Translation of article 10.22146/jnteti.v13i4.13289

MULTI-WD: Multilingual Completion Tool for Wikidata Data

Mohammad Yani[1], Lilyan Arhatia Agustine[1], Iryanto[1]

[1] Program Studi Rekayasa Perangkat Lunak, Jurusan Teknik Informatika, Politeknik Negeri Indramayu, Indramayu, Jawa Barat 45252, Indonesia

[Submitted: 4 June 2024, Revised: 19 August 2024, Accepted: 1 October 2024]
Corresponding Author: Mohammad Yani (email: mohammad.yani@polindra.ac.id)

ABSTRACT — Wikidata, a rapidly expanding knowledge graph (KG), owes its growth to two primary factors. First, Wikidata allows open access and editing by anyone. Second, it offers a multilingual feature that enables data entities to be accessed in various languages worldwide. However, the issue of incomplete information across multiple languages remains a significant challenge. For instance, the description of the entity "bada reuteuk" (ID: Q100606305) is currently available only in Indonesian as "a traditional food in Indonesia," but it lacks descriptions in other languages. Consequently, these data are not accessible or recognizable in languages other than Indonesian. The system incorporates two primary features: language profiling and data translation. Language profiling, implemented using SPARQL queries via the Wikidata API, provides an overview of the multilingual status of Wikidata entities. For data translation, the system utilized the Translated Labs library, chosen for its open access, cost-free availability, and high-quality translation outputs. The translated results are subsequently saved into Wikidata. System evaluation involved five respondents from the Wikidata community, using a black-box testing approach. Results demonstrated that MULTI-WD's core functionalities—including category selection, data statistics display, translation, and data updates—achieved 100% operational success. Furthermore, the tool enhanced data translation efficiency by up to 300% compared to manual translation directly through the Wikidata interface.

KEYWORDS — Wikidata Profiling, Wikidata Multilingualism, Multilingual Profiling, Data Completeness.

I. INTRODUCTION

Over the past decade, the use of knowledge graphs (KGs) in semantic web technology research has grown significantly. A KG is a secondary database that contains a collection of descriptions of entities interconnected through relationships [1], [2]. Entities in a KG represent objects, situations, events, and abstract concepts in the real world. Several major KGs exist today, with Wikidata being one of the most prominent and widely used. Wikidata is an open-access KG that can be accessed and edited by both humans and machines. It contains over 100 million entities and thousands of properties [3]. Wikidata is one of the largest knowledge graphs currently available [4].

Wikidata is particularly appealing for research on knowledge graph question answering (KGQA) systems for two key reasons. First, Wikidata is open access, allowing anyone to edit and access its data [5]–[7], enabling rapid data expansion. Second, Wikidata supports multilingual features, making it accessible in multiple languages. Due to these two factors, Wikidata is widely utilized in research on semantic web technologies, particularly in KGQA systems that implement multilingual approaches [8], [9].

However, the multilingual feature of Wikidata currently faces a significant limitation in terms of completeness. This issue arises when specific data in Wikidata are available in one language but not in others. For instance, the description of the entity "bada reuteuk" with the ID "Q100606305" is only available in Indonesian, as "makanan khas di Indonesia" (a traditional food in Indonesia), but lacks descriptions in other languages. Such cases hinder the recognition of information about "bada reuteuk" in languages other than Indonesian. The incomplete information typically includes the "Label," "Description," and "Also known as" fields of a Wikidata entity.

At least two factors contribute to the incompleteness of multilingual information in Wikidata. First, the number of Wikidata users and editors varies significantly across languages. The number of contributors from Asian countries is considerably lower than those from European countries [10], resulting in a dominance of Wikidata information in European languages. Second, there is limited time available to complete the information for a Wikidata entry in multiple languages. Figure 1 illustrates that the data for "bada reuteuk" is complete only in Indonesian but remains incomplete in other languages. As shown in Figure 1, the English data are available only in the "Label" section, with no information provided in the description or "Also known as" fields. Similarly, data in other languages also lacks completeness.

Figure 1 illustrates that the information about the entity "bada reuteuk" is only complete in Indonesian (represented by the "Indonesian" field in the language column), while the information in other languages remains incomplete. The missing fields in the Label," "Description," and "Also known as" columns highlight the core issue addressed in this study.

One prior study has addressed the multilingual aspect of Wikidata [11], focusing on completing Wikidata descriptions using data from Wikipedia. However, this approach heavily depends on the completeness of Wikipedia entries, as it employs a point-to-point approach within a single language rather than translating data [11].

Additionally, two studies have explored data completeness in Wikidata: COOL-WD [12] and ProWD [13]. However, both focus exclusively on nonmultilingual cases. COOL-WD emphasizes the completeness of Wikidata's lifecycle, while ProWD proposes a tool to measure the completeness of Wikidata using a class-facet-attribute (CFA) framework. Neither study offers a solution for addressing the issue of


Bada reuteuk (Q100606305)

No description defined

In more languages

Configure

Language	Label	Description	Also known as
English	Bada reuteu	No description defined	
Indonesian	Bada reuteu	variasi makanan khas Indone	bada reuteuk
Javanese	Bada reuteu	No description defined	
Sundanese	Bada reuteu	No description defined	

All entered languages

Figure 1. A screenshot showcasing an example of incomplete data on Wikidata presented in a multilingual format.

multilingual incompleteness in Wikidata, leaving a research gap that this study aims to address.

Building on this premise, this research proposed a tool capable of performing data profiling through a semantic-based query approach and enhancing Wikidata's information in a multilingual context using machine translation libraries. While the semantic-based query approach is widely utilized to extract data from knowledge graphs, the literature review reveals that it has not yet been applied to address the challenge of data incompleteness in Wikidata's multilingual support.

This research aimed to address the issue of data incompleteness in Wikidata's multilingual context by utilizing a semantic-based query approach combined with a translation engine library. The study makes two key contributions. First, it seeks to provide users with a convenient means to enrich Wikidata data in languages that are currently unavailable. Second, it aims to deliver a comprehensive multilingual Wikidata dataset to support research in KGQA systems.

Additionally, this research is expected to facilitate data profiling for low-resource languages on Wikidata. Low-resource languages are defined as those with significantly less data available compared to other languages, presenting unique challenges for knowledge representation and accessibility [14].

II. MULTILINGUALISM OF WIKIDATA DATA

Previous research has explored aspects of multilingualism in Wikidata. One study focused on event linking for Wikidata description data across multiple languages by linking Wikipedia [11]. In this approach, Wikidata descriptions were enriched using search results for similar events found on Wikipedia. However, this method has limitations: not all data in Wikidata is represented on Wikipedia, and it only retrieves descriptions in the same language without translating content between languages.

Additional research has addressed tools for assessing Wikidata data completeness, such as COOL-WD [12] and ProWD [13]. COOL-WD is an annotation tool designed to evaluate the completeness of the Wikidata data lifecycle, whereas ProWD provides a dashboard for profiling statement completeness of Wikidata entities using the CFA approach. However, neither study specifically targets data completeness in the multilingual context. Instead, they focus on assessing the completeness of subjects and predicates in statements, ensuring the minimum requirements for data completeness are met. In response to these gaps, this research proposes a novel tool that profiles Wikidata data using a semantic-based query approach and enriches multilingual information through machine translation libraries. The primary contribution of this research is to create a comprehensive multilingual dataset for Wikidata. This enriched dataset is intended to support KGQA systems that rely on multilingualism. The KGQA system enables users to access data from knowledge graphs using natural language question inputs [15].

III. MULTI-WD

This section presents the architecture and implementation of MULTI-WD, along with an overview of its main program components.

A. MULTI-WD ARCHITECTURE

The MULTI-WD architecture is specifically designed to profile Wikidata data using a semantic query-based approach and to translate Wikidata information through a translation engine library. MULTI-WD incorporates two primary features: multilingual profiling of Wikidata data and the translation of Wikidata entries. To implement these features, the system is composed of three main components: SPARQL, Translated Labs, and Pywikibot. The MULTI-WD architecture design is illustrated in Figure 2. SPARQL is utilized to retrieve data from Wikidata, Translated Labs facilitates the translation of data from the source language to the target language, and Pywikibot enables access to and modification of Wikidata data using Python.

1) MULTILINGUAL PROFILING

This feature conducts multilingual profiling by retrieving information about the "Label," "Description," and "Also known as" fields of a Wikidata record. MULTI-WD employs SPARQL for its semantic query-based operations. Semantic-based queries are used to extract data represented as graphs within a KG. These graphs are typically structured as triples, consisting of a subject, predicate, and object.

SPARQL is utilized by MULTI-WD to perform multilingual profiling of Wikidata data. SPARQL is a query language specifically designed to retrieve data in the form of triples from a KG [16]. It is capable of answering queries by extracting information from a KG, whether represented in the resource description framework (RDF) or web ontology language (OWL) format [17]. Additionally, SPARQL can query a KG via an application programming interface (API), enabling further processing of the retrieved data [18]. SPARQL is a widely used tool for facilitating communication between users and KGs in KGQA systems [19].

In this research, data from Wikidata were retrieved in several official languages of the United Nations (UN), including English, French, Spanish, and Indonesian, which is also recognized as an official instructional language by UNESCO [20]. These four languages were selected because they are globally recognized as official languages of both the UN and UNESCO and have the highest number of Wikidata contributors. Figure 3 provides an example of retrieving data, such as the ID and label, for all Indonesian food types in the Indonesian language via the Wikidata endpoint. To enable the use of SPARQL in Python, MULTI-WD employs the SPARQLWrapper library.

In Figure 3, the query command is divided into two parts: the SELECT clause, which specifies the variables to be included in the query results, and the WHERE clause, which defines the basic graph patterns to be matched with the knowledge graph (KG). The syntax in Figure 3 can be modified

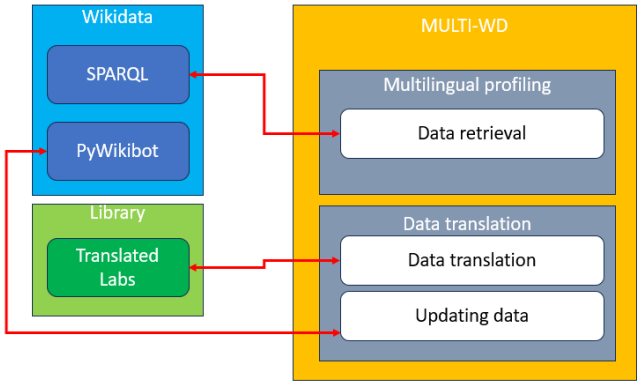

Figure 2. Architecture of MULTI-WD.

to retrieve data in other languages by adjusting the language settings within the SERVICE wikibase:label graph pattern.

In the example shown in Figure 3, the query outputs two values, represented by the variables ?makanan and ?makananLabel, which are derived from two graph patterns. These graph patterns are expressed as triples in the format (subject, predicate, object). The two graph patterns are detailed in rows 3 and 4 of Figure 3.

The first graph pattern, shown in row 3, refers to all entities that are "instances of" (P31) the entity "food" (Q2095). The second graph pattern, shown in row 4, specifies entities of type "food" that originate from the "country" (P17) "Indonesia" (Q252). In Wikidata, IDs beginning with the letter "P" represent properties, while IDs starting with the letter "Q" represent entities.

The SERVICE wikibase:label { bd:serviceParam wikibase:language "id". } is a graph pattern used to filter information specifically in the Indonesian language ("id"). The language setting can be adjusted by replacing the language code "id" with the code for another language, such as "en" for English. The query results are subsequently displayed in the variables ?food and ?foodLabel, as illustrated in Figure 4.

2) DATA TRANSLATION

This feature consists of two subfeatures: data translation and data update. For the data translation feature, MULTI-WD utilized the Translated Labs translation engine library. Translated Labs was selected due to its open-source nature, making it accessible to users globally. It is an artificial intelligence (AI)-based tool capable of translating text between languages via an API. Translated Labs employed the Modern Machine Translation (MMT) model developed by MMT in 2018 [21]. In this feature, the label, description, and also known as fields from the original Wikidata data served as input for Translated Labs. Using the MMT model, these fields were translated into the target language. However, the translation results from this feature were not immediately stored in Wikidata. To update Wikidata with the translated information, MULTI-WD included a data update feature. This feature ensures that the translated data is integrated into Wikidata. To update the information from the translated data, MULTI-WD implemented a data update feature. This feature updates the data for the information translated into Wikidata. To perform this process, MULTI-WD used PyWikibot. MULTI-WD utilized PyWikibot, a Python library specifically designed to facilitate interaction between Python applications and Wikidata. The overall process of MULTI-WD is illustrated in Figure 5.

Figure 3. SPARQL query to retrieve all Indonesian food type data in Indonesian language.

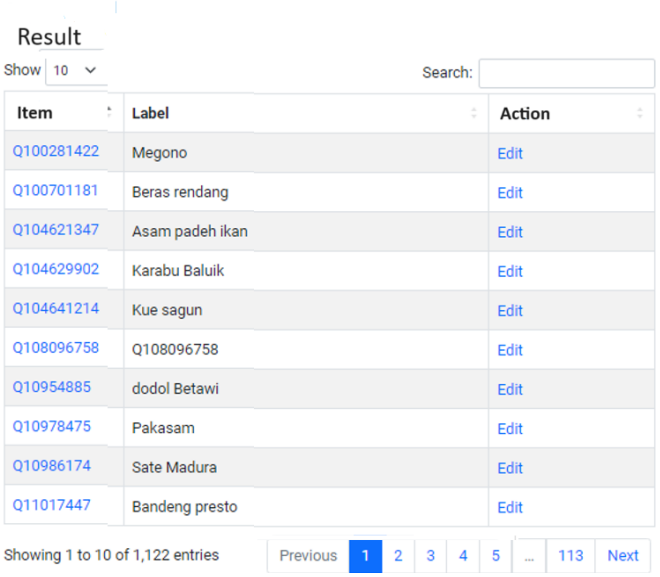

Figure 4. Screenshot of Indonesian food data query result in Indonesian language.

B. IMPLEMENTATION OF MULTI-WD

This section details the implementation of MULTI-WD, including its system setup and use cases. MULTI-WD operated on a 64-bit computer equipped with an 11th Gen Intel® Core™ i7-1165G7 processor running at 2.80 GHz and 16 GB of RAM. The back-end of MULTI-WD was developed using Python, with SPARQL serving as the query service to retrieve data from Wikidata via the PyWikibot library. For the front end, MULTI-WD utilized the Laravel 8 framework.

1) SYSTEM IMPLEMENTATION

MULTI-WD is a web-based application designed for data profiling and the translation of Wikidata information. It performs data profiling using SPARQL queries, while the translation of Wikidata information is facilitated by the Translated Labs translation engine library. To enable communication between Python and Wikidata, MULTI-WD utilizes the PyWikibot library.

2) USE CASE ON MULTI-WD

This subsection outlines several use cases that MULTI-WD supports, including retrieving data from Wikidata, selecting specific data types, monitoring data statistics by language, translating data, and updating it in Wikidata.

The first use case involved data retrieval from Wikidata. This process was performed using SPARQL queries executed via Python. Data were retrieved based on the criteria specified in the query. An example query for data retrieval is shown in Figure 3.

The second use case focuses on selecting the type of data to retrieve from Wikidata. Users can filter data based on specific categories, such as dances, food, songs, or traditional houses. Once a data type is selected, MULTI-WD retrieves and

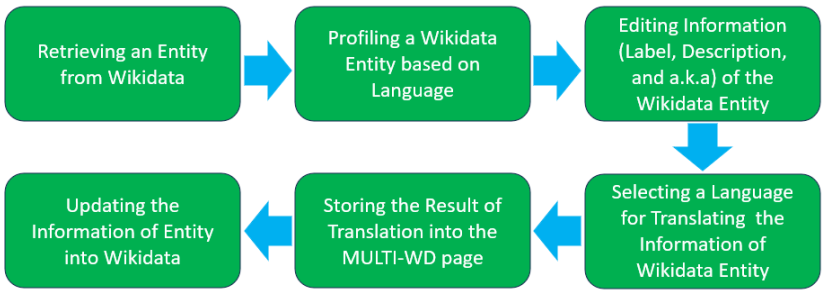

Figure 5. MULTI-WD Wikidata entity profiling process flow.

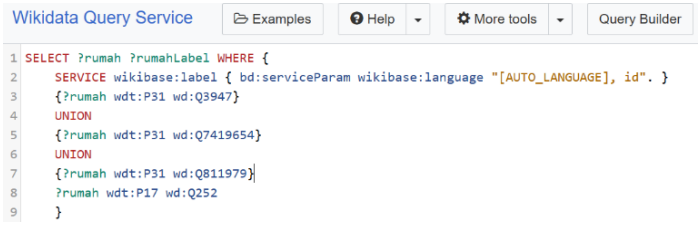

Figure 6. SPARQL query to display traditional house data.

displays all records associated with that type. For instance, Figure 6 presents an example of data categorized under the traditional house type.

In the SELECT clause, the query retrieves two pieces of data, which are stored in the variables ?rumah and ?rumahLabel. In the WHERE clause, the query matches several graph patterns. Specifically, this query matches four graph patterns, as outlined in rows 3, 5, 7, and 8 of Figure 6. Each pattern is structured as a triple, consisting of a subject, predicate, and object. For instance, the triple ?rumah wdt:P31 wd:Q3947 indicates that ?rumah is the subject, wdt:P31 is the predicate, and wd:Q3947 is the object. In this context, two PREFIXes are used: "wdt" and "wd".The "wdt" prefix is employed to denote Wikidata properties, while "wd" refers to Wikidata entities.

The graph pattern shown in row 3 of Figure 6 shows that query matches all data in Wikidata that corresponds to the triple ?rumah wdt:P31 wd:Q3947. Here, the property P31 represents the Wikidata property "instance of," and the entity Q3947 corresponds to the Wikidata entity "house." By using this graph pattern, all data identified as an "instance of" the entity "house" is stored in the variable ?rumah.

The graph pattern in the fifth row of Figure 6 retrieves all data that is an instance of (P31, "instance of") the entity Q7419654 ("traditional house"), which is then stored in the variable ?rumah. Similarly, the graph pattern in the 7th row retrieves all data that is an instance of (P31, "instance of") the entity Q811979 ("architectural structure"), which is also stored in the variable ?rumah. Additionally, the graph pattern in the 8th row retrieves all data associated with the property P17 ("country") for the entity Q252 ("Indonesia"), and these data are likewise stored in the variable ?rumah.

The graph patterns in lines 3–5 of Figure 6 each generate a separate set of query results, which are then combined using the UNION syntax. However, for the combined results to be valid, they must all satisfy the graph pattern specified in line 7 of Figure 6, ensuring that all entities originate from the country Indonesia (wdt:P17 wd:Q252).

The syntax SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],id". } wsd used to retrieve data written in Indonesian. To retrieve data in English, the language code can be changed from "[AUTO_LANGUAGE],id" to "[AUTO_LANGUAGE],en". The final two characters, id, and en, represent the Indonesian and English languages, respectively. The result of this SERVICE syntax is stored in the variable ?rumahLabel and displayed accordingly.

The third use case involved the statistical presentation of data quantity based on type and language. This statistical data was generated by aggregating information by language and visualized as a bar graph. The statistics were further categorized based on the "Label," "Description," and "Also known as" fields.

The fourth use case focused on translating information into a target language. MULTI-WD translates the "Label," "Description," and "Also known as" fields from the original language to the destination language using the Translated Labs library.

The fifth use case involved updating Wikidata with the translated data generated in the fourth use case. To perform these updates or edits, MULTI-WD used the PyWikibot library.

IV. SYSTEM DEMONSTRATION

In this section, a demonstration of the MULTI-WD system is presented, covering the demonstration scenario, running examples, testing, and validation of the results. The complete source code for MULTI-WD is available at https://github.com/moh-yani/multi-wd/.

A. DEMONSTRATION SCENARIO

The demonstration was carried out on the existing use cases of MULTI-WD, including statistical analysis of Wikidata data, data translation, and data updates. The subsections below provide further details on the demonstration process.

B. RUNNING EXAMPLE

This section uses the example shown in Figure 3, which features data on Indonesian specialties. The first step in using MULTI-WD is to select the type of data to be profiled. In this example, the selected data type is typical food. After selecting typical food, MULTI-WD displays detailed information and statistical analysis related to this data type.

After selecting the data type, MULTI-WD generates Wikidata data statistics in four languages, categorized by "Label," "Description," and "Also known as." Figure 7 provides an example of these statistics, specifically focusing on

the completeness of the description field across the four languages.

From Figure 7, it is evident that the description of typical food in Spanish has the lowest level of completeness compared to the other three languages, followed by French and English. In the context of KGs, languages like Spanish and French, which have limited data availability, are considered low-resource languages. Based on the data presented in Figure 4, users can perform language editing through the "edit" action. This process involves translating the data from the original language into the target language using the "Translate Text" button. Once the translation and editing are completed, MULTI-WD allows real-time updates to Wikidata data through the "Edit Data" button. The workflow for editing and translating data is illustrated in Figure 8. The results of the translation process for the "bada reuteuk" data shown in Figure 1, processed via the workflow in Figure 8, can be accessed at https://www.wikidata.org/wiki/Q100606305.

C. TESTING

This stage involved a functional test of MULTI-WD by five members of the Wikidata Indonesia community, who are also active Wikidata users and editors. Wikidata Indonesia is a nonprofit organization and a local partner of the Wikimedia Foundation, responsible for managing Wikipedia and other related wiki projects. The test focused on evaluating the use cases of MULTI-WD through black-box testing. The primary objective was to assess whether the system could perform data profiling (statistical presentation), data translation, and real-time updates to Wikidata data or not.

The test was conducted on four categories of Wikidata instances: typical foods, traditional dances, traditional weapons, and traditional houses. For each category, ten instances were randomly selected to undergo profiling (statistical presentation), translation, and real-time updates. Testers were given the flexibility to select instances randomly based on the category. Additionally, they were allowed to choose both the source and target languages during the translation process and decide whether to save the translated results to Wikidata. The following section outlines some of the test cases for MULTI-WD, including the objectives, expected outcomes, and actual results for each test case.

1) CATEGORY SELECTION TESTING

The purpose of this testing is to evaluate whether MULTI-WD can select data categories and display the corresponding data. The expected outcome is that MULTI-WD accurately displays data based on the selected category, including the "Label," "Description," and "Also known as" fields. The testing results, as summarized in Table I, confirm that the process of retrieving data by category was successful. The detailed scenario for the category selection testing is presented in Table I.

2) MULTILINGUAL STATISTICS DISPLAY TESTING FOR WIKIDATA ENTITIES

The purpose of this testing is to verify whether MULTI-WD can display Wikidata data statistics for the selected data type across four languages in the form of a bar graph. The expected outcome is that MULTI-WD generates data statistics and presents them as bar charts. The testing results, as outlined in Table II, demonstrate that this functionality operates as intended, successfully displaying Wikidata data statistics in bar graph form.

Figure 7. Display of statistical description of "traditional food" data in four languages.

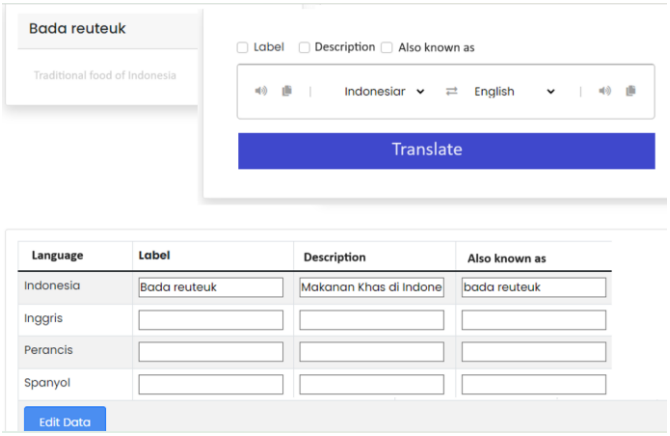

Figure 8. Interface of data translation and data update pages.

3) ENTITY DATA TRANSLATION TESTING

The purpose of this test is to evaluate MULTI-WD's ability to translate data from the original language to the target language and to assess whether the translation results comply with Wikidata data editing rules. The testing results, summarized in Table III, demonstrate that the system can successfully translate data between languages while adhering to Wikidata's editing guidelines. The editing rules are documented and available at https://github.com/moh-yani/multi-wd/. Details of the data translation test scenario are provided in Table III.

4) TESTING DATA UPDATES TO WIKIDATA

The purpose of this test is to verify the system's ability to update Wikidata data in real-time and ensure that no bugs occur during the process. The expected outcome is that the system updates Wikidata data in real-time without encountering any issues. The testing results, as demonstrated by the data available at https://www.wikidata.org/wiki/Q100606305, confirm that the system performs its intended function as a language editor for Wikidata data in real-time and operates without bugs. Details of this data update test scenario are presented in Table IV.

The process of updating entity data to Wikidata is generally feasible. However, the success of this update largely depends on the status of the Wikidata API. Data updates may fail if the API experiences a timeout condition. The test results for updating data to Wikidata, as shown in Table IV, were obtained under conditions where the API was not in a timeout state.

D. VALIDATION

Validation was conducted by five members of the Wikidata Indonesia community using a direct observation approach and

TABLE I
CATEGORY SELECTION TEST-CASE SCENARIO

Identification				**Test Case**
Test Case				Category selection
Objective				Reviewing the display after selecting a category, including the category data results and the complete data for "Label," "Description," and "Also known as" fields, categorized by language
Scenario				
1. Click select option to choose a cultural category 2. Display the results of the selected cultural category 3. Display the data completeness results in the form of a diagram				
Expected Outcome				
1. Display entity data based on the selected category 2. Display a diagram presenting the statistical data of an entity				
Statistics				
Number of Respondents	***Number of Category Test Data Per Respondent***	***Succeed***	***Failed***	***Description***
5	4	20	0	As per the test objectives

TABLE II
TEST-CASE SCENARIO OF MULTILINGUAL STATISTICS DISPLAY OF WIKIDATA ENTITY

Identification				**Test Case**
Test Case				Multilingual statistics interface of Wikidata entities
Objective				Reviewing the display after the user selects an entity (the system will present its data statistics in multiple languages)
Scenario				
1. Click show entity list on the "Select option" menu 2. Click the entity name on the row containing the list of entities				
Expected Outcome				
The system displays the statistics of an entity's multilingual completeness data in the form of a diagram				
Statistics				
Number of Respondents	***Number of Entity Test Data Per Respondent***	***Succeed***	***Failed***	***Description***
5	40	200	0	As per the test objectives

by reviewing the results of data editing performed through MULTI-WD. The objective of this stage was to assess whether the system functions as intended. A manual validation method was chosen to ensure results closely align with the gold standard, relying on human expertise for accuracy.

Validation for the data selection and data display use cases was conducted by comparing results with data retrieved directly thr ough the Wikidata endpoint at https://query.wikidata.org/. Results are considered valid if and only if the data generated by MULTI-WD matches the data retrieved directly from the Wikidata endpoint. The validation statistics for the data selection and display use cases are presented in Table I.

TABLE III
ENTITY DATA TRANSLATION TEST-CASE SCENARIO

Identification				**Test Case**
Test Case				Entity data translation
Objective				Verifying whether the system can translate an entity's data from the original language to the target language and place the translation result in the edit field
Scenario				
1. Select an entity to be translated 2. Choose the source and target languages from the menu 3. Click the "Translate" menu				
Expected Outcome				
The system successfully translates entity data from the original language to the target language and places it in the edit field				
Statistics				
Number of Respondents	***Number of Entity Test Data Per Respondent***	***Succeed***	***Failed***	***Description***
5	40	200	0	As per the test objectives

TABLE IV
TEST-CASE SCENARIO FOR UPDATING DATA TO WIKIDATA

Identification				
Test Case				Entity data update to Wikidata
Objective				Verifying the system's ability to successfully update entity data on Wikidata pages.
Scenario				
1. Click the "Edit Data" button, as shown in Figure 8 2. Verify the updated entity data on the Wikidata page				
Expected Outcome				
The system successfully updates entity data on the Wikidata page				
Statistics				
Number of Respondents	***Number of Entity Test Data Per Respondent***	***Succeed***	***Failed***	***Description***
5	40	200	0	As per the test objectives

Validation for the data translation use case was performed by directly observing the translation results to determine whether they conform to the Wikidata data editing rules, which are available at https://github.com/moh-yani/multi-wd/. The results are deemed valid if and only if the translations align with these rules. Validation statistics for the data translation use case are provided in Table III.

Validation for the data update use case was conducted by verifying the updated data directly on the relevant Wikidata page. For instance, the results of updates to the "bada reuteuk" data can be verified at https://www.wikidata.org/wiki/Q100606305. The results are considered valid if and only if the data updated on the Wikidata page matches the data generated by MULTI-WD. Validation statistics for this use case are provided in Table IV.

E. DISCUSSION

According to the test results presented in Tables I through IV, the functionality of MULTI-WD, including both unit

testing and system testing, generally performs as expected. However, specific observations can be made regarding the data translation and update features for Wikidata. First, the data translation feature relies on machine translation by default. While functional, these machine-generated translations may not meet gold-standard natural language quality and may require user refinement. To address this, MULTI-WD includes an edit feature that allows users to modify and improve the translated data before finalizing it.

Second, the data update feature may experience temporary interruptions due to exceeding the API access limits imposed by Wikidata. As this is an external limitation beyond the control of MULTI-WD, the system implements error-handling mechanisms to minimize user disruption. Users are notified of any issues encountered during the update process through clear messages, ensuring they understand the nature of the problem and preventing the impression of a critical system error.

In terms of efficiency, MULTI-WD is significantly more effective, being 300% more efficient than manual translation through the Wikidata website. Unlike conventional methods that require three separate processes to translate the "Label," "Description," and "Also known as" fields, MULTI-WD consolidates these into a single, streamlined process.

If MULTI-WD is adopted on a large scale by Wikidata users or communities, it has the potential to significantly enhance the multilingual completeness of Wikidata data. This improvement would contribute to advancing the multilingual capabilities of Wikidata, particularly benefiting research in the development of Wikidata-based KGQA systems.

V. CONCLUSION

MULTI-WD provides a solution for multilingual profiling of Wikidata data, addressing the issue of incomplete multilingual information in Wikidata—a problem that has not yet been explored in existing research. Its implementation can contribute to improving the completeness of datasets used in research on Wikidata-based KGQA systems. MULTI-WD has been tested using a black-box approach across four main functions: category selection, multilingual statistics display, data translation, and data updates to Wikidata. These tests were conducted by five respondents from the Wikidata community, and the results indicate that all functions performed successfully, achieving 100% functionality. In terms of efficiency, MULTI-WD demonstrated the ability to increase translation efficiency by 300% compared to manual methods. However, the system still faces limitations in generating grammatically perfect translations that align with natural human language. Future research could focus on enhancing MULTI-WD by integrating more advanced open language models that reflect the current state-of-the-art while considering cost-effective solutions. Additionally, MULTI-WD could be expanded to support other forms of language profiling as KGQA systems and their associated datasets continue to evolve.

CONFLICTS OF INTEREST

The authors declare that this research was conducted and written with no conflict of interest.

AUTHORS' CONTRIBUTIONS

Conceptualization, Mohammad Yani; methodology, Mohammad Yani; writing-original draft, Mohammad Yani and Lilyan Arhatia Agustine; writing-reviewing and editing, Mohammad Yani; software, Lilyan Arhatia Agustine; visualization, Lilyan Arhatia Agustine; supervision, Mohammad Yani and Iryanto; project administration, Lilyan Arhatia Agustine; funding acquisition, Mohammad Yani; resources, Mohammad Yani and Lilyan Arhatia Agustine; validation, Mohammad Yani and Lilyan Arhatia Agustine; investigations, Mohammad Yani and Iryanto.

ACKNOWLEDGMENT

Gratitude is extended to the Indramayu State Polytechnic Research and Community Service Center for providing research funding. This study serves as an additional output from research activities funded under contract number: 0720/PL42.PL42.9/AL.04/2024.

REFERENCES

[1] A. Hogan *et al*., "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, May 2022, doi: 10.1145/3447772.

[2] A.A. Krisnadhi, M. Yani, and I. Budi, "Entity and relation linking for knowledge graph question answering using gradual searching," *J. Nas. Tek. Elek. Teknol. Inf.*, vol. 13, no. 2, pp. 139–146, May 2024, doi: 10.22146/jnteti.v13i2.9184.

[3] M. Yani and A.A. Krisnadhi, "Challenges, techniques, and trends of simple knowledge graph question answering: A survey," *Information*, vol. 12, no. 7, pp. 1–31, Jul. 2021, doi: 10.3390/info12070271.

[4] S. Figueroa, "Knowledge discovery in Wikidata with machine learning in graph," in *Inf. Syst. Technol.*, A. Rocha dkk., Eds., Cham, Switzerland: Springer, 2024, pp. 3–12, doi: 10.1007/978-3-031-45645-9_1.

[5] K. Tharani, "Much more than a mere technology: A systematic review of Wikidata in libraries," *J. Acad. Librariansh.*, vol. 47, no. 2, pp. 1–8, Mar. 2021, doi: 10.1016/j.acalib.2021.102326.

[6] K. Shenoy *et al*., "A study of the quality of Wikidata," *J. Web Semant.*, vol. 72, pp. 1–10, Apr. 2022, doi: 10.1016/j.websem.2021.100679.

[7] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Oct. 2014, doi: 10.1145/2629489.

[8] M.U. Akhtar *et al*., "Entity alignment based on relational semantics augmentation for multilingual knowledge graphs," *Knowl.-Based Syst.*, vol. 252, pp. 1–10, Sep. 2022, doi: 10.1016/j.knosys.2022.109494.

[9] A. Perevalov, D. Diefenbach, R. Usbeck, and A. Both, "QALD-9-plus: A multilingual dataset for question answering over DBpedia and Wikidata translated by native speakers," in *2022 IEEE 16th Int. Conf. Semant. Comput. (ICSC)*, 2022, pp. 229–234, doi: 10.1109/ICSC52841.2022.00045.

[10] Z. Shaik, F. Ilievski, and F. Morstatter, "Analyzing race and country of citizenship bias in Wikidata," dalam *2021 IEEE 18th Int. Conf. Mob. Ad Hoc Smart Syst. (MASS)*, 2021, pp. 665–666, doi: 10.1109/MASS52906.2021.00099.

[11] A. Pratapa, R. Gupta, and T. Mitamura, "Multilingual event linking to Wikidata," in *Proc. Workshop Multiling. Inf. Access (MIA)*, 2022, pp. 37–58, doi: 10.18653/v1/2022.mia-1.5.

[12] F. Darari, R.E. Prasojo, S. Razniewski, and W. Nutt, "COOL-WD: A completeness tool for Wikidata," in *CEUR Workshop Proc.*, 2017, pp. 1–4.

[13] A. Wisesa *et al*., "Wikidata completeness profiling using ProWD," in *K-CAP 2019 - Proc. 10th Int. Conf. Knowl. Capture*, 2019, pp. 123–130, doi: 10.1145/3360901.3364425.

[14] L.-A. Kaffee *et al*., "Multilingual knowledge graphs and low-resource languages: A review," *Trans. Graph Data Knowl. (TGDK)*, vol. 1, No. 1, pp. 1–19, Dec. 2023, doi: 10.4230/TGDK.1.1.10.

[15] M. Yani, A.A. Krisnadhi, and I. Budi, "A better entity detection of question for knowledge graph question answering through extracting position-based patterns," *J. Big Data*, vol. 9, pp. 1–26, Jun. 2022, doi: 10.1186/s40537-022-00631-1.

[16] E. Prud'hommeaux and A. Seaborne. "SPARQL query language for RDF." W3C. Access date: 15-Jan-2024. [Online]. Available: https://www.w3.org/TR/rdf-sparql-query/

[17] G. Xiao and J. Corman, "Ontology-mediated SPARQL query answering over knowledge graphs," *Big Data Res.*, vol. 23, pp. 1–25, Feb. 2021, doi: 10.1016/j.bdr.2020.100177.

[18] M. Mosser *et al*., "Querying APIs with SPARQL," *Inf. Syst.*, vol. 105, pp. 1–14, Mar. 2022, doi: 10.1016/j.is.2020.101650.

[19] M. Bakhshi, M. Nematbakhsh, M. Mohsenzadeh, and A.M. Rahmani, "Data-driven construction of SPARQL queries by approximate question graph alignment in question answering over knowledge graphs," *Expert Syst. Appl.*, vol. 146, pp. 1–19, May 2020, doi: 10.1016/j.eswa.2020.113205.

[20] K. Syamsi *et al*., "Developing a culture-based Indonesian language textbook for non-native speakers for academic purposes," *Cakrawala Pendidik.*, vol. 43, no. 1, pp. 115–126, Feb. 2024, doi: 10.21831/cp.v43i1.60321.

[21] O. Bojar *et al*., "Findings of the 2018 conference on machine translation (WMT18)," in *Proc. 3rd Conf. Mach. Transl. (WMT)*, 2018, pp. 272–303, doi: 10.18653/v1/W18-64028.