Received: 29 August 2022 | Revised: 12 November 2022 | Accepted: 12 December 2022

DOI: 10.1111/exsy.13223

ORIGINAL ARTICLE

Expert Systems WILEY

# Structured knowledge creation for Urdu language: A DBpedia approach

Shanza Rasham[1] | Habib Ullah Khan[2] | Fahad Maqbool[1] | Saad Razzaq[1] | Sajid Anwar[3] | Muhammad Ilyas[1]

[1]Department of Computer Science & IT, University of Sargodha, Sargodha, Pakistan

[2]Department of Accounting and Information Systems, Qatar University, Doha, Qatar

[3]Center of Excellence in IT, Institute of Management Sciences, Peshawar, Pakistan

**Correspondence**
Muhammad Ilyas, Department of Computer Science & IT, University of Sargodha, Sargodha, Pakistan.
Email: muhammad.ilyas@uos.edu.pk

## Abstract

Wikipedia information is extracted by DBpedia and linked to other web resources as Linked Open Data, which is an important contribution to the field of semantics. As part of its internationalisation endeavour, DBpedia now has 20 language chapters that have been mapped to it; nonetheless, there have been very few attempts from Urdu. This article outlines the procedures and highlights the efforts put forward as the first contribution to the manual creation of Urdu mappings with DBpedia Ontology classes. Our approach led to an increase in the number of mapped infoboxes, thus enhancing the DBpedia. The mapping procedure is broken down into two parts. The infobox template is first mapped to the DBpedia ontology's relevant class, and then the attributes of the infobox are mapped to the properties of that class. In addition, alongside other mapped languages, Urdu labels are included to the description of Ontology classes. We have covered around a thousand properties and attributes of Urdu with English DBpedia Ontology on DBpedia mapping server.

KEYWORDS

DBpedia, infobox, knowledge graph, RDF

## 1 | INTRODUCTION

In order to organise and extract content from Wikipedia, the DBpedia project started as a community effort. A knowledge graph is used to present and make accessible the structured knowledge on DBpedia. A knowledge graph is a kind of data set that organises data into a machine-readable structure. This structured data not only makes the data machine readable but also make it machine understandable. DBpedia was initially confined only to the English edition of Wikipedia, but later it integrated information from multiple linguistic versions of Wikipedia via cross-language links Morsey et al. (2012). The DBpedia knowledge base covers various domains like famous individuals, places, companies, institutes, and much more. It offers complex and advanced query search utilising the SPARQL query language, in contrast to Wikipedia. In initial editions of DBpedia, it used to extract structured data and all updates from Wikipedia dumps by the end of every month Morsey et al. (2012). Modifications made to Wikipedia were not immediately recorded and mapped to DBpedia. This leads to inconsistency in the information given to the user on both ends until DBpedia completed its dump extraction. This issue was later addressed by using the continuous fetching of updates from Wikipedia to DBpedia in real time. For example, if any person's personal information is updated in Wikipedia, such as an actor's new project, this will be updated in DBpedia's database as well. Currently DBpedia automatically updates information whenever there is a change in Wikipedia.

In web 1.0 and web 2.0 the purpose is to provide user information in a precise, dynamic and interactive manner. Later on, in web 3.0 the demand is to have a machine understandable data that helps in making web applications as an intelligent application. Virtual assistance (VA) is one

https://doi.org/10.1111/exsy.13223

of the example of intelligent web application. VA assist us in our daily routine and in making the weekly or monthly planners based on events added in calendar. Availability of structured data in the form of DBpedia helps in the development of web 3.0 based applications. The DBpedia also preserves Resource Description Framework (RDF) linkages from Wikipedia's extracted content to other external data resources. This helps in linking DBpedia database to other linked data sets. Machine-decipherable search engines, such as the LodLive browser, can crawl semantics data via rdfs. Users ask a variety of inquiries, and the results are generated by following the information through rdfs linkages. DBpedia employs the Foaf (friend of a friend) method of ontology to obtain information from Wikipedia about a person, his friends, relatives, locations, interests, or anything else related to that person's profile.[1]

Currently, in DBpedia dumps, there are around 100 million RDFs. DBpedia retrieves these triples from multilingual Wiktionaries. It functions as similar to a thesaurus in 171 human and machine languages, including Greek, German, English, and a variety of others[2]. The instances or entities used in the DBpedia data sets are represented by classes and properties in the DBpedia ontology. The information and data from Wikipedia's info-boxes are used to create the ontology.[3] Wikipedia maintains a template structure to store its factual and vital information about the article in the info-boxes. In order to include improvements, DBpedia has released several versions throughout time. Dbpedia first offered rules and instructions in version 3.2 for manually creating mappings between the info-box data and the ontology classes. It also addressed a number of problems with Wikipedia's info-box structure, including the existence of several info-boxes with the same class name, the use of the same name for distinct attributes, and the use of different names for the same kind of characteristics. Additionally, info-boxes did not follow a uniform format or data-type structure. Due to the aforementioned problems, DBpedia enabled manually extracting data from info-boxes and mapping that data to the appropriate classes. As a result, DBpedia's data became more organised and precise than that of Wikipedia. With the release of version 3.5, DBpedia made it possible for the community to contribute by offering mapping instructions, allowing them to create the info-box with ontology mapping, and allowing them to edit the already-existing mappings and classes in the ontology. As a result of more classes and attributes being introduced by other contributors, mapping statistics improved. The classes in the Dbpedia Ontology are organised in a hierarchical structure of parent classes with sub-classes as of the release of DBpedia version 3.7. As a musician, for instance, 'Michael Jackson' corresponds to the class 'Musician,' which is a subclass of 'Artist,' which is a subclass of 'Person,' where 'Thing' is the parent class and 'Person' is a child of 'Agent.' Any instance that does not belong to or map to a specific class is by default mapped with the class 'Thing.' The current ontology contains a total of 768 classes and 3000 different properties[4]. The info-boxes are the most significant and valuable part of the Wikipedia articles for the extractions and DBpedia mappings. Info-boxes are commonly used to describe structured information about an article's facts and figures in the form of a table. The relevant facts are enlisted in the form of attributes and their values. The info-box for languages like Urdu and Arabic that begin on the left and end on the right is located on the right side of the Wikipedia page, while the info-box for languages that start on the left is located on the left side of the article. Lehmann et al. (2015).

Urdu Wikipedia[5] is a tiny hub of knowledge containing a minimal count of articles of around 164,887, which is increasing with time. Still, the Urdu-speaking community requires effort to add value to Urdu Wikipedia through authentic and updated information by following a uniform and proper template structure. Most of the time, info-boxes are placed on the right-hand side of the Wikipedia article, but for Urdu and Arabic, they are found on the left side, as shown in Figure 1.

In Urdu Wikipedia, the attributes in the info-box can be written in Urdu or English, but Urdu is preferred; attribute values should also be written in Urdu, but a few are also found in English. One such example can be seen in Urdu[6]. In this article, our focus is to map Urdu info-box properties, attributes, and classes with English DBpedia Ontology. Moreover, we are unable to find any mapping directions in the literature for Urdu info-box mapping. In this regard we have also published an article regarding the challenges and case of Urdu DBpedia Rasham et al. (2022) in which we have highlighted the complexities in the creation of Urdu Dbpedia such as lack of consistent template structure, unorganised and unstructured content in Urdu Wikipedia, missing attributes, and lack of Urdu dependencies in DBpedia extraction framework. Since the existing solutions applied by other language editions could not work on Urdu due to limited and unavailable resources, we opted first to adopt the manual mapping approach. Moreover, DBpedia also lacked significant support in mapping from Urdu Wikipedia. We have also presented a case for the integration of the Urdu mappings with DBpedia. In this article, we covers the following research questions.

- How Urdu mapping of templates is done with English ontology classes?
- How the attributes and properties of Urdu Wikipedia Info-box are mapped with English?
- How Urdu labels are added in the definition of ontology classes?
- how many mappings are performed?
- What are the challenges faced by the research community in the mapping process?

The rest of the article is organised as, Section 2 summarised the related work. Section 3 explains the proposed methodology. In Section 4 we have summarised the challenges and results. Finally, the conclusion and recommendations are discussed in Section 5.

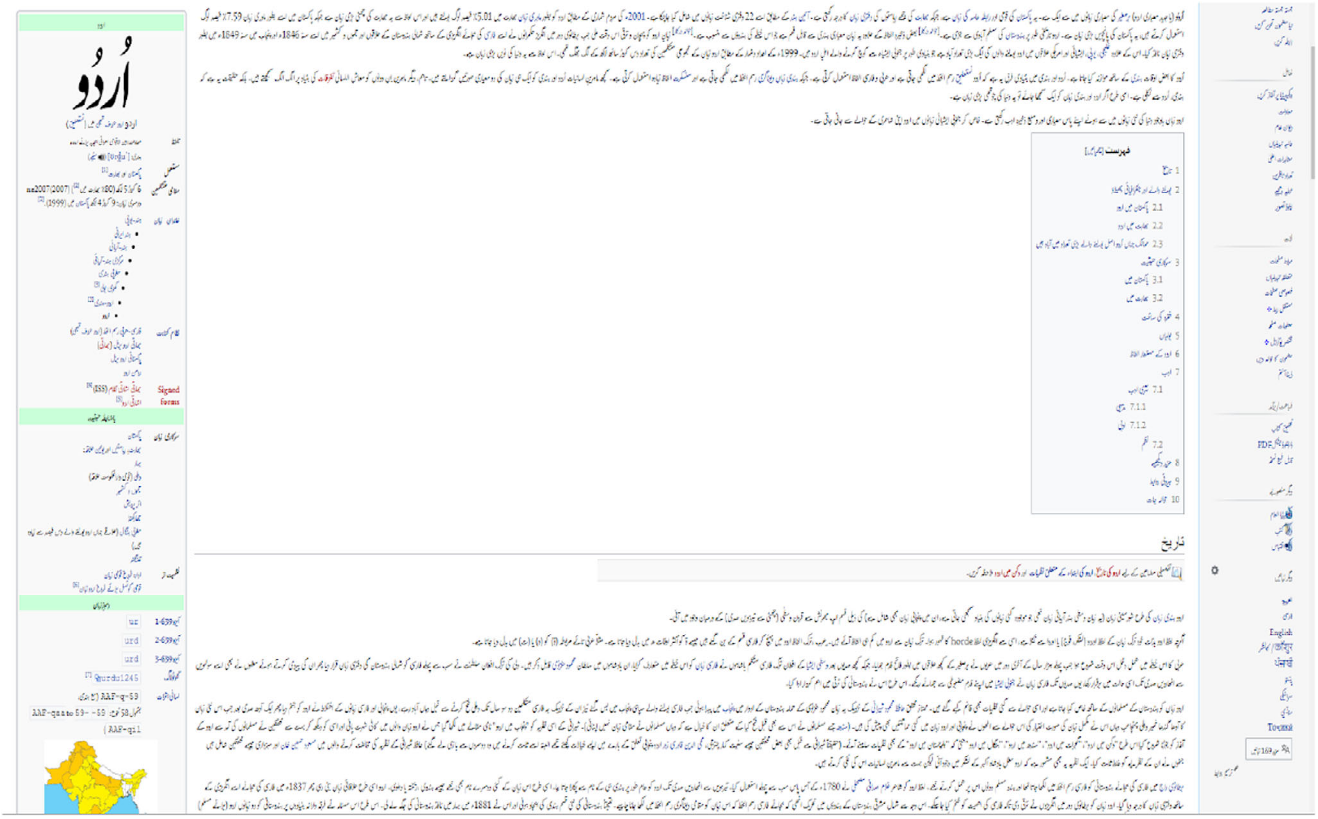


FIGURE 1 Position of info-box in Urdu Wikipedia

## 2 | RELATED WORK

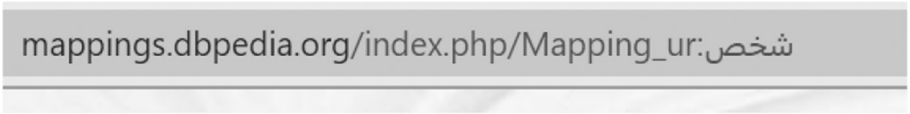
The knowledge base of DBpedia grows as Wikipedia is updated and represents the current status of Wikipedia by extracting live Wikipedia changes. The quality of information was improved by mapping Wikipedia's info-box templates to the DBpedia ontology. Editors were facilitated in integrating their data with DBpedia by constructing RDF links from DBpedia to various additional data sources. DBpedia contains details of different fields but not limited to personalities, celebrities, corporations, movies, music, pharmaceuticals, books, and scientific publications. The extraction framework also performs live and dump-based information extraction from info-boxes. Various people define the same things differently or use different labels for identical features, such as birthplace and placeofbirth. DBpedia addresses this situation by employing two separate extraction strategies in parallel: general info-box extraction, which covers all info-boxes and their properties, and mapping-based info-box extraction, which ensures sound data quality Bizer et al. (2009). DBpedia applications also provides community members with access to various open data sets and interfaces, including Linked-Data, RDF Dumps, the SPARQL query language, and many others. Various web editors can use data sets provided by DBpedia to incorporate information into their web pages Auer et al. (2008). This strategy also connects DBpedia to other databases, serving as a hub for the emerging open data web. Moreover DBpedia connects with numerous databases and other project resources such as YAGO Tanon et al. (2020), Media-Wiki Rogushina and Grishanova (2020), Freebase Färber et al. (2018), and SPARQL endpoint Bonifati et al. (2020) as it enhances the way that information is presented and integrated with other semantic and link data technologies. Furthermore, data from DBpedia is being used in applications such as Chatbots Følstad et al. (2020) which give a controlled interface for simulating and generating intelligent human discussions. Additionally, the community can access Linked-Data, RDF dumps, SPARQL query protocol, and many other open data sets and interfaces through DBpedia applications Auer et al. (2007). Many web editors can use DBpedia's data sets to add information into their websites. This strategy also links DBpedia to other databases, enabling it to serve as a center for the developing Open Data Web. Other than DBpedia, various knowledge graphs like Wikidata, centrally stores all structured data of Wiki-media applications, and the data is accessible by SPARQL queries. Wikidata is thus grouped with other basic knowledge graphs like DBpedia, and their comparison made by Abián et al. (2017) clearly showed that Wikidata is more open and centralised, while DBpedia is more established in the Web Of data and Linked Open Data groups and is based on Wikipedia's various multilingual editions.

DBpedia has made several multilingual data sets available, including data from several Wikipedia language versions. Using mappings from the DBpedia network, the data retrieved from these Wikipedia instances is transformed into RDF. Nonetheless, not all mappings are accurate and uniformly consistent across all the DBpedia datasets. Because the incorrect mappings are scattered among many mappings, personally examining each one to ensure accuracy is impractical. As a result, by assessing information from both entity data and ontological constraints, data analysis and interpretation-based technique for automatically detecting erroneous and improper mappings was proposed by Rico et al. (2018). A prediction-based machine-learning approach was presented for detecting incorrect and false mappings. Various supervised classification methods were tested for this task, and the proposed model attained a 93 percent accuracy. These results aid in the identification of incorrect mappings as well as the building of a high-quality DBpedia mapping.

Spanish DBpedia recently adopted the approach of using the DBpedia data bus for its periodic updates, and automatic data extractions Sanz-Lucio et al. (2021). The concept behind utilising the Databus is to retrieve the data so that it can develop metadata of the datasets and then publish it on the Databus, allowing the users to query, download, extract, and create applications on top of the data.

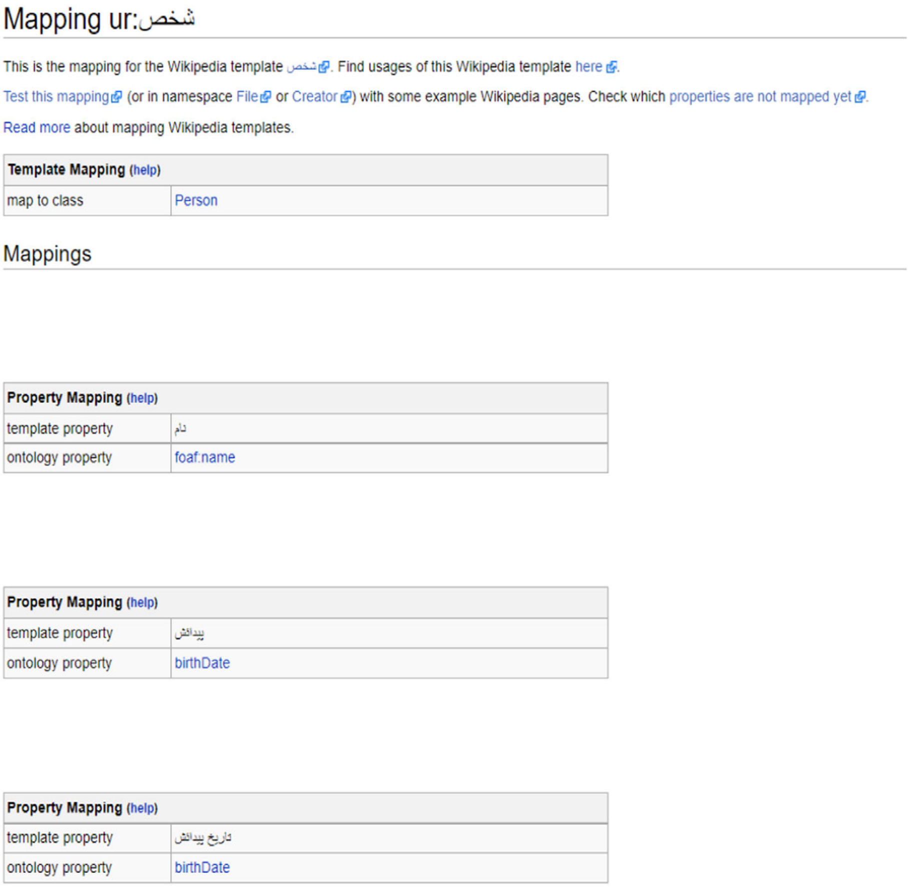
Moreover, DBpedia also supports different crowd-sourcing techniques to enrich its knowledge base and facilitate mappings from Wikipedia in other linguistic editions with the ontology classes. One such method used by Aprosio et al. (2013) is to align the DBpedia classes across Wikipedia languages by using categorization and classification. However, lack of focus on assigning new entities and knowledge to the categories, this approach did not gather further information. Similarly, the integration of Urdu Wikipedia with DBpedia also needs efforts from Urdu speaking community to enrich its knowledge graph. Our approach also aligns with this aim by adding attributes to classes. Under the umbrella of Natural Language Processing (NLP), the instances and concepts from Wikipedia and DBpedia are mapped with the WordNet so that the synsets and descriptions can be replaced with the link to Wikipedia articles and further can discover and explore machine-readable knowledge from resources such as DBpedia McCrae (2018). In the same manner, Paulheim and Ponzetto (2013) used NLP methods to extend the coverage of DBpedia instances and ontology through list pages of Wikipedia.

In one research over Chatbots, Kondylakis et al. (2020) introduced R2D2, a semantic web-based intelligent chatbot that provides a natural language interface with intelligent control for using DBpedia to retrieve the data. The core engine of this chatbot responded to structured input



mappings.dbpedia.org/index.php/Mapping_ur:شخص

FIGURE 2 URI resource of mapping Urdu



Mapping ur:شخص

This is the mapping for the Wikipedia template شخص. Find usages of this Wikipedia template here.

Test this mapping (or in namespace File or Creator) with some example Wikipedia pages. Check which properties are not mapped yet.

Read more about mapping Wikipedia templates.

| Template Mapping (help) | |
|---|---|
| map to class | Person |

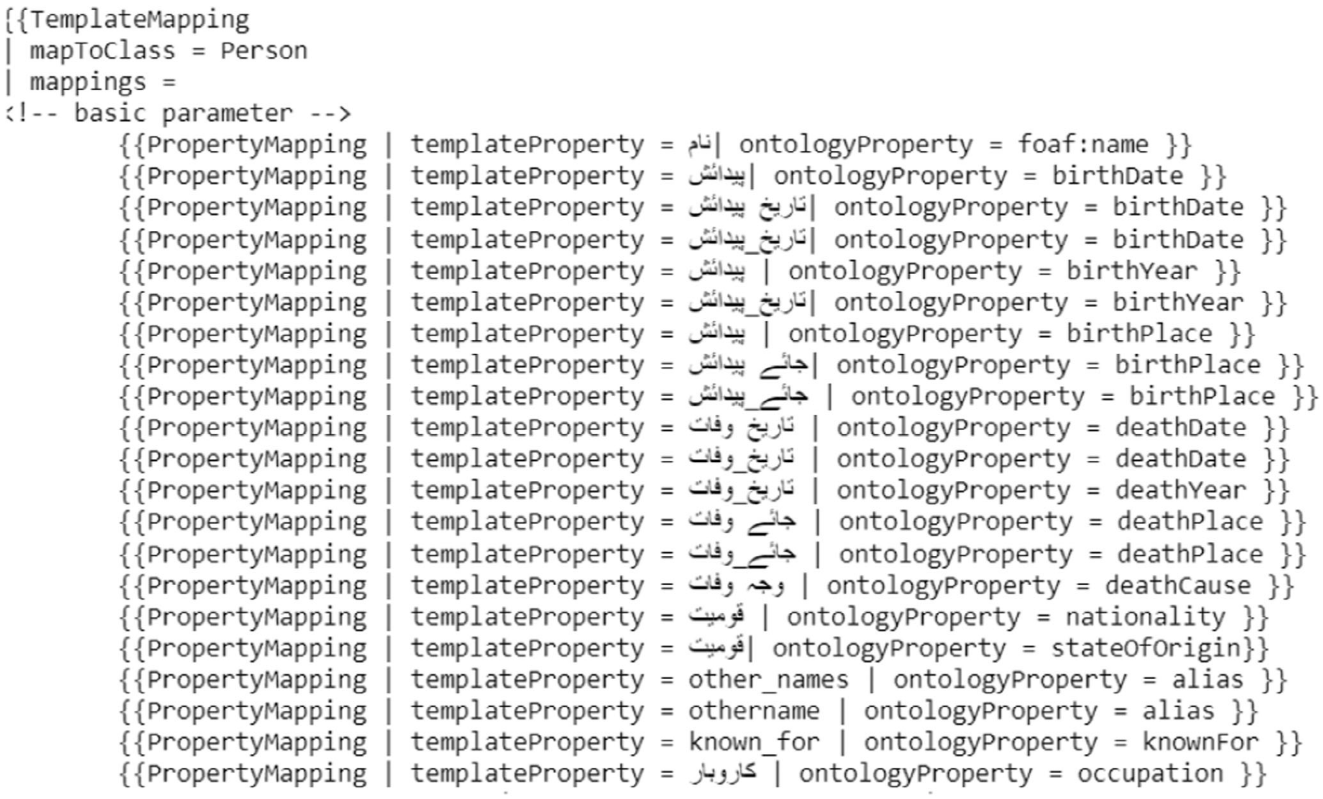
Mappings

| Property Mapping (help) | |
|---|---|
| template property | نام |
| ontology property | foaf:name |

| Property Mapping (help) | |
|---|---|
| template property | پیدائش |
| ontology property | birthDate |

| Property Mapping (help) | |
|---|---|
| template property | تاریخ پیدائش |
| ontology property | birthDate |

FIGURE 3 The already mapped info-box mapping Ur



```
{{TemplateMapping
| mapToClass = Person
| mappings =
<!-- basic parameter -->
        {{PropertyMapping | templateProperty = نام| ontologyProperty = foaf:name }}
        {{PropertyMapping | templateProperty = پیدائش| ontologyProperty = birthDate }}
        {{PropertyMapping | templateProperty = تاریخ پیدائش| ontologyProperty = birthDate }}
        {{PropertyMapping | templateProperty = تاریخ_پیدائش| ontologyProperty = birthDate }}
        {{PropertyMapping | templateProperty = پیدائش | ontologyProperty = birthYear }}
        {{PropertyMapping | templateProperty = تاریخ_پیدائش| ontologyProperty = birthYear }}
        {{PropertyMapping | templateProperty = پیدائش | ontologyProperty = birthPlace }}
        {{PropertyMapping | templateProperty = جائے پیدائش| ontologyProperty = birthPlace }}
        {{PropertyMapping | templateProperty = جائے_پیدائش | ontologyProperty = birthPlace }}
        {{PropertyMapping | templateProperty = تاریخ وفات | ontologyProperty = deathDate }}
        {{PropertyMapping | templateProperty = تاریخ_وفات | ontologyProperty = deathDate }}
        {{PropertyMapping | templateProperty = تاریخ_وفات | ontologyProperty = deathYear }}
        {{PropertyMapping | templateProperty = جائے وفات | ontologyProperty = deathPlace }}
        {{PropertyMapping | templateProperty = جائے_وفات | ontologyProperty = deathPlace }}
        {{PropertyMapping | templateProperty = وجہ وفات | ontologyProperty = deathCause }}
        {{PropertyMapping | templateProperty = قومیت | ontologyProperty = nationality }}
        {{PropertyMapping | templateProperty = قومیت| ontologyProperty = stateOfOrigin}}
        {{PropertyMapping | templateProperty = other_names | ontologyProperty = alias }}
        {{PropertyMapping | templateProperty = othername | ontologyProperty = alias }}
        {{PropertyMapping | templateProperty = known_for | ontologyProperty = knownFor }}
        {{PropertyMapping | templateProperty = کاروبار | ontologyProperty = occupation }}
```

**FIGURE 4** Mapping template and properties format of person to ontology class

that allowed users to ask inquiries using triple-patterns. As they type, an auto-complete service offers DBpedia resources to help users create the triple patterns. User input in the form of triple-pattern queries was automatically used to generate the relevant SPARQL queries. R2D2 received the results from the respective DBpedia SPARQL endpoint, augmented them with maps and other graphics, and then displayed them to the user. A similar interactive chatbot that can converse with people in real time and respond to challenging queries using a basic graphical user interface can be implemented in Urdu. Users can explore the available data in real-time by asking queries to the chatbot in Urdu. The user's request will be automatically translated to a SPARQL query and then sent to the DBpedia SPARQL endpoint for an Urdu response.

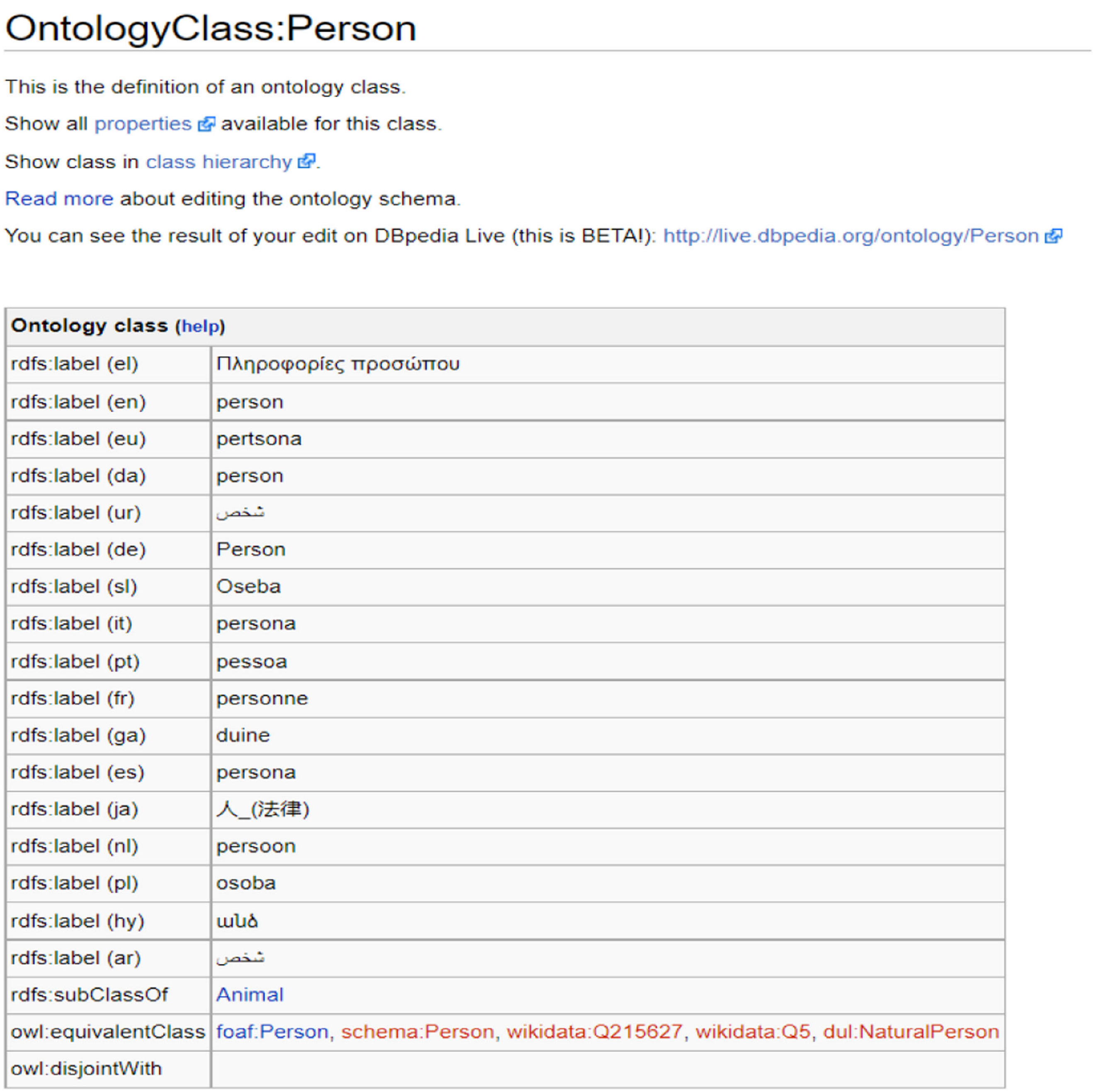
As social media and micro-blogging networks have grown, a significant number of brief textual collection of documents are produced every day; this necessitates the adoption of efficient organised and classification techniques. Flisar and Podgorelec (2020) suggested a novel method in which short text documents were used to locate relevant concepts using the DBpedia Spotlight framework. and then added the information from the DBpedia ontology to the text to reduce its sparsity. According to the findings, the suggested text enrichment method greatly enhanced the categorization of short texts and was resilient to a variety of input sources, domains, and training data sizes.

Despite the existence of various medical standard vocabularies, it is still difficult to accurately identify the concepts contained in electronic medical records. The coverage and abstraction of these texts' annotations may have varied due to the annotation methodology and knowledge graphs used, leading to noticeably differing outcomes in the annotations that were automatically processed. Gazzotti et al. (2020) suggested a semi-supervised approach based on DBpedia for extracting medical topics from electronic medical records (EMRs),[7] and assessing the effect of including these topics in the parameters used to describe EMRs in the task of hospitalisation prediction.

A lot of work has to be done to establish a complete chapter of Urdu DBpedia. Earlier, there was nothing done in Urdu DBpedia since the essential resources required to perform the automatic mapping techniques and extraction were not available for Urdu. So, we started from the initial step and laid the foundation of mapping by adopting a manual mapping approach. We have also enlisted all the challenges related to mapping for infoboxes and related attributes.

## 3 | PROPOSED METHODOLOGY

DBpedia maps the data from the info-boxes with its Ontology classes. In the case of Urdu Wikipedia, in some info-boxes, the attributes are missing or vary as compared to English Wikipedia info-box, different template structure and there exist such articles which are directly translated from English to Urdu through a translator which may not guarantee the authentication of data as mentioned in.[8] Also, there are so many articles with



# OntologyClass:Person

This is the definition of an ontology class.

Show all properties available for this class.

Show class in class hierarchy.

Read more about editing the ontology schema.

You can see the result of your edit on DBpedia Live (this is BETA!): http://live.dbpedia.org/ontology/Person

| Ontology class (help) | |
|---|---|
| rdfs:label (el) | Πληροφορίες προσώπου |
| rdfs:label (en) | person |
| rdfs:label (eu) | pertsona |
| rdfs:label (da) | person |
| rdfs:label (ur) | شخص |
| rdfs:label (de) | Person |
| rdfs:label (sl) | Oseba |
| rdfs:label (it) | persona |
| rdfs:label (pt) | pessoa |
| rdfs:label (fr) | personne |
| rdfs:label (ga) | duine |
| rdfs:label (es) | persona |
| rdfs:label (ja) | 人_(法律) |
| rdfs:label (nl) | persoon |
| rdfs:label (pl) | osoba |
| rdfs:label (hy) | անձ |
| rdfs:label (ar) | شخص |
| rdfs:subClassOf | Animal |
| owl:equivalentClass | foaf:Person, schema:Person, wikidata:Q215627, wikidata:Q5, dul:NaturalPerson |
| owl:disjointWith | |

FIGURE 5 Adding labels in the ontology class definition-I

no info-box, such as the Wikipedia article of Harappa Museum[9] of Archaeology has no info-box. Being low on article count, Urdu Wikipedia still requires a lot of community support to increase the number of articles with proper template structure and referencing. As there are such redirect links that point to non-existing articles. For example: from the list of Museums in Pakistan, the Wikipedia articles pages, 18 out of 31 redirects links to the Urdu Wikipedia articles of the mentioned museums do not exist[10] Similarly, 09 out of 40 links point to the pages that have not shown any English Wikipedia articles[11] We have also found museum names both in English and Urdu without having any Wikipedia pages for them. This is an example of one such inconsistency. Urdu Wikipedia article count can be increased by the collaborative effort of Urdu community and Language based research centres for Urdu. The mappings from Urdu to English DBpedia ontology needed to be created from scratch. The mapping statistics for Urdu are also not available to date, which makes us unable to track the current ratio of mapped and un-mapped templates, properties and classes.[12] As far as linking of Urdu Wikipedia with DBpedia is concerned. This research is the first attempt in creating the mappings of Urdu with English DBpedia ontology classes. The mappings are created manually due to the discussed issues of Urdu Wikipedia. The steps of the proposed mapping is as under:

- Mapping of properties and templates with English ontology classes.
- Mapping of labels with ontology classes.

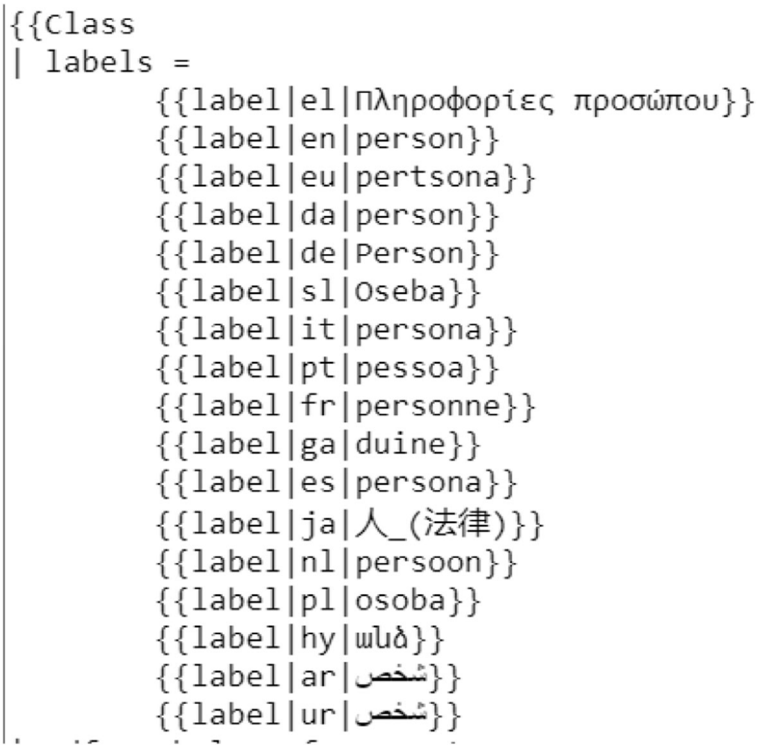
In order to generate mappings of Urdu with DBpedia, we acquired authorization from DBpedia officials and then continued with our methodology explained further. First, it is needed to check whether the info-box template is already mapped or not. To validate an info-box template's mapping, we need the info-box name concatenated with the mapping of the Urdu chapter's URI in the mapping ur web page[13]. Such an example for ‘Person’ class in Urdu can be seen in Figure 2.

If the mapping to the class of ontology already exists then we do not need to re-map as the template will be displayed that shows the mapping to the class within the ontology as well as the mapping of different properties, as shown in mapping[13] in Figure 3. If the Wikipedia info-box is not mapped already, we create the mappings page in DBpedia by mapping the info-box template with its property attributes to the appropriate ontology class characteristics, as shown in Figure 4. The info-box of a person in Urdu is mapped to the DBpedia Ontology class ‘Person’ at ‘mapToClass,’ whereas in the later part, the attributes of the info-box are mapped with the properties of the Ontology class of DBpedia such as name, birth date, occupation, and so on.

Moreover, in the definition of ontology classes and their sub-classes, the Urdu labels are also added alongside other established DBpedia chapters[14] as shown in Figure 5. It can be seen in the syntax shown in Figure 6 that we have edited the class definition of ontology and added its Urdu label of the respective ‘Person’ class with other linguistic labels like in Arabic, English, French, and many more.

This process of mapping will be continued until all the existing infoboxes are mapped. Further, it will lead to mappings of ontology classes, properties, and datatypes of Urdu with English DBpedia. The Figure 7 reflects the imagery view of the mapping process discussed above and shows in detail the complete step-by-step process of how we have achieved the mapping the data from infoboxes with DBpedia ontology classes. These mappings and labels will play a meaning-full role for other Dbpedia frameworks and semantics. Furthermore, these mapping guidelines would be the road-map for future extensions and they would be automated as framework for mapping once we have a structured and an appropriate number of articles in Urdu Wikipedia for every domain. After which SPARQL query language framework will be applied to generate responses to the end user.

We have also added Urdu labels in the ontology class definition. Examples are presented in Figure 9.



```
{{Class
| labels =
        {{label|el|Πληροφορίες προσώπου}}
        {{label|en|person}}
        {{label|eu|pertsona}}
        {{label|da|person}}
        {{label|de|Person}}
        {{label|sl|Oseba}}
        {{label|it|persona}}
        {{label|pt|pessoa}}
        {{label|fr|personne}}
        {{label|ga|duine}}
        {{label|es|persona}}
        {{label|ja|人_(法律)}}
        {{label|nl|persoon}}
        {{label|pl|osoba}}
        {{label|hy|անձ}}
        {{label|ar|شخص}}
        {{label|ur|شخص}}
```

FIGURE 6 Adding labels in the ontology class definition-II

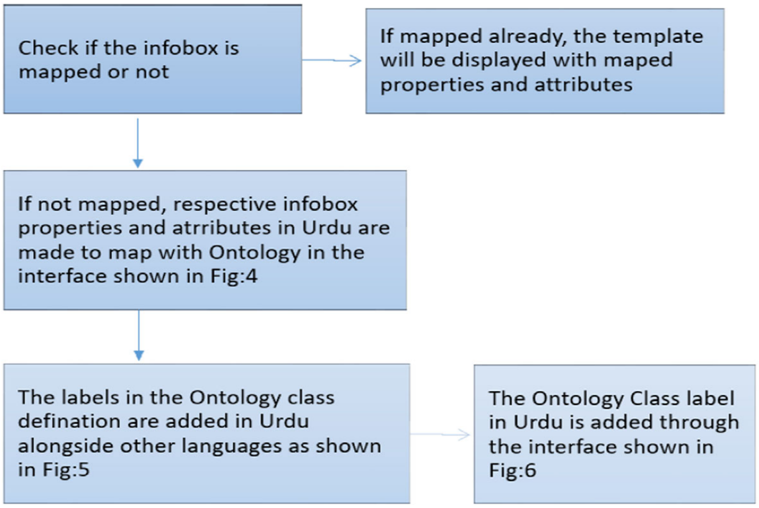
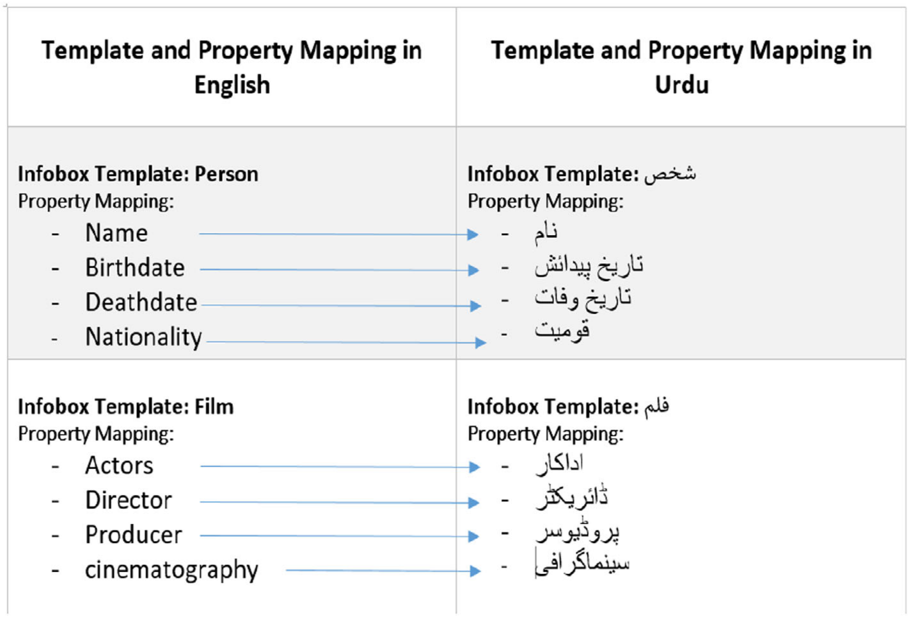

Check if the infobox is mapped or not
If mapped already, the template will be displayed with maped properties and attributes
If not mapped, respective infobox properties and atrributes in Urdu are made to map with Ontology in the interface shown in Fig:4
The labels in the Ontology class defination are added in Urdu alongside other languages as shown in Fig:5
The Ontology Class label in Urdu is added through the interface shown in Fig:6

FIGURE 7 Process of mapping Urdu with DBpedia ontology



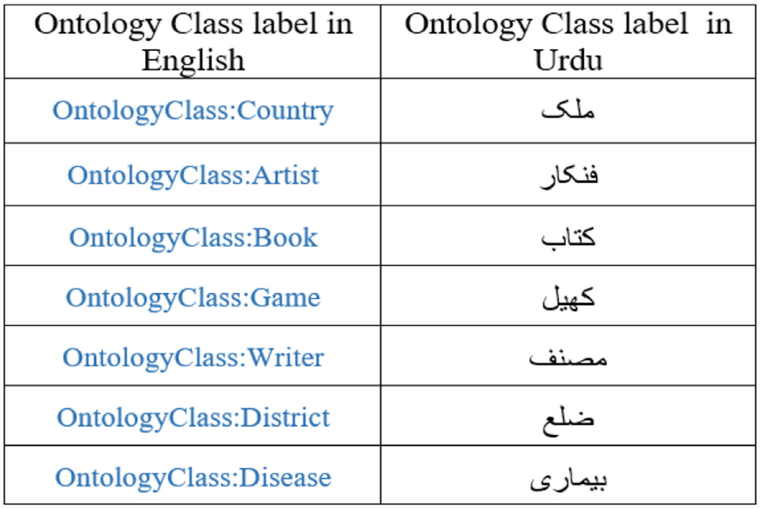
| Template and Property Mapping in English | Template and Property Mapping in Urdu |
|---|---|
| **Infobox Template: Person**<br>Property Mapping:<br>- Name<br>- Birthdate<br>- Deathdate<br>- Nationality | **Infobox Template:** شخص<br>Property Mapping:<br>- نام<br>- تاریخ پیدائش<br>- تاریخ وفات<br>- قومیت |
| **Infobox Template: Film**<br>Property Mapping:<br>- Actors<br>- Director<br>- Producer<br>- cinematography | **Infobox Template:** فلم<br>Property Mapping:<br>- اداکار<br>- ڈائریکٹر<br>- پروڈیوسر<br>- سینماگرافی |

FIGURE 8 The info-box template and property mappings of Urdu with English

## 4 | RESULTS AND DISCUSSION

We have mapped around 1000 info-box templates to their corresponding classes and sub-classes of DBpedia ontology. The mapping example is shown in Figure 8.

These mappings can be viewed at[15]. Moreover, have summarised the following challenges while mapping classes and properties. There is no standardised template structure in the Urdu Wikipedia. This makes it challenging to automatically extract and map info-box features to ontology classes. Moreover, Urdu Wikipedia is neither as well-organised nor as well-resourced with material as English Wikipedia, particularly the info-boxes, which are considerably different in both languages and often use distinct templates. The Urdu Wikipedia has a lot of unstructured information. To initiate mapping, the info-box table values must first be extracted via Wikipedia dumps or live extraction. Due to changes in template structure, missing attributes, and different and insufficient information in Urdu Wikipedia info-boxes compared to its English equivalent of info-boxes, the extraction framework is unable to correctly extract and map properties with ontology classes. In presence of these infrastructure issues, automated mapping is not possible. Because the DBpedia framework for extraction is typically used for mapping in other language



| Ontology Class label in English | Ontology Class label in Urdu |
|---|---|
| OntologyClass:Country | ملک |
| OntologyClass:Artist | فنکار |
| OntologyClass:Book | کتاب |
| OntologyClass:Game | کھیل |
| OntologyClass:Writer | مصنف |
| OntologyClass:District | ضلع |
| OntologyClass:Disease | بیماری |

**FIGURE 9** The Urdu labels added in ontology class definition

chapters of DBpedia, some languages' dependencies are included in extraction files for the extraction framework, while dependencies for Urdu are not available. Urdu does not have as many resources as other languages such as English, Greek, German, and so forth. In the presence of all these challenges, we are able to map few info-box templates and their attributes in Urdu Wikipedia to DBpedia ontology.

## 5 | CONCLUSIONS

The mapping of info-boxes and their related attributes in Urdu DBpedia is one of its unique efforts. We have mapped around a thousand properties and attributes of Urdu with English DBpedia ontology. Our contribution leads the first step in the internationalisation of DBpedia for Urdu. This mapping plays a vital role in the evolution of DBpedia editions and integrating the structured data from an unstructured and raw form in Urdu. To further support the DBpedia internationalisation project, additional work is still needed to expand the mapping coverage and automate the process.

### AUTHOR CONTRIBUTION

All authors have equally contributed to the completion of the article.

### CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

### DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

### ORCID

*Habib Ullah Khan* https://orcid.org/0000-0001-8373-2781

*Fahad Maqbool* https://orcid.org/0000-0003-3969-5551

### ENDNOTES

[1] https://www.dbpedia.org/resources/linked-data/

[2] http://downloads.dbpedia.org/wiki-archive/wiktionary-rdf-extraction.html

[3] https://www.dbpedia.org/resources/ontology/

[4] https://www.dbpedia.org/resources/ontology/

[5] https://en.wikipedia.org/wiki/Urdu_Wikipedia

[6] https://ur.wikipedia.org/wiki/%D8%A7%D8%B1%D8%AF%D9%88

[7] Electronic medical records (EMRs) provide vital information on a patient's health, and their examination should enable the prevention of diseases that could impact the patient in the future.

[8] https://meta.wikimedia.org/wiki/Requests_for_comment/Concerned_about_Urdu_Wikipedia_articles%27_truthiness_and_neutrality

[9] https://en.wikipedia.org/wiki/Archaeological_Museum_Harappa

[10] https://ur.wikipedia.org/wiki/%D9%BE%D8%A7%DA%A9%D8%B3%D8%AA%D8%A7%D9%86_%D9%85%DB%8C%DA%BA_%D8%B9%D8%AC%D8%A7%D8%A6%D8%A8_%DA%AF%DA%BE%D8%B1%D9%88%DA%BA_%DA%A9%DB%8C_%D9%81%DB%81%D8%B1%D8%B3%D8%AA

[11] https://en.wikipedia.org/wiki/List_of_museums_in_Pakistan

[12] http://mappings.dbpedia.org/server/statistics/ur/

[13] http://mappings.dbpedia.org/index.php/Mapping_ur:%D8%B4%D8%AE%D8%B5

[14] http://mappings.dbpedia.org/index.php/OntologyClass:Person

[15] http://mappings.dbpedia.org/index.php/Mapping_ur

## REFERENCES

Abián, D., Guerra, F., Martínez-Romanos, J., & Trillo-Lado, R. (2017). Wikidata and DBpedia: A comparative study. In *Semanitic keyword-based search on structured data sources* (pp. 142–154). Springer.

Aprosio, A. P., Giuliano, C., & Lavelli, A. (2013). Automatic expansion of DBpedia exploiting wikipedia cross-language information. Extended semantic web conference (pp. 397–411). Springer.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The semantic web* (pp. 722–735). Springer.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2008). DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th international semantic web conference (ISWC)* (Vol. 4825, pp. 722–735). Springer.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia-a crystallization point for the web of data. *Journal of Web Semantics*, *7*(3), 154–165.

Bonifati, A., Martens, W., & Timm, T. (2020). An analytical study of large sparql query logs. *The VLDB Journal*, *29*(2), 655–679.

Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked data quality of DBpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, *9*(1), 77–129.

Flisar, J., & Podgorelec, V. (2020). Improving short text classification using information from DBpedia ontology. *Fundamenta Informaticae*, *172*(3), 261–297.

Følstad, A., Araujo, T., Papadopoulos, S., Law, E. L.-C., Granmo, O.-C., Luger, E., & Brandtzaeg, P. B. (2020). *Chatbot research and design*. Springer.

Gazzotti, R., Faron-Zucker, C., Gandon, F., Lacroix-Hugues, V., & Darmon, D. (2020). Injection of automatically selected DBpedia subjects in electronic medical records to boost hospitalization prediction. In Proceedings of the 35th annual ACM symposium on applied computing (pp. 2013–2020). The Association for Computing Machinery.

Kondylakis, H., Tsirigotakis, D., Fragkiadakis, G., Panteri, E., Papadakis, A., Fragkakis, A., Tzagkarakis, E., Rallis, I., Saridakis, Z., Trampas, A., Pirounakis, G. (2020). R2d2: A DBpedia chatbot using triple-pattern like queries. *Algorithms*, *13*(9), 217.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., et al. (2015). DBpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, *6*(2), 167–195.

McCrae, J. P. (2018). Mapping wordnet instances to wikipedia. Proceedings of the 9th Global Wordnet Conference (pp. 61-68). Global Wordnet Association.

Morsey, M., Lehmann, J., Auer, S., Stadler, C., & Hellmann, S. (2012). DBpedia and the live extraction of structured data from wikipedia. *Program*, *46*(2), 157–181.

Paulheim, H.), & Ponzetto, S. P. (2013). Extending DBpedia with wikipedia list pages. NLP-DBpedia@ ISWC, 13.

Rasham, S., Naz, A., Afzal, Z., Ahmed, W., Abbas, Q., Anwar, M. H., Ejaz, M., & Ilyas, M. (2022). The challenges and case for urdu DBpedia. In Proceedings of international conference on information technology and applications (pp. 439–448). Springer.

Rico, M., Mihindukulasooriya, N., Kontokostas, D., Paulheim, H., Hellmann, S., & Gómez-Pérez, A. (2018). Predicting incorrect mappings: A data-driven approach applied to DBpedia. In Proceedings of the 33rd annual ACM symposium on applied computing (pp. 323–330). The Association for Computing Machinery.

Rogushina, J., & Grishanova, I. (2020). Ontological methods and tools for semantic extension of the media wiki technology. *Problems in Programming*, *2-3*, 61–73.

Sanz-Lucio, S., Tahiri-Alaoui, O., & Rico, M. (2021). Latest enhancements in the spanish DBpedia.

Tanon, T. P., Weikum, G., & Suchanek, F. (2020). Yago 4: A reason-able knowledge base. In European semantic web conference (pp. 583–596). Springer.

## AUTHOR BIOGRAPHIES

**Shanza Rasham** is a student of MSCS program at University of Sargodha. Her research interest includes semantic web and ontology engineering. Email: shanzyrasham@gmail.com

**Habib Ullah Khan** is working as a Professor of MIS in the Department of Accounting & Information Systems, College of Business and Economics, Qatar University, Qatar. He completed his PhD degree in Management Information Systems from Leeds Beckett University, UK. He has nearly 25 years of industry, teaching and research experience. His research interests are in the area of IT Adoption, Social Media, Internet Addiction, Mobile Commerce, Computer Mediated Communication, IT Outsourcing, Big data, and IT Security. Email: habib.khan@qu.edu.qa

**Fahad Maqbool** is working as assistant professor of computer science at University of Sargodha, Pakistan. His research interest includes large-scale global optimization, cluster computing, and FAIRification. Email: fahad.maqbool@uos.edu.pk

**Saad Razzaq** received MSCS degree in 2006 from NUCES, Islamabad. He has 15+ years teaching experience and currently working as assistant professor at University of Sargodha. His research interest includes semantic web, intelligent systems, and ontology designing. Email: saad.razzaq@uos.edu.pk

**Sajid Anwar** received the B.Sc. (comp. sc.) and M.Sc. (comp. sc.) degrees from the University of Peshawar, in 1997 and 1999, respectively, and the M.S. (comp. sc.) and PhD degrees in software architecture from the University of NUCES-FAST, Pakistan, in 2007 and 2011, respectively. He is currently an Associate Professor of computing science with the Institute of Management Sciences, Peshawar, Pakistan. His research interests include software architecture, software requirement engineering, searched-based software engineering, and mining software repository. Email: sajid.anwar@imsciences.edu.pk

**Muhammad Ilyas** is an assistant professor at department of computer science and information technology, University of Sargodha, Pakistan. He received his M.S. in Software Project Management (in 2004) from NUCES-FAST, Lahore, Pakistan, and PhD (in 2010) in Computer Science from Institute for Application Oriented Knowledge Processing, Johannes Kepler University of Linz, Austria. Currently, his featured administrative roles include Director of Information Technology at the University of Sargodha. His research interests include information retrieval and information and knowledge management systems. He has been a reviewer of numerous national and international journals. He has supervised to completion of many M.S. research students. Furthermore, he has published over 50 research articles in prestigious conferences and journals. Email: muhammad.ilyas@uos.edu.pk

**How to cite this article:** Rasham, S., Khan, H. U., Maqbool, F., Razzaq, S., Anwar, S., & Ilyas, M. (2023). Structured knowledge creation for Urdu language: A DBpedia approach. *Expert Systems*, e13223. https://doi.org/10.1111/exsy.13223