Open Cultural Heritage Data in University Programming Courses

Tabea Tietz and Harald Sack

Karlsruhe Institute of Technology, Institute AIFB, Germany and
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
firstname.lastname@fiz-karlsruhe.de

Abstract. Cultural heritage data are not only an important research subject for the semantic web community, but also provide interesting material for practical programming courses in universities. In this paper, four projects created by master students at Karlsruhe Institute of Technology (KIT) show how open cultural heritage data can be used to develop creative and ambitious applications which improve the students' knowledge and experience working with semantic web technologies, linked data, natural language processing techniques and machine learning. Furthermore, challenges and lessons learned are discussed.

Keywords: education · cultural heritage · semantic web · linked data · data exploration · machine learning · natural language processing

1 Introduction

Cultural heritage data have become an important subject of research in many domains. Especially in the field of semantic web research, the possibilities of providing means to explore the growing amount of these data and the tasks for knowledge engineering and knowledge mining are numerous [7][6]. At the same time, students of semantic web (and related) courses in bachelor and master programs at universities need to be provided with interesting topics and projects to build their skills and spark their enthusiasm for the field. The idea to combine real world research problems with university-level education has been implemented widely as it not only provides students with realistic insights into the academic life but also allows tutors to integrate results obtained into their own research efforts. The presented poster paper follows this line of argumentation. In the paper, outcomes of a master student course are presented in which student groups chose open cultural heritage datasets, developed and evaluated their own web-based applications, thereby utilizing and improving their knowledge in semantic web technologies, natural language processing (NLP), and machine learning (ML). The presented results consist of an exploration framework for historical texts, a gamification approach for history lessons enriched with linked data, a content-based book recommendation system and a museum chatbot. A specialty about this particular programming course is the use of open cultural

heritage datasets made available by the Coding da Vinci initiative, which confronted the students with interesting and mostly uncharted data to explore [1]. Coding da Vinci is the first German open cultural data hackathon, founded in 2014. Its goal is to bring together cultural heritage institutions with the hacker designer community to develop ideas and prototypes for the cultural sector as well as for the public. The paper furthermore provides lessons learned from the practical seminar including challenges the student groups have dealt with.

2 Course Structure and Data

The seminar project course 'Information Service Engineering' (ISE) was held at KIT for master students of information management. As a prerequisite, all students previously attended the ISE lecture, which prepared them with foundations in linked data engineering, ML, and NLP. The goal of the project course was to work on a common research problem in groups of 3 students each and to come up with and implement a web-application using state-of-the art scientific research methods. All groups received the tasks to select datasets part of Coding da Vinci and to implement their own project idea, supported by at least 2 tutors from the teaching team. Coding da Vinci originally is a German hackathon which supports interdisciplinary work on cultural heritage data. However, the hackathon concept does not really fit the university curriculum, i.e. with a timeframe of 5 weeks for the current edition of Coding da Vinci, the obtained workload would significantly influence the students' other courses. Therefore, it was decided to decouple the Coding da Vinci competition from the 14 weeks' course work. However, the goal was nevertheless to establish a competitive environment with the student teams working on real world datasets, comparing their achieved results also with the original hackathon results. In general, the experimental ISE programming course was conducted to give proof that current research problems of the cultural heritage and digital humanities domain can successfully be integrated into the teaching curriculum by the development of semantic web based solutions.

3 Applications

Overall, 4 cultural heritage data projects have been implemented (cf. Fig. 1). All applications can be accessed on the web [2]

① **Exploring Historical Text using Word & Document Embeddings** In this project a framework is presented which combines unsupervised ML techniques and NLP on the example of historical text documents on the $19th$ century USA. Named entities are identified and extracted from semi-structured text, which is enriched with complementary information from Wikidata. Word embeddings are leveraged to enable the analysis of the text corpus, which is visualized in a web-based application. Experiments provide evidence that the method used to create the semantic representations using word and document

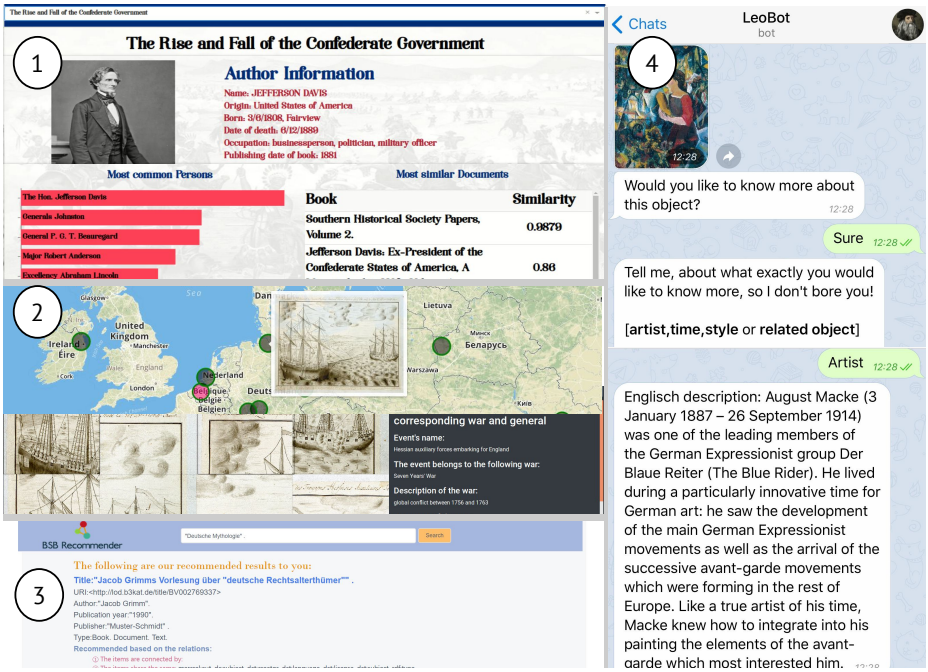

Fig. 1. Screenshots of the four presented ISE applications

vectors are especially promising for the analysis of unstructured historical English text collections, as long as the amount of text is sufficiently large to train meaningful semantic representations with the neural net.

② **Learning History through Gamification.** This project contributes to the area of improving history lessons through gamification. Users interested in studying historical wars in Europe can select from a set of pictures presented on a map. Once chosen, a puzzle game based on historical pictures of European wars opens. Upon completion, the user receives complementary information automatically generated from DBpedia and Wikidata about the event and persons involved. The maps and landscapes used are part of the Hessian State Archives [4]. An evaluation of the game involved 37 participants and revealed that the users are interested in the presented combination of solving a game and receiving further knowledge about historical events, especially in the context of history lessons in schools and museums. The main challenge for the students involved the often sparse amount of data available for historical wars and battles.

③ **Content-based Book Recommender System.** A content-based book recommender system was developed using a dataset published by the Bavarian State Library[3]. A main challenge for the students was handling the large size of the 135GB dataset consisting of more than 100 million entities. Recommendations are generated based on semantic similarities of the documents and on relatedness of specific entities in the corpus being connected to Wikidata.

④ **A Linked Data Enhanced Museum Chatbot.** In this project, a chatbot for the Städel Museum Frankfurt, Germany [5] was developed to make museum visits more interesting for younger generations. The application is based on the Telegram messenger app and lets the user converse with a chatbot to receive information about art pieces and background information about artists, places or paintings integrated using Wikidata and DBpedia.

4 Lessons Learned and Conclusion

In this paper, the outcomes of a university programming course are presented, where methods have been developed to successfully explore cultural heritage data collections using semantic web and linked data technologies, as well as NLP and ML techniques. An evaluation of the course performed by the university reveals that the students genuinely enjoyed the course. It is especially welcomed that the students implemented their own ideas using the data they have chosen themselves. The evaluation also reveals that the workload the students encountered was very high, which is a disadvantage of the groups choosing the project topics on their own, because the amount of work to be completed was always underestimated. This will be counteracted in the future by examining the datasets more closely before the seminar and pointing out possible problems with concrete solutions. The most valuable lessons learned in this course are that the open nature of the course fostered refreshing and interesting ideas and perspectives from students who are not yet part of the semantic web research community. The seminar enabled to build bridges between traditional GLAM institutions who own the data and young researchers who provide the necessary ideas on how to explore these data using scientific methods and new technologies. Last but not least, the seminar helped to increase the students' interest in semantic web technologies, because the application on cultural heritage data revealed a broader context on why these technologies are important after all.

Acknowledgement. We would like to thank all seminar students and tutors, who invested a great amount of work to make each project a successful one, and the Coding da Vinci initiative for providing the extensive amount of data.

References

1. Coding da Vinci (2014, accessed March 19,2019), `https://codingdavinci.de/`
2. Course Projects (accessed April 23,2019), `https://ise-fizkarlsruhe.github.io/CourseProjects2019`
3. Bavarian St. Library (accessed March 19,2019), `https://www.bsb-muenchen.de/`
4. Hessian St. Archives (accessed March 19,2019), `https://landesarchiv.hessen.de/`
5. Städel Museum (accessed March 19,2019), `https://www.staedelmuseum.de/en`
6. Dragoni, M., Tonelli, S., Moretti, G.: A knowledge management architecture for digital cultural heritage. JOCCH **10**(3), 15 (2017)
7. Simou, N., Chortaras, A., Stamou, G., Kollias, S.: Enriching and publishing cultural heritage as linked open data. In: Mixed Reality and Gamification for Cultural Heritage, pp. 201–223. Springer (2017)