Scientific Research Publishing

Journal of Computer and Communications, 2023, 11, 151-160
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Creating Bengali Freebase Using Wikidata

Rukaiya Habib[1], Mahmuda Ferdous[2], Md Musfique Anwar[1]

[1]Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, University of Development Alternative, Dhaka, Bangladesh
Email: rukaiya.habib45@gmail.com, ferdousmahmuda@yahoo.com, manwar@juniv.edu

How to cite this paper: Habib, R., Ferdous, M. and Anwar, M.M. (2023) Creating Bengali Freebase Using Wikidata. *Journal of Computer and Communications*, **11**, 151-160.
https://doi.org/10.4236/jcc.2023.115011

Received: April 20, 2023
Accepted: May 27, 2023
Published: May 30, 2023

Abstract

Freebase is a large collaborative knowledge base and database of general, structured information for public use. Its structured data had been harvested from many sources, including individual, user-submitted wiki contributions. Its aim is to create a global resource so that people (and machines) can access common information more effectively which is mostly available in English. In this research work, we have tried to build the technique of creating the Freebase for Bengali language. Today the number of Bengali articles on the internet is growing day by day. So it has become a necessary to have a structured data store in Bengali. It consists of different types of concepts (topics) and relationships between those topics. These include different types of areas like popular culture (e.g. films, music, books, sports, television), location information (restaurants, geolocations, businesses), scholarly information (linguistics, biology, astronomy), birth place of (poets, politicians, actor, actress) and general knowledge (Wikipedia). It will be much more helpful for relation extraction or any kind of Natural Language Processing (NLP) works on Bengali language. In this work, we identified the technique of creating the Bengali Freebase and made a collection of Bengali data. We applied *SPARQL* query language to extract information from natural language (Bengali) documents such as Wikidata which is typically in RDF (Resource Description Format) triple format.

Keywords

Knowledge-Base, Structured Data, NLP, RDF

1. Introduction

In reality, we can see that the World Wide Web does not provide all the knowledge in consistent and uniformly structured way. For that reason, access of the Web is difficult for the purpose of sophisticated data analysis, search and organ-

ization [1]. Some of the specific problems are mentioned below:

- Structured information is "trapped" in unstructured documents which are unable to be easily read by automated processes.
- Multiple, unconnected representations are available of the same real-world entity.
- Often, a lack of separation of meaning from presentation is there. Most unstructured and many structured data sources mix semantics with display information.
- Often, a heterogeneous representation of information is existed across data sources and even within a single data source across time.
- Sometimes Web-based information is presented without the explicit context of other, potentially helpful information sources. This is a condition arising from the one-way nature of Web links.

So to solve the above problem, a useful knowledge base Freebase has been built in English. We can create a Bengali Freebase by making query through Wikidata. Wikidata is a document-oriented database, focused on items which represent topics, concepts or objects. Wikidata has RDF triple format. Actually, a mapping has been built from Freebase properties to Wikidata. Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. We can easily get our required seed tuple automatically rather than giving input manually by using Wikidata query service [2]. In this way, large scale information integration, entity extraction and data reconciliation problems will be solved by automatically perform structuring, extraction and reconciliation tasks over large, messy, already existing datasets [1].

Most existing works in Bangla language processing focused on research areas such as machine translation [3] [4] [5], reading compression [6] [7], sentiment analysis etc. Few research works focused on information retrieval (such as relation extraction [8] [9]) where the authors tried to use knowledge base.

2. Related Work

The value of Wikipedia's data has long been obvious, with many efforts to use it. The Wikidata approach is to crowd-source data acquisition, allowing a global community to edit the data. This extends the traditional wiki approach of allowing users to edit a website. Wiki is a Hawaiian word for fast; Ward Cunningham, who created the first wiki in 1995, used it to emphasize that his website could be changed quickly [10].

Some existing popular such system is Semantic MediaWiki, or SMW [11], which extends MediaWiki, the software used to run Wikipedia [12], with data-management capabilities. SMW was originally proposed for Wikipedia but was quickly used on hundreds of other websites as well. Unlike Wikidata, SMW manages data as part of its textual content, thus hindering creation of a multilingual, single knowledgebase supporting all Wikimedia projects. Moreover, the data model of Wikidata is more elaborate than that of SMW, allowing users to

capture more complex information. In spite of these differences, SMW has had a great influence on Wikidata, and the two projects share code for common tasks.

Other examples of free knowledgebase projects are OpenCyc and Freebase. OpenCyc is the free part of Cyc [13], which aims for a much more comprehensive and expressive representation of knowledge than Wikidata. OpenCyc is released under a free license and available to the public, but, unlike Wikidata, is not editable by the public. Freebase, acquired by Google in 2010, is an online platform that allows communities to manage structured data [14]. Objects in Freebase are classified by types that prescribe what kind of data an object can have; for example, Freebase classifies Einstein as a "musical artist" since it would otherwise not be possible to refer to recordings of his speeches. Wikidata supports the use of arbitrary properties on all objects. Other differences from Wikidata are related to multi-language support, source information, and the proprietary software used to run the site. The latter is critical for Wikipedia, which is committed to running on a fully open source software stack to allow all to fork, or copy and create one's own version of the project.

3. Freebase

Freebase is designed to facilitate high "collaborative density" among its users in the organization, representation and integration of large, diverse data sets. Freebase has some own characteristics. They are given below.

3.1. A Huge Data Store

This is a scalable, tuple store with some built-in query planning and optimization capabilities which allow deep, naively constructed queries to be satisfied quickly. This assists users in query optimization in building high performing systems.

3.2. A Large Data Object Store (LOB)

This is a store of large data objects such as text documents. LOB objects are indexed and annotated in the store.

3.3. A Substantial Seed Data Set

An emphasis has been placed on the early seeding of Freebase with data sets of interest to the general population rather than those that are highly esoteric and specialized. This hopefully will result in greater heterogeneity of structure and content that is more representative of the world's sum of general knowledge. It consists of different types of concepts (topics) and relationships between those topics. Topics are:

- popular culture (e.g. films, music, books, sports, television);
- location information (restaurants, geolocations, businesses);
- scholarly information (linguistics, biology, astronomy);
- general knowledge (Wikipedia).

While this data is already useful, we are making efforts for it to grow quickly over time in both quantity and density of relationships.

4. Wikidata Service

Wikidata is a website that belongs to the Wikimedia family of websites. The most famous site in that family is Wikipedia. Data from Wikidata is available in RDF dumps. Actually RDF stands for Resource Description Framework which is a general method for describing data by defining relationships between data objects and it allows data integration from multiple sources. RDF has triple format which is a set of three entities that codifies a statement about semantic data in the form of subject-predicate-object expressions [15].

Wikidata is a place to store structured data in many languages. The basic entity in Wikidata is an item. An item can be a thing, a place, a person, an idea or anything else. The subject of each Wikipedia article corresponds to a Wikidata item but the definition of a Wikidata item is more flexible and inclusive and there are many items about which there are no Wikipedia articles. Wikidata has identifier numbers for entities and properties.

4.1. Entity Identifier Number

As Wikidata treats all languages in the same way, items don't have names, but generic identifiers. Each identifier is the letter Q that is followed by a number. For example, the item about the capital of Japan is called neither "Tokyo" nor "anything" but Q1490. But to give it a human-readable name, each item has a list of labels in each language associated with it. So we'll see that the English (en) label at Q1490 is "Tokyo", also has corresponding word for the Japanese (ja) label, the Bengali (bn) label and so on.

4.2. Property Identifier Number

Every item has a list of statements associated with it. Each statement has a "property" and a "value". There is a long list of possible properties. Like items, properties have generic identifiers, but they begin with the letter P and not Q. For example, the property to indicate the country is P17, and it has the label "country" in English. The value of P17 (country) for Q1490 (Tokyo) is Q17 (Japan, etc.). There are many other statements about Tokyo: flag (Q20900820, which points to an image at File:Flag of Tokyo), population (13,686,371), mayor (Q389617) and many others.

This way of organizing the information in a structured way makes it easy for computers to process. For an example,

Rukaiya is the ***citizen of*** Bangladesh.

- Here, Rukaiya is a person, an entity which is called "human". In Wikidata human entity has an identifier no Q5.
- Here, citizen of is the relation between two entites which is called a property and this property name is "country of citizenship" and has an identifier no

P27.

- Here, Bangladesh is a country which is an entity. In Wikidata Bangladesh entity has an identifier no Q902.

5. SPARQL Query Process

It is a necessary to extract information from complaints, either scraped from the Web or received directly from the client for many companies nowadays. The aim is to find inside them some actionable knowledge. To this purpose, verbal phrases must be analyzed, as many complaints refer to actions improperly performed. The Semantic Roles of the actions (who did what to whom) and the Named Entities involved need to be extracted. Moreover, for the correct interpretation of the claims, the software should be able to deal with some background knowledge (for example, a product's ontology). Although there are already many libraries and out of the shelf tools that allow tackling these problems singularly, it may be hard to find one that includes all the needed tasks. There is a query language, SPARQL to extract information from natural language documents, pre-annotated with NLP information. A query language is much easier. Moreover, the adoption of the SPARQL syntax allows to seamlessly mix, inside the same query, NLP patterns with traditional RDF, simplifying the integration with Semantic Web technologies [16].

SPARQL stands for SPARQL Protocol and RDF Query Language. It is an RDF query language and able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. It was made a standard by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium. SPARQL allows a query to consist of triple patterns. SPARQL allows users to write queries against what can loosely be called "key-value" data [17]. Some important key points have been mentioned below.

- "?" sign has been used before a variable.
- In SPARQL query, by the **SELECT** word, it will return the result by using variables.
- We use **WHERE** clause to write the main query code. Here, the query is in triple format.
- Here, for using the property identifier, we use **"wdt"** keyword before it like wdt: P19 (place of birth) and for entity identifier we use **"wd"** keyword like wd: Q902 (Bangladesh).
- We can find the identifier number from the Wikidata search box by typing the required entity or property there. Besides this, while using Wikidata online query service, we can use control and spacebar simultaneously and type it there.
- IN SPARQL, for all languages the property has the same identifier number and same as for entity. Then it will be labeled with the required language.
- Here, we return our query output in Bengali. For which result there has been an available Bengali Wikipedia, it will be returned in Bengali.

- If it is not labeled with the required language, then identifer number will be returned as the query result.
- The query result can be downloaded in different format like JSON, CSV, TSV.

6. Methodology

We have followed the given procedures:

- We can retrieve data according to our need by making query and for this, we have an online query service engine known as Wikidata query service. The URL of online query service is https://query.wikidata.org/.
- SPARQL is a standard query language technology which is endorsed by the World Wide Web consortium for querying any linked data information source. For making a query, we have to understand the SPARQL query.

Example Section

We have added an example of a query of online query service. The query is about that **we like to find out all the poets who are the citizens of Bangladesh**.

SPARQL query code:

SELECT? item? item Label? occupation Label? citizenship Label WHERE

{

?item wdt:P31 wd:Q5.

?item wdt:P106?occupation.

?item wdt:P27?citizenship. FILTER (?citizenship=wd:Q902). FILTER (?occupation=wd:Q49757).

SERVICE wikibase:label { bd:serviceParam wikibase:language "bn".}

}

Below, we explain the query:

- Here, P31 is the property "instance of" and Q5 is the entity which is "human".
- P106 is id of property of "occupation" and Q49757 is the id of "poets".
- Here, P27 is the "country of citizenship" property and it is filtered by the id of Bangladesh which is Q902.

Figure 1 and Figure 2 show sample SPARQL Query to Wikidata Service with sample outcome.

7. Experiment Results

Here, in the example section, we have added a query. The output contains 103 poets' names who have the citizenship of Bangladesh. So we can see that we can easily identify the poets' names. Thus, by making different queries, we can build a large collection of data which will be much more helpful for the researchers while working with Bengali language. A small portion of output result is given in Figure 3.

So we can say that Wikidata is scalable, tuple store with the help of some query planning and optimization capability. We can find out results on different topics which will be much more beneficial.

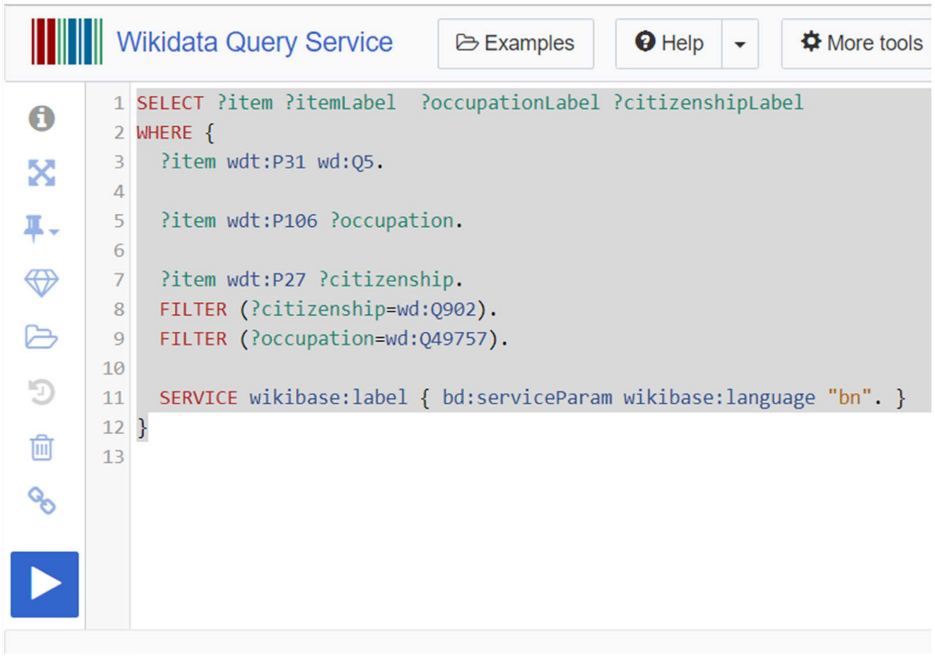

Figure 1. Sample SPARQL Query to Wikidata Service.

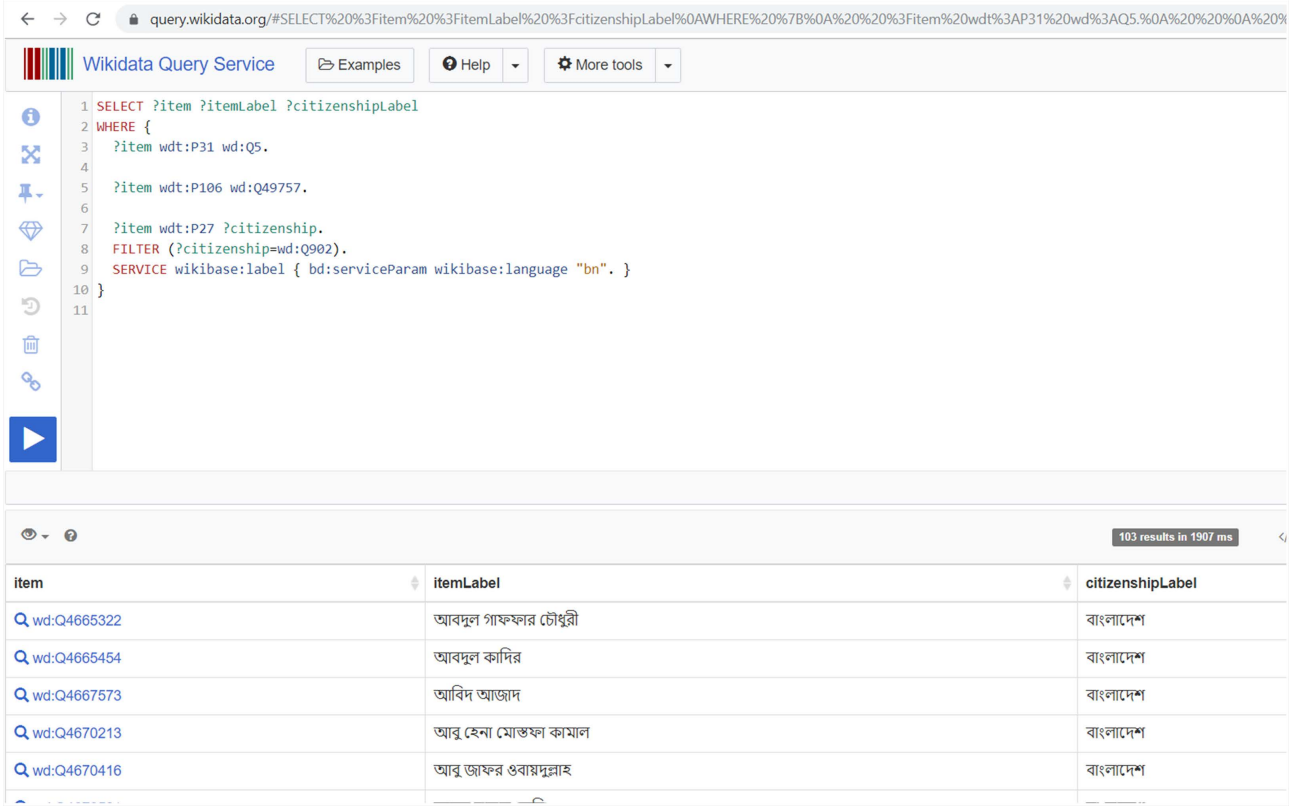

Figure 2. Sample SPARQL Query to Wikidata Service with sample outcome.

item	itemLabel	occupationLabel	citizenshipLabel
http://www.wikidata.org/entity/Q4665322	আবদুল গাফফার চৌধুরী	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4665454	আবদুল কাদির	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4667573	আবিদ আজাদ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4670213	আবু হেনা মোস্তফা কামাল	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4670416	আবু জাফর ওবায়দুল্লাহ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4670581	আবুল হাসান (কবি)	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q5277095	দিলওয়ার	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q5299406	শামীম আজাদ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q5436352	ফররুখ আহমদ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q5529165	গাজীউল হক	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q3031309	সুফিয়া কামাল	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q3143161	হুমায়ুন আজাদ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q3348651	আল মাহমুদ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q3350219	আলাউদ্দিন আল আজাদ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4696034	আহমদ ছফা	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4696347	আহসান হাবীব (কবি)	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4802171	অরুণাভ সরকার	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4854476	বন্দে আলী মিয়া	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4890119	বেনজির আহমেদ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q4913677	বিমল গুহ	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q263503	সৈয়দ আলী আহসান	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q319578	শ্রী চিন্ময়	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q334818	কাজী নজরুল ইসলাম	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q734564	শামসুর রাহমান	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q1683785	জসীম উদ্দীন	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q24572770	জাহানারা আরজু	কবি	বাংলাদেশ
http://www.wikidata.org/entity/Q24893168	হাবীবুল্লাহ সিরাজী	কবি	বাংলাদেশ

Figure 3. This is the portion of the output of the poets who have the citizenship of Bangladesh.

8. Conclusion

Freebase is a large collection of structured data. It is available in English. In this work, we have tried to find out the technique of creating Freebase for Bengali language which will be much more helpful for further research work like in Bengali language processing, where the researchers need to get a seed tuple. This research work demonstrated the technique of how to make queries for their required result using Wikidata service rather than making a database by giving input manually which is very time-consuming task. Researchers in areas such as entity extraction and reconciliation, data mining, Semantic Web, information retrieval, ontology creation and analysis can use this technique to support their research works.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bollacker, K., Tufts, P., Pierce, T. and Cook, R. (2007) A Platform for Scalable, Collaborative, Structured Information Integration. *Workshop on Information Integration on the Web* (*IIWeb*-07), Vancouver, 23 July 2007, 22-27.

[2] Hernández, D., Hogan, A. and Krötzsch, M. (2015) Reifying RDF: What Works Well with Wikidata? *Proceedings of the* 11*th International Workshop on Scalable Semantic Web Knowledge Base Systems*, Bethlehem, 32-47.

[3] Anwar, M.M., Zabed Anwar, M. and Al-Amin Bhuiyan, M. (2009) Syntax Analysis and Machine Translation of Bangla Sentences. *International Journal of Computer Science and Network Security*, **9**, 317-326.

[4] Ashrafi, S.S., Kabir, M.H., Anwar, M.M. and Noman, A.K.M. (2013) English to Bangla Machine Translation System Using Context-Free Grammars. *International Journal of Computer Science Issues*, **10**, 144-153.

[5] Anwar, M.M. (2018) Bangla to English Machine Translation Using Fuzzy Logic. *International Journal of Computer Science and Information Security*, **16**, 156-165.

[6] Aurpa, T.T., Rifat, R.K., Ahmed, M.S., Anwar, M.M. and Ali, A.B.M.S. (2022) Reading Comprehension Based Question Answering System in Bangla Language with Transformer-Based Learning. *Heliyon*, **8**, E11052. https://doi.org/10.1016/j.heliyon.2022.e11052

[7] Aurpa, T.T., Ahmed, M.S., Rifat, R.K., Anwar, M.M. and Ali, A.B.M.S. (2023) UDDIPOK: A Reading Comprehension Based Question Answering Dataset in Bangla Language. *Data in Brief*, **47**, Article ID: 108933. https://doi.org/10.1016/j.dib.2023.108933

[8] Habib, R. and Anwar, M.M. (2020) Finding Out Noisy Patterns for Relation Extraction of Bangla Sentences. *International Journal on Natural Language Computing*, **9**, 9-20. https://doi.org/10.5121/ijnlc.2020.9102

[9] Mahfuz, T., Suha, T.F. and Anwar, M.M. (2020) Reducing Wrong Labels Using Conflict Score in Distant Supervision for Relation Extraction in Bangla Language. *IEEE Asia-Pacific Conference on Computer Science and Data Engineering*, Gold Coas, 16-18 December 2020, 1-6. https://doi.org/10.1109/CSDE50874.2020.9411604

[10] Leuf, B. and Cunningham, W. (2001) The Wiki Way: Quick Collaboration on the Web. Addison-Wesley, Boston.

[11] Krötzsch, M., Vrandčić, D., Völkel, M., Haller, H. and Studer, R. (2007) Semantic Wikipedia. *Journal of Web Semantics*, **5**, 251-261. https://doi.org/10.1016/j.websem.2007.09.001

[12] Barrett, D.J. (2008) MediaWiki. O'Reilly Media, Inc., Sebastopol.

[13] Lenat, D.B. and Guha, R.V. (1989) Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Boston.

[14] Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. (2008) Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. Proceedings of the ACM SIGMOD International Conference on Management of Data, Vancouver, 9-12 June 2008, 1247-1250. https://doi.org/10.1145/1376616.1376746

[15] Hernández, D., Hogan, A., Riveros, C., Rojas, C. and Zerega, E. (2016) Querying Wikidata: Comparing Sparql, Relational and Graph Databases. *The Semantic Web-ISWC* 2016: 15*th International Semantic Web Conference*, Kobe, 17-21 October 2016, 88-103. https://doi.org/10.1007/978-3-319-46547-0_10

[16] Quintavalle, B. and Orlando, S. (2019) SPARQL/T, A Query Language with SPARQL's Syntax for Semantic Mining of Textual Complaints. 10t*h Italian Infor-*

mation Retrieval Workshop (*IIR*), Padova, 16-18 September 2019, 32-37.

[17] Weise, M., Lohmann, S. and Haag, F. (2016) Ld-Vowl: Extracting and Visualizing Schema Information for Linked Data. 2*nd International Workshop on Visualization and Interaction for Ontologies and Linked Data*, Kobe, 17-18 October 2016, 120-127.