Wikibase as an Infrastructure for Community Documents: The Example of the Disability Wiki Platform

Kushagra Singh Bisen[1,5], Sara Assefa Alemayehu[1], Pierre Maret[1,4,*], Alexandra Creighton[2], Rachel Gorman[2], Bushra Kundi[2], Thumeka Mgwgwi[3], Fabrice Muhlenbach[1], Serban Dinca-Panaitescu[2], Dennis Diefenbach[1,4], Kunpeng Guo[1,4] and Christo El Morr[2]

[1]*Université Jean Monnet Saint Etienne, France*
[2]*York University, School of Health Policy and Management, Canada*
[3]*School of Gender, Sexuality and Women's Studies, York University, Canada*
[4]*The QA Company, France*
[5]*IDLab, Ghent University - imec, Belgium*

Abstract

The questions that can arise from the users searching for domain-specific answers can hardly be answered with Web search engines. A corpus-dedicated platform is generally needed. In this paper, we present how the Wikibase environment can be employed to make documents searchable efficiently. We use this environment for the Disability Wiki platform. Search for information can be both on the metadata as well as on the content of the documents.

Keywords

Wikibase, Question Answering, Document Corpus, Domain-Specific Question Answering

1. Introduction

The need for information for users is achieved by search techniques over the web. There are at least 2.46 billion pages indexed on the World Wide Web. Communities express their need to search for available domain-specific data and documents. This is the case for the disability and human rights community, where individuals, advocacy groups, as well as decision-makers seek better information access[1][2]. Searching for disability information on the Web returns information which is far too general for advocacy. Therefore, there is a need for online search engines that can target data and documents in this domain, and that can answer the queries of stakeholders. This demo paper shows the result of the Disability Wiki project which aims at providing a platform for improving the access to information in this field. We present how the Wikibase environment[1] is used as a platform for domain information in the form

SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, September 13-15, 2022, Vienna, Austria
*Corresponding author.
0000-0003-0950-6043 (K. S. Bisen)

[1]https://wikiba.se/, Wikibase is the software behind Wikidata

of data and documents (structured and unstructured data). We further demonstrate how a question-answering system can be applied to provide useful access to this corpus.

2. Related Work

Question Answering over documents is an open research topic. There are existing solutions employing computer vision to query the documents [3] [4] and solutions to express the content of a document as a knowledge graph which can be searched [5]. The paper [6] makes a domain-specific knowledge graph from the documents in biology, but it is not extendable to other domains. Anteghini et al. [7] represents community information in Wikibase and proposes an ontology which is only domain-specific to slavery documents. As far as we know, no proposal has been made yet to use the Wikibase infrastructure to model document meta-data, and to query simultaneously both the meta-data and the document contents.

3. The Disability Wiki platform

The overall Disability Wiki platform consists roughly of a Wikibase instance, a website to upload documents and to question the platform, and a question-answering engine.

3.1. Representing documents in Wikibase

The Wikibase environment uses an RDF-compliant data modelling approach. Therefore, we propose an ontology to represent the metadata related to documents (Figure 1). In this ontology the concepts *author*, *institution*, *upload date*, *topic* (etc.) are proposed, as well as the concept of *content box* which corresponds to any part of a document. The ontology uses the data types offered by Wikibase[2]

3.2. Setting up a Wikibase instance

An instance of a Wikibase is created and initiated with items and properties[3]. These items and properties correspond respectively to the concepts and the relations of the ontology. Properties represent relations between the document and the metadata in the Disability Wikibase. We employ the glossary which is generated manually by domain-experts containing disability-specific topics and their aliases and is added to Wikibase as items.

3.3. Question-answering engine

The Disability Wiki platform implements the QAnswer question-answering engine. It can query over both knowledge graphs[8] and documents[9] to provide an answer with the maximum confidence score. There is a fallback to elastic search[10] in case the confidence level is below a certain defined threshold. A domain expert in disability studies can provide feedback to validate the answer to the question, which is used to train the QA engine further.

[2]https://www.wikidata.org/wiki/Help:Data_type
[3]https://disabilitywiki.univ-st-etienne.fr/wiki/The_Disability_Wikibase

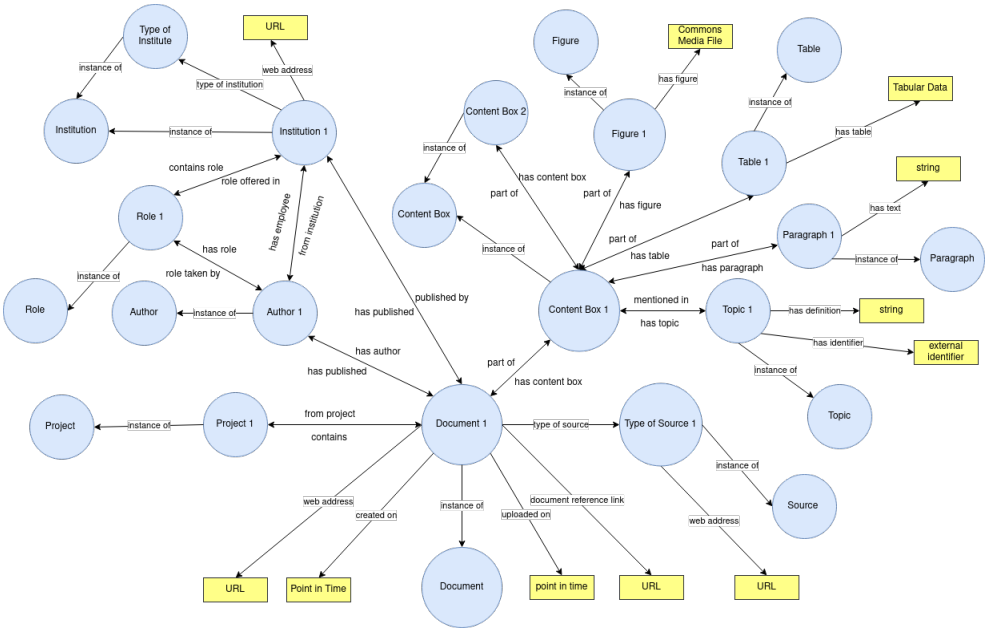

Figure 1: Ontology to represent the documents in Wikibase

3.4. Disability Wiki Website

The Disability Wiki platform appears to the users as a website[4]. Authorised users can log in and upload documents. Document meta data are stored in the Wikibase. Any web user can visit the website to query the platform, i.e. write text questions sent over the Wikibase and the uploaded documents.

4. System implementation

4.1. System description

The question answering system utilises two datasets, which are the RDF data and documents. The RDF data is generated from the Wikibase and dumped to QAnswer KG[8] service for indexing. The domain documents are uploaded as PDF in a separate dataset.

4.2. Querying the Disability Wiki platform

The system can answer different types of questions on the platform,

- Searching for single words and definitions. The answers are brought from Wikibase for it matches knowledge expressed by domain experts in their glossary. For example, "What is the definition of Disability?" and "rural area definition"

[4]http://disabilityrightsweb.univ-st-etienne.fr/

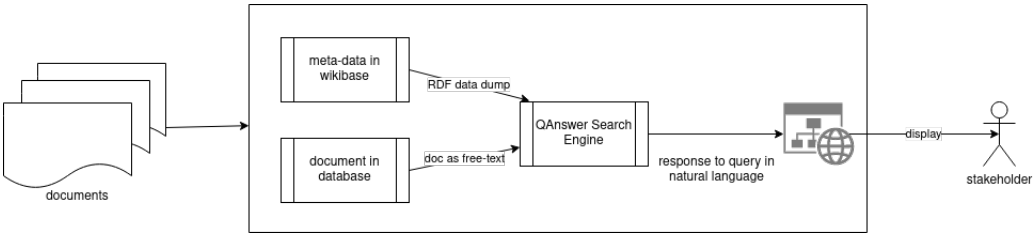

Figure 2: Description of the system employed for Question Answering over domain specific corpus

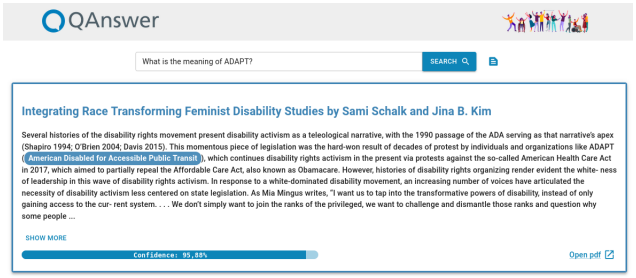

Figure 3: Result of the query: "What is the meaning of ADAPT?". The screenshot shows the response which was retrieved employing QAnswer on free-text of the documents.

- Querying metadata of a document. For example, "give me text related to discrimination and addiction" or "subject of crpd article 11" (See figure 4)
- Getting answers from the documents. The system can give answers with high confidence like for example: "What is the meaning of ADAPT?" (see Figure 3) or it can lead to an exploratory search with answers having comparable confidence, for instance: "What is ableism?"
- Searching for keywords in the document corpus. For example, "ableism mortality" will be answered back with documents containing these words.

Demonstration

We provide the link to the current version of the system at:

http://demo-disabilityrightsweb.univ-st-etienne.fr/

Moreover, by clicking on the above figure or examples you will be redirected to the demo.

5. Conclusion

With this demo paper, we demonstrate how the Wikibase environment can be used to model a document's metadata and how question answering over community document corpora can be done by combining the document's metadata in RDF(Resource Description Framework) with the content of the document in free-text. We also demonstrate the different types of questions that can arise on searching over document-specific corpora, and how our system can answer these types of questions.

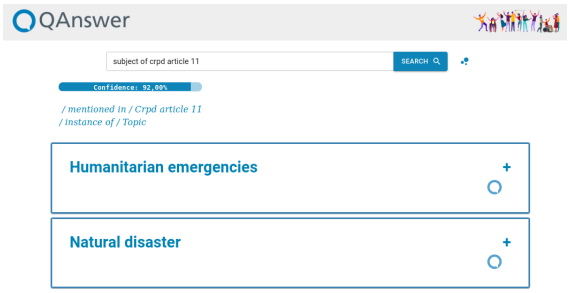

Figure 4: Result of the query: "subject of crpd article 11". The screenshot shows the response which was retrieved employing QAnswer on the Disability Knowledge Graph.

References

[1] M. Loeb, Disability statistics: an integral but missing (and misunderstood) component of development work, Nordic journal of human rights 31 (2013) 306–324. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4766593/.

[2] O. Abualghaib, N. Groce, N. Simeu, M. T. Carew, D. Mont, Making visible the invisible: Why disability-disaggregated data is vital to "leave no-one behind", Sustainability 11 (2019). URL: https://www.mdpi.com/2071-1050/11/11/3091. doi:10.3390/su11113091.

[3] O. Tüselmann, F. Müller, F. Wolf, G. A. Fink, Recognition-free question answering on handwritten document collections, 2022. URL: https://arxiv.org/abs/2202.06080. doi:10.48550/ARXIV.2202.06080.

[4] Y. Ding, Z. Huang, R. Wang, Y. Zhang, X. Chen, Y. Ma, H. Chung, S. C. Han, V-doc: Visual questions answers with documents, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21492–21498.

[5] B. R. Andrus, Y. Nasiri, S. Cui, B. Cullen, N. Fulda, Enhanced story comprehension for large language models through dynamic document-based knowledge graphs, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 10436–10444. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21286. doi:10.1609/aaai.v36i10.21286.

[6] M. Anteghini, J. D'Souza, V. A. P. M. d. Santos, S. Auer, Easy semantification of bioassays, 2021. URL: https://arxiv.org/abs/2111.15182. doi:10.48550/ARXIV.2111.15182.

[7] C. Shimizu, P. Hitzler, Q. Hirt, D. Rehberger, S. G. Estrecha, C. Foley, A. M. Sheill, W. Hawthorne, J. Mixter, E. Watrall, et al., The enslaved ontology: Peoples of the historic slave trade, Journal of Web Semantics 63 (2020) 100567.

[8] D. Diefenbach, J. Giménez-García, A. Both, K. Singh, P. Maret, Qanswer kg: Designing a portable question answering system over rdf data, in: A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, M. Cochez (Eds.), The Semantic Web, Springer International Publishing, Cham, 2020, pp. 429–445.

[9] K. Guo, C. Defretiere, D. Diefenbach, C. Gravier, A. Gourru, Qanswer: Towards question answering search over websites, ACM, 2022.

[10] C. Gormley, Z. Tong, Elasticsearch: the definitive guide: a distributed real-time search and analytics engine, " O'Reilly Media, Inc.", 2015.