Recommendation-Assisted Data Curation for Wikidata

Position Paper

Eva Zangerle
University of Innsbruck
eva.zangerle@uibk.ac.at

Claudia Müller-Birn
Freie Universität Berlin
clmb@inf.fu-berlin.de

ABSTRACT

The Wikidata project provides a structured knowledge base that is curated by bots and humans alike. The quality and completeness of data contained is naturally influenced by the users who enter and maintain the data in the form of items described by statements on the platform. Users who are new to the Wikidata environment and its underlying data model, but are, nonetheless, experts in their fields, are confronted with a steep learning curve when aiming to enter information on Wikidata (e.g. regarding the choice of suitable properties for creating statements). In this work, we propose a recommendation-based annotation platform where users who currently work with or on a text are supported in finding suitable Wikidata entities for data extracted from the underlying text source to ultimately feed this structured information to the Wikidata platform. Such recommendations not only support users in annotating their data according to Wikidata's terminological knowledge, but also expand the number of references on the Wikidata platform that reveal the origin of existing statements.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**; • **Information systems → Recommender systems**;

KEYWORDS

Wikidata, Annotation, Recommendation

ACM Reference Format:
Eva Zangerle and Claudia Müller-Birn. 2018. Recommendation-Assisted Data Curation for Wikidata: Position Paper. In *Proceedings of Wiki Workshop 2018*. Lyon, France. https://doi.org/10.5281/zenodo.1194790

1 INTRODUCTION

The Wikidata project aims to create a free, structured knowledge base that can be read and edited by humans and machines alike. Wikidata was launched in 2012 as a sister project of Wikipedia with the specific purpose of managing structured data represented in Wikipedia articles across all language versions [9]. Wikidata currently comprises data beyond the original Wikipedia scope and has become one of the largest and most active Wikimedia projects in terms of editing. As of January 2018, Wikidata has stored information about 42 million items and has received contributions from over 18,000 currently active contributors. More than 376 million statements describe and interrelate these entities. Statements consist of properties (e.g. date of birth or location) that have specific values which exhibit either other Wikidata items (e.g. Berlin) or values (e.g. 01/01/1970). Statements include additional contextual information that refers, for example, to the origin of the data provided [3].

Wikidata is already being used outside the Wikimedia community context, for example, in Apple's voice assistant Siri. Thus, Wikidata has recently become an increasingly important topic by measuring and increasing the quality and completeness of knowledge graphs. Ahmeti et al. [1], for example, introduced Recoin - the Relative Completeness Indicator, which computes the relative completeness of items.[1] The CoolWD tool introduced by Darari et al. [2] and recently the AMIE system by Galarraga et al. [4] provide approaches for measuring the completeness of knowledge graphs.[2] However, computing completeness scores can only serve as an indicator regarding to which extent information is complete; the actual curation of missing information is not tackled by these tools. Furthermore, these tools provide users with no assistance when adding information. Users, especially nontechnical experts, i.e. users who are domain experts in their respective domain, are crucial contributors to the Wikidata platform. As opposed to existing peer production systems such as Wikipedia, users need to have an understanding of Wikidata's underlying data model (terminological knowledge) when contributing to Wikidata [6].

The lack of such an understanding creates a natural entry barrier for new editors who aim to extend the knowledge base (e.g. when it comes to choosing suitable properties) or might lead to incorrectly inserted information. This is particularly critical, as Piscopo et al. [8] have shown that the diversity of contributors affects the quality of collaboratively built knowledge graphs positively. Thus, the inclusion of a diverse user group, especially of nontechnical experts, is a major goal for sustaining and building Wikidata's community. Wikidata currently provides users with a property suggester that aims at assisting users when adding new statements to a given Wikidata item[3].

However, this property suggester extends Wikidata's user interface and does not provide suggestions outside the Wikidata context. This is especially disadvantageous, since one of Wikidata's core strengths is the verifiability of its statements, i.e. the origin of data is documented by references. An approach to increase the number of referenced statements, on the one hand, and lowering the barriers to provide information to Wikidata, on the other hand, can be to enable nontechnical experts to provide data from their respective work contexts.

In this position paper, we propose a novel human-algorithmic approach that will allow nontechnical experts reading a given text, i.e. a research paper, to contribute statements to Wikidata despite

Wiki Workshop 2018, April 2018, Lyon, France
https://doi.org/10.5281/zenodo.1194790

[1] https://www.wikidata.org/wiki/Wikidata:Recoin
[2] http://cool-wd.inf.unibz.it/
[3] https://github.com/wikimedia/mediawiki-extensions-PropertySuggester

being unaware of or not familiar with its data model. Our contributions are as follows: (1) we present a prototype that allows users to annotate texts semantically, while being supported by recommendations that point the user to suitable properties and, ultimately, submit these structured annotations to Wikidata; (2) we discuss how such an approach of providing information to enhance the verifiability of Wikidata's knowledge graph; and (3) we elaborate on future extensions to this prototype which aim to enhance the user experience and foster human-computer collaboration.

2 RECOMMENDATION-ASSISTED DATA CURATION

We envision the proposed prototype to enable (nontechnical) experts who read or work on textual resources online to contribute to Wikidata more easily by providing them with recommendations for annotations that can be fed directly to Wikidata. For this goal, we combine two existing tools into a new prototype: *"neonion"* is a web application that provides an environment for the semantic annotation of textual resources [7]. The core of the tool is an intuitive browser-based user interface to add semantic annotations in a triple format to texts manually. The semantic annotation capabilities of neonion feature a flexible terminological knowledge model-based and contextual metadata for each annotation, such as timestamp or owner. We combine neonion with a recommender system that assists users by proposing properties that are suitable for the text string the user is currently annotating. We rely on the *Snoopy system* [5] for the recommender system, as it has been shown to work well for the recommendation of properties on the Wikidata platform [10]. Snoopy is a recommender system that assists users in entering and editing information in semi-structured information systems (i.e. systems that are based on the triple-format, as used on Wikidata). In this particular case, Snoopy is used to recommend properties that are suitable for adding additional annotations to the text based on a seed set of annotations. Thus, the user needs to know only a small part of Wikidata's terminological knowledge to create some first (seed) annotations. By computing the co-occurrence of properties on items, we can provide recommendations for further properties that might be used when working on a given text. These annotations are internally stored in neonion in a RDF-based triple format and can in a further step, be fed into Wikidata with a reference that links to the data origin, for example, the URL of an online resource or the DOI of a research article.

3 FUTURE DIRECTIONS AND CONCLUSION

Based on the current prototype, we aim to expand our research in the following directions:

(i) Text Context: In the current prototype, recommendations are based solely on the co-occurrence of properties on data items. The actual text underlying the annotation and curation process has not yet been considered. Recommendations can further be contextualized and improved by utilizing, for example, named entity recognition and its subsequent matching with Wikidata items and properties. Extracting the topic of a given textual resource can also help alleviate the so-called cold start problem, where the recommender system does not have sufficient data about a new entity or user to compute relevant recommendations.

(ii) Personalization: We do not provide personalization for users in the current prototype. The computation of recommendations is only based on user-agnostic data. In future, we aim to extend the current prototype such that, for example, the extent to which recommendations are shown can be based on the user's experience to distinguish between novice users and heavy users, who naturally require a different level of support.

(iii) Transparency and Feedback: We aim to lay a particular emphasis on making the recommender system transparent to the user by providing them with explanations and justifications [10]. This not only allows the user to understand the rationale behind choosing certain properties in a given context, but also learn and get an understanding of Wikidata's data model, which contributes directly to the quality of future edits. In turn, we aim to gather feedback from the user (implicitly and/or explicitly) about the perceived usefulness of certain items to optimize the recommender algorithm.

(iv) Evaluation: We aim to evaluate the proposed prototype. We will rely a questionnaire suggested by Pu et al. [9] on user-centric aspects of the recommender system to get an understanding of recommendation quality and usability. Naturally, we aim to perform a user study to capture user experience with the system and evaluate the impact of the system on the quality of the statements created.

All these measures will contribute to an increased data quality and completeness on Wikidata, as (i) the system allows nontechnical experts to enter data into Wikidata directly from original resources, (ii) the user is educated by providing justifications for recommendations and, hence, more informed when entering data, and (iii) recommender systems that incorporate not only data from Wikidata, but also from the textual resource yield more precise recommendations and, thus, higher quality and completeness on the Wikidata platform.

REFERENCES

[1] Albin Ahmeti, Simon Razniewski, and Axel Polleres. 2017. *Assessing the Completeness of Entities in Knowledge Bases.* Springer, 7–11.

[2] Fariz Darari, Raditiyo Eko Prasojo, Simon Razniewski, and Werner Nutt. 2017. COOL-WD: A Completeness Tool for Wikidata. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks.*

[3] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the Linked Data Web. *Semantic Web Conference* (2014).

[4] Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M Suchanek. 2017. Predicting completeness in knowledge bases. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining.* ACM, 375–383.

[5] Wolfgang Gassler, Eva Zangerle, and Günther Specht. 2011. The Snoopy Concept: Fighting heterogeneity in semistructured and collaborative information systems by using recommendations. In *2011 International Conf. on Collaboration Technologies and Systems (CTS).* 61–68.

[6] Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. 2015. Peer-production System or Collaborative Ontology Engineering Effort: What is Wikidata?. In *Proc. of the 11th International Symposium on Open Collaboration.* ACM, New York, NY, USA, 20:1–20:10.

[7] Claudia Müller-Birn, Tina Klüwer, André Breitenfeld, Alexa Schlegel, and Lukas Benedix. 2015. Neonion: Combining Human and Machine Intelligence. In *Proc. of the 18th ACM Conf. Companion on CSCW & Social Computing.* ACM, 223–226.

[8] Alessandro Piscopo, Chris Phethean, and Elena Simperl. 2017. What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *International Conference on Social Informatics.* Springer, 305–322.

[9] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[10] Eva Zangerle, Wolfgang Gassler, Stefan Steinhauser, and Günther Specht. 2016. An Empirical Evaluation of Property Recommender Systems for Wikidata and Collaborative Knowledge Bases. In *Proc. of the 12th International Symposium on Open Collaboration (OpenSym '16).* ACM.