, we plan to replace the currently used deduplication approach with a novel approach based on neural networks. Another planned improvement is entirely replacing the fuzzy matching approach for entity linking with CohEEL.

REFERENCES

[1] Chen Chen, Behzad Golshan, Alon Y. Halevy, Wang-Chiew Tan, and AnHai Doan. 2018. BigGorilla: An Open-Source Ecosystem for Data Preparation and Integration. *IEEE Data Engineering Bulletin Issues* 41 (2018), 10–22.
[2] Toni Grütze, Gjergji Kasneci, Zhe Zuo, and Felix Naumann. 2016. CohEEL: Coherent and efficient named entity linking through random walks. *Journal of Web Semantics* 37–38 (2016), 75–89.
[3] Michael Loster, Zhe Zuo, Felix Naumann, Oliver Maspfuhl, and Dirk Thomas. 2017. Improving Company Recognition from Unstructured Text by using Dictionaries. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*. 610–619.
[4] Renee Miller. 2018. Open Data Integration. *Proceedings of the International Conference on Very Large Databases (VLDB)* 11 (2018), 2130–2139.
[5] Christopher De Sa, Alexander Ratner, Christopher Ré, Jaeho Shin, Feiran Wang, Sen Wu, and Ce Zhang. 2017. Incremental knowledge base construction using DeepDive. *Proceedings of the International Conference on Very Large Databases (VLDB)* 26 (2017), 81–105.
[6] Michael Stonebraker, Daniel Bruckner, Ihab F. Ilyas, George Beskales, Mitch Cherniack, Stanley B. Zdonik, Alexander Pagan, and Shan Xu. 2013. Data Curation at Scale: The Data Tamer System. In *Conference on Innovative Data Systems Research CIDR*.
[7] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 207–2013.
[8] Zhe Zuo, Michael Loster, Ralf Krestel Krestel, and Felix Naumann. 2017. Uncovering Business Relationships: Context-sensitive Relationship Extraction for Difficult Relationship Types. In *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen" (LWDA)*. 271–283.