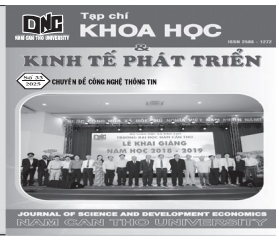
Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn



Build an open link database of Mekong Delta tourism information by extracting data from website and Wikipedia

Do Thi Bich Tram[1*], Tran Van Thoai[1], Tran Van Thien[1]

[1]*Nam Can Tho University*

[*]*Corresponding author: Do Thi Bich Tram (email: dotram210602@gmail.com)*

Received: 8/1/2025
Revised: 3/2/2025
Accepted: 15/2/2025

Keywords: *data mining, open link data, ontology, semantic web, travel data warehouse*

Từ khóa: *dữ liệu liên kết mở, kho dữ liệu du lịch, trích xuất dữ liệu, web ngữ nghĩa*

ABSTRACT

In today's era of development, tourism has become one of the most thriving and advanced industries globally. Numerous travel information search apps have been developed to assist travelers in preparing and planning for their trips. However, tourism data is often scattered across various platforms and lacks semantic links. To address this issue, this paper proposes a method for building an open link database (LOD) for tourism information in the Mekong Delta, utilizing data from travel websites and Wikipedia. The data is extracted and linked using data mining techniques and resource description language (RDF). The research aims to enhance the retrieval of tourism information by implementing open link data technology, providing users with a comprehensive and easily accessible source of information. The results of this study not only benefit the tourism industry, but also have potential applications in fields such as cultural studies, ecology, and resource management.

TÓM TẮT

Trong thời đại phát triển ngày nay, du lịch là một trong những ngành công nghiệp phát triển và sôi động nhất trên toàn cầu. Rất nhiều ứng dụng tìm kiếm thông tin du lịch được xây dựng để hỗ trợ du khách, giúp họ chuẩn bị và lập kế hoạch cụ thể cho chuyến đi của mình. Dữ liệu du lịch rất đa dạng và phong phú, nhưng lại phân tán và nằm rải rác trên nhiều nền tảng khác nhau, đặc biệt là nội dung không được liên kết về mặt ngữ nghĩa. Bài báo này đề xuất phương pháp xây dựng cơ sở dữ liệu liên kết mở (LOD) cho thông tin du lịch

ở Đồng bằng sông Cửu Long, sử dụng dữ liệu từ các trang web du lịch và Wikipedia. Dữ liệu được trích xuất và liên kết bằng các kỹ thuật khai thác dữ liệu và ngôn ngữ mô tả tài nguyên (RDF). Nghiên cứu nhằm mục đích cải thiện việc truy xuất thông tin du lịch thông qua việc ứng dụng công nghệ dữ liệu liên kết mở, tạo ra nguồn thông tin phong phú và dễ tiếp cận cho người dùng. Kết quả không chỉ có lợi cho ngành du lịch mà còn có thể ứng dụng trong các lĩnh vực như nghiên cứu văn hóa, sinh thái và quản lý tài nguyên.

1. INTRODUCTION

1.1 Mekong Delta travel data

Tourism is a key industry for Vietnam's economic development, the number of domestic and foreign tourists has continuously increased sharply in recent years, especially in the Mekong Delta. It is one of the large and tourism-rich deltas, with a diversity of culture, history and natural landscapes, attracting a lot of domestic and foreign tourists. But current tourism information is still scattered and lacks coherence. The construction of an linked open data warehouse can help centralize and connect different sources of information, creating a richer data ecosystem that is convenient for both visitors and researchers.

1.2 Reason for choosing the topic

A huge treasure trove of data is currently distributed across various platforms, including websites, Wikipedia, and social networks. This data not only brings practical value to businesses and organizations, but also makes an important contribution to improving daily life through smart applications and personalized services in the Mekong Delta region. However, the management and exploitation of tourism information in this region still faces many difficulties due to the fragmentation and lack of synchronization in data storage and sharing. To overcome these problems, this study proposes to build an linked open data system, with the ability to integrate and access tourism information more easily and efficiently.

1.3 Research objectives

The goal of this study is to collect and extract data from public sources such as websites and Wikipedia, and then integrate them into the LOD database. Next is to build an linked open data warehouse system for tourism information in the Mekong Delta, in order to create a synchronous, integrated and easily accessible information platform.

1.4 Theoretical basis

Previous research on ontology matching has employed various methods to align entities between ontologies. For example, [12] utilized a method called TEXTO (TEXT-based Ontology Matching). This approach focuses on using the textual descriptions of classes within ontologies for matching, rather than relying on the structural information of the ontologies.

1.4.1 Ontology

Ontology is a representation of knowledge that allows information to be shared between applications and is important for the semantic web [5]. In other words, ontology provides a clear

and structured semantic model for representing knowledge in a given field [3]. It is a powerful tool in the field of artificial intelligence and information technology, helping to organize and classify information in a systematic and structured way.

Ontology can be built using two methods: manual or automated. Both of these methods go through four main steps: analyzing information, designing concepts and relationships, and then implementing them in specialized language.

The components of Ontology include [4]:

- Individuals - Representation: Represents specific objects or fields
- Classes - Concept: considered as groups, sets of abstract objects. These classes can contain individuals, other classes, or a combination of both.
- Properties: used to indicate the relationship between entities (object attributes) or between entities and the data type described (data type attributes).
- Relationships: Used to determine the relevance of one object to another object using corresponding attributes.

1.4.2 Linked Open Data (LOD)

Linked Open Data (LOD) is a set of design principles for connecting and sharing data on the web using open standards [10]. LOD allows data from different sources to be linked, Creating a global data network that allows data to be connected and reused across various contexts [1]. Open Link Data (LOD) is an important part of the Semantic Web, allowing data to be connected and reused across a variety of contexts. Numerous studies have demonstrated the potential of LOD in increasing data access and integration, especially in the travel industry, where information from multiple sources needs to be aggregated coherently. LOD allows data from different sources to be linked together, creating a global network of data that can be efficiently queried and mined [1]. Linked Open Data brings many benefits in data management and sharing [6] such as: Enhancing data access and discovery; Enhancing data linkage and integration; Enhanced reusability and scalability; Openness and retention; Interoperability and linkage. In order to help data be linked and shared on the web effectively, facilitating easier discovery and use of information. Tim Berners-Lee has identified four basic principles for Linked Open Data including:

- Use URIs (Uniform Resource Identifiers) as names for objects.
- Use HTTP URIs so people can look them up.
- Provides useful information when someone looks up a URI.
- Include links to other URIs so people can explore more.

LOD is not merely a data connection technique, but also a supporting philosophy for the development of semantic web and open data ecosystems, bringing many significant benefits to the user community and application developers. There are many important advantages in data management and sharing:

- Data connectivity and integration: LOD allows data from various sources to be connected into an information network, enhancing data search and analysis capabilities.
- Standardization and Homogenization: LOD uses open standards such as RDF (Resource

Description Framework) and SPARQL (SPARQL Protocol and RDF Query Language), which help ensure consistency and ease of use of data.

- Easy access and use: Thanks to open protocols, LOD provides a flexible way to access and use data from different sources efficiently, making it possible for users to reuse data quickly and conveniently.
- Ensure consistency and reliability: LODs often come with guidelines and norms for data modelling and representation, helping to ensure the consistency and reliability of data when shared and used.
- Expand the searchability of information: LODs facilitate the enhancement of information searchability by linking data together in streamlined relationships, which improves the efficiency and accuracy of search services.

Structure of Linked Open Data in Mekong Delta data extraction: Ontology is developed to consist of 45 classes, 1 Object Properties, and 23 Data Properties, ensuring that it provides enough information to assist users in finding and selecting suitable tourist destinations.

The program consists of 5 main classes:

- luru_trú: Including information about accommodation services during travel such as motels, hotels, homestays and other diverse types of resorts, in the provinces in the Mekong Delta. Visitors can choose a place to stay that suits their needs and budget.
- mua_sắm: Contains information and addresses related to people's shopping needs such as ATMs, markets, stores,... in the provinces of the Mekong Delta.
- vận_chuyển: Including transportation services in the provinces of the Mekong Delta region. Make it easy for tourists to move from one point to another to explore the rich and diverse land
- ăn_uống: Contains basic information about restaurant and cafe services,... in the provinces located in the Mekong Delta region. Catering to tourists from rich local dishes to diverse international dishes, all here to meet the needs of tourists.
- điểm_đến: Contains information about entertainment places such as beaches, pagodas, mountains, islands,... In the Mekong Delta region, one of the rich lands with diverse landscapes and Vietnamese culture.

1.4.3 Web Semantic

The semantic web is an extension of a website such as adding semantic elements, aiming to help computers better exploit the information and data on the website. On the Semantic web, the resources provided will ensure semantic accuracy and flexibility so that computers and humans can collaborate more smoothly and efficiently.

RDF (Resource Description Framework): RDF is a labeled directional graph format used to represent information on the web [9], which plays an extremely important role in building an LOD, especially in the context of extracting and organizing travel information from various sources such as websites and Wikipedia.RDF is used as the most general method for conceptual descriptions or modeling of information interpreted in web resources, using in various syntax formats. RDF allows linking pieces of data from a variety of sources, such as from a travel website and a Wikipedia article, by using URIs (Uniform Resource Identifiers) to identify

entities. This helps to create a consistent data warehouse and can connect relevant information from heterogeneous sources. The data stored in RDF can be easily queried using the SPARQL query language. This allows complex queries to be performed to efficiently extract information from the open federated data warehouse.

RDF is built on basic website standards such as XML and URLs. Developed by the W3C (World Wide Web Consortium) to ensure the ability to link and exchange data on the web. At the same time, the "set of 3 elements" model is used: Each triad consists of three components: Subject, Predicate, and Object. This structure is like a simple sentence in natural language, where the subject performs an action (predicate) on the object.

SPARQL (SPARQL Protocol and RDF Query Language): SPARQL stands for SPARQL Protocol and RDF Query Language, which is an important query language in the Semantic Web and plays a key role in building an open federated data warehouse. It enables efficient and flexible querying, mining, and integration of RDF data, regardless of whether the data is stored natively as RDF or transformed via middleware [13]. SPARQL supports the combination of information from a variety of data sources, creating a unified and comprehensive view of open link data. In addition, SPARQL provides query optimization tools, which improve performance when handling large and complex data warehouses, with features such as pagination and conditional expression optimization. As an open and widely supported standard in RDF data management systems, this makes it simple and easy to integrate an open federated data warehouse with other applications and services.

OWL (Web Ontology Language): OWL is part of the Semantic Web and is the latest ontology language standardized by the World Wide Web Consortium (W3C). OWL is built on top of RDF and RDFS, providing additional vocabulary for defining classes and relationships. However, RDFS has many limitations, especially in describing resources in detail due to the lack of local constraints to define scopes and domains, as well as not supporting complex data types but only basic data types such as strings, Numbers and Boolean. OWL is considered an important technique in installing the Semantic Web in the future. In addition, OWL is designed specifically to provide a wide range of powerful features for describing and representing complex knowledge on the Semantic Web. The OWL language is expected to make computer systems readable and capable of replacing humans. Because OWL is designed based on XML, the data information in OWL can be easily exchanged with other computer systems, can use different languages and operating systems. The main important function of OWL is to provide standards from which to create an underlying platform for asset management, enterprise-level integration, and for the purpose of sharing and reusing data on the Web.

2. RESEARCH METHODS

2.1 Processing model

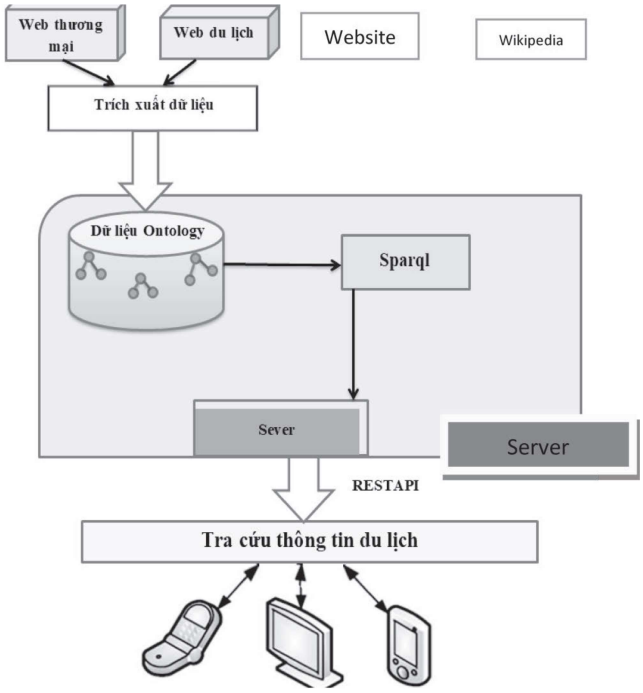


Figure 1. Processing model of the system

The first is data collection: Use data collection tools to get information from both Wikipedia and websites. For Wikipedia, data collection focuses on articles and link structures. For travel websites, destinations, reviews, and service information are the main goals.

Next is analysis and extraction: Apply text analysis tools to extract concepts, entities, and relationships from the collected data. For Wikipedia, the focus is on identifying the associated entities and how they are described. Whereas with websites, this process focuses on identifying service information, and the relationships between them.

Conversion: The extracted data is converted into an Ontology format, where entities and relationships from both Wikipedia and Web sites are represented as graphs. This helps to create a comprehensive and federated data structure. The data is then stored in an ontology database. This database contains standardized concepts, entities, and relationships, allowing for easy query and reuse in other applications.

SPARQL (SPARQL Protocol and RDF Query Language) is used to query ontology databases. This process allows specific information to be retrieved from the database based on query requests, which makes it possible to fully exploit information from both Wikipedia and websites.

Ontology is created from extracted data that can be used anywhere and on any type of device. It is possible to connect directly to Ontology or build a server as in the model.

- Server: Process SPARQL queries, prepare data, and perform the necessary operations to output the results.
- RESTful API: Used to make processed data easily accessible. This API allows different devices to send requests and receive information from the server efficiently.
- The data has been processed and the query will be available for lookup from various types of devices. Users can send queries through the interface on a laptop, desktop, or mobile phone and receive results directly from the system.

2.2 Extract data from travel websites

2.2.1 Libraries used

Selenium Library: Selenium is an open-source suite of tools designed to automate web browsers. This library is extremely useful for extracting data from travel websites, especially those with complex interactive and dynamic content. By using Selenium, developers and researchers can automate the data collection process, save time and effort, and ensure the accuracy and completeness of the extracted data [7]. Selenium offers benefits such as.

- Dynamic Content Processing: Selenium can handle websites with dynamic content loaded by JavaScript well, where other simple data extraction tools may struggle.
- Multi-Platform and Browser: Supports multiple browsers and operating systems, allowing data to be tested and extracted across a variety of environments.
- High Flexibility: Easy integration with programming languages such as Python, Java, C#, and many other frameworks to extend functionality and data processing.

Scrapy is one of the most popular tools for web scraping and web crawling, designed to help users extract data efficiently and flexibly from websites. Written in Python, it provides the tools and frameworks needed to build effective spiders, which automates the data collection process, saves time and effort, and ensures the accuracy and completeness of the data. By using Scrapy, developers and researchers can collect and process large amounts of data from travel websites for analysis and application in a variety of fields [8].

Benefits of using Scrapy as:

- High Performance: Scrapy is designed to collect data quickly and efficiently, optimizing the crawling and scraping process, Scrapy allows multiple HTTP requests to be made simultaneously, which speeds up the crawling process.
- Easy to scale and customize: Scrapy can handle websites with dynamic content generated by JavaScript, which other simple data extraction tools may struggle with.
- Data Integration Support: Scrapy supports exporting data to various formats such as JSON, CSV, XML, making it easy to integrate with other data analysis systems. Ability to manage and process large volumes of data from a variety of websites, suitable for large-scale data collection projects.
- Complex Content Processing: Scrapy can handle complex structured websites and dynamic content through middleware and add-ons.

2.2.2 The process of collecting and processing data from the website

Use powerful tools and libraries like Scrapy and Selenium to extract valuable travel information from a variety of web sources, including travel websites, social media, and travel review sites. The process of extracting travel data from a website includes key steps to ensure accurate and efficient information collection. First, the identification of the data source is done by selecting travel websites with the necessary information and identifying the types of data to be collected such as destination names, descriptions, addresses, reviews, and prices. The next step is to use a tool such as the browser's Developer Tools to view the HTML structure and find the location of the data to be extracted through exploring HTML and CSS to identify the elements that contain the necessary data. After identifying the HTML elements, the data extraction tool is suitable such as a programming language such as Python with libraries such as selenium to send HTTP requests to the website. And libraries like Scrapy for HTML analysis and information extraction. Once the data is collected, it will be processed to remove errors, duplicates, and unnecessary information and save it to a JSON file. Processed data is temporarily stored in

files before being imported into an open-link data warehouse.

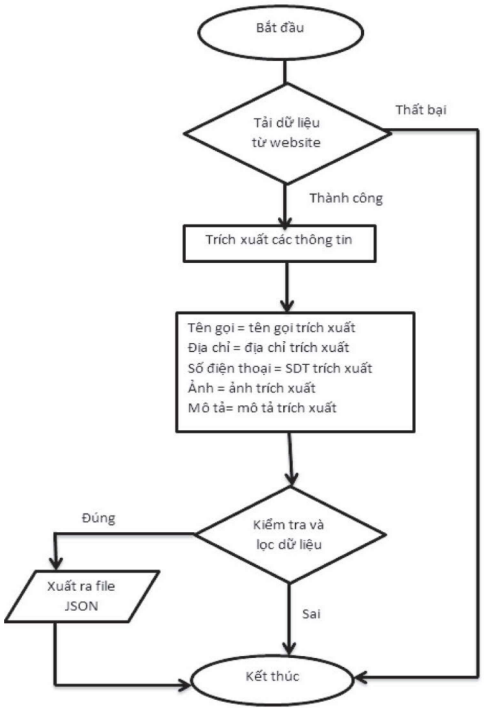


Figure 2. Description of the algorithm for extracting data from the website

Extraction algorithm example:

```
ten = response.xpath('//div[@class="verticleilist listing-shot"]/div[@class="container-fluid"]//h4/a/text()').get()
item['ten_goi']=ten
diachi=response.xpath('//i[@class="fafa-map-marker"]/following-sibling::text()').get()
 if (diachi):
        diachi = diachi.strip()
        diachi = re.sub(r'\s+', ' ',diachi)
item['dia_chi'] =diachi
sodienthoai=response.xpath('//i[@class="fafa-phone"]/following-sibling::text()').get()
if (sodienthoai):
        sodienthoai = sodienthoai.strip()
        sodienthoai = re.sub(r'\s+', ' ', sodienthoai)
item['so_dien_thoai'] = sodienthoai
trangweb = response.url
item['trang_web'] = trangweb
gia = 'tùy phòng'
item['gia'] = gia
soluongdanhgia = 'không có'
item['so_luong_danh_gia'] = soluongdanhgia
hangsao = response.xpath('//div[@class="col-md-8 h-100 cslt-detail"]/span[1]/img/@src').get()
```

Example results obtained:

- Name: Nhà hàng Lion City
- Address: 09 Nam Kỳ Khởi Nghĩa, Phường Tân An, Quận Ninh Kiều, Thành phố Cần Thơ
- Phone number: 0292 0914 781 313
- Website: https://csdl.vietnamtourism.gov.vn/rest/?item=2595
- Price: tùy món
- Number of Reviews: không có
- Assess: không có
- Image:http://csdl.vietnamtourism.gov.vn/uploads/logo/01_3/CSDLNHAHANG2021/TPCanTho/LionCity/lion5.jpg
- Describe: null

2.3 Extract data from Wikipedia

Use powerful tools and libraries like SPARQLWrapper to extract Wikipedia travel information. Build an algorithm to extract data from Dbpedia: Use the SPARQLWrapper library in Python to make a connection to Dbpedia's federated data warehouse and then execute the SPARQL statement to query and retrieve the required list of data. Finally, save the results obtained to a json file as the output of the destination to take advantage of for later steps.

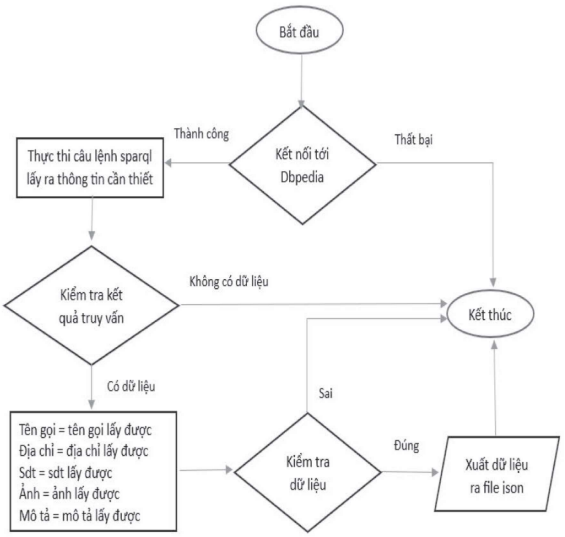


Figure 3. Description of the algorithm for extracting data from Wikipedia

Example of a data extraction sparql statement:

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?subject ?name ?mota ?anh

WHERE {

?subject rdf:type dbo:Restaurant .

?subject rdfs:label ?name .

?subject dbo:abstract ?mota .

?subject dbo:thumbnail ?anh .}

Example results obtained:

- Name: Mì cay 1992.
- Address: 198 Nguyễn Ngọc Sanh khóm 3, Phường 6, Thành phố cà Mau, Cà Mau.
- Phone number: null
- Website: https://csdl.vietnamtourism.gov.vn/rest/?item=2116
- Price: null
- Number of Reviews: không có
- Assess: không có
- Image:http://csdl.vietnamtourism.gov.vn/uploads/logo/01_3/CSDLNHAHANG2020/CAMAU/1992/mi-cay-1992-340466.jpg
- Describe: Hiện nay mì cay đang dần "phủ sóng" ở khắp các tỉnh thành trên cả nước,…

3. RESULTS AND DISCUSSION

3.1 Results of the data warehouse construction process

The process of extracting data from travel websites and Wikipedia has yielded significant results, providing insight into travel destinations and a wealth of relevant information. The organization and standardization of data helps improve information retrieval and analysis, supporting research and applications in the tourism sector in the Mekong Delta region. An open link database has been built with more than 49,000 data points collected from travel websites and Wikipedia. Includes information about landmarks, restaurants, hotels, transportation, markets, restaurants,... However, there are still some challenges in maintaining the up-to-date and accurate data.

In Vietnam, the use of Wikipedia compared to other websites has a marked difference in popularity and application. Although Wikipedia is a valuable resource with rich and diverse information, the use of this platform in Vietnam is not widely available compared to other countries. One of the main reasons is the lack of familiarity and widespread acceptance of Wikipedia by users in the user community in Vietnam. This leads to a much more limited amount of information on the Wikipedia platform than the Website, namely the extraction of tourism information in the Mekong Delta using

Wikipedia is more than 1000 elements, but the information extracted from the Website is up to more than 47000 elements.

Below are the results before and after adding the data:

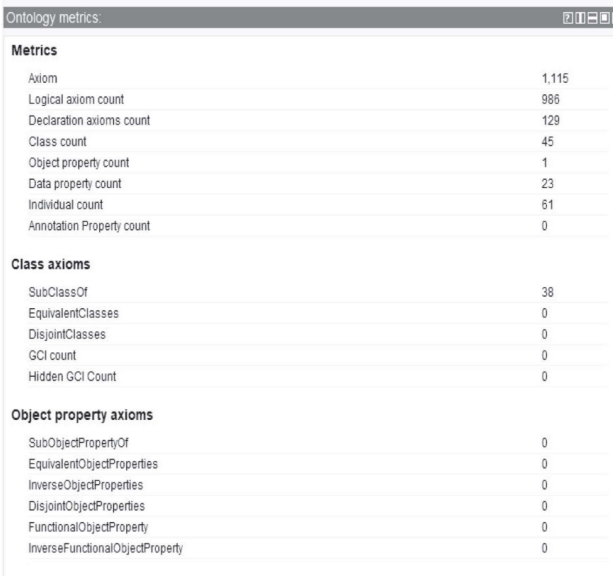


Figure 4. Extracted using Wikipedia

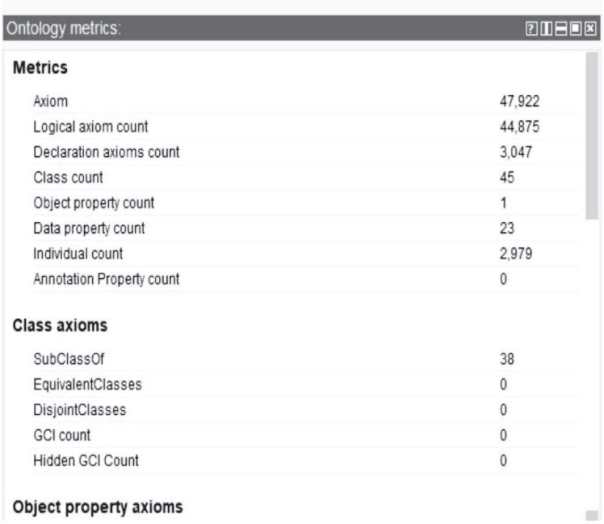


Figure 5. Extract by website

3.2 Benefits, limitations, and application potential of the open link data warehouse

The ontology project on travel data is a very interesting and promising topic, which is an important step forward in improving the way data is managed and mined. The project will bring practical benefits and contribute to promoting the sustainable development of the tourism industry.

3.2.1 Benefits of an Open Link Data warehouse

The project has succeeded in creating an open federated data repository. This helps to structure and standardize data, making it easier to manage and retrieve information. The Open Link Database provides a rich and easily accessible source of information for visitors and researchers alike. Including important entities and relationships in the field of tourism such as destinations, hotels, means of transportation, entertainment activities and support services... Supporting the development of smart travel applications, analyzing travel data, and creating customized services for tourism. Ontology was developed to help different travel information systems communicate and exchange data effectively, minimize incompatibilities, optimize the process of integrating data from multiple sources, and enhance data interoperability. Travel data can be analyzed in greater depth and detail, helping to make accurate business decisions and marketing strategies.

3.2.2 Limitations of Open Link Data warehouse

The data in an linked open data warehouse often comes from a variety of sources with heterogeneous quality standards, which can lead to data being inaccurate, incomplete, or not up to date. To improve the reliability of data, it is necessary to implement quality control mechanisms and update them regularly. Open-linked data repositories often use a variety of vocabulary and ontology, making it difficult to integrate data from sources without standardization and synchronization. The application of standard vocabulary and ontology such as schema.org or OWL is essential to ensure

compatibility and convenience in data integration [11]. The process of building, maintaining, and expanding open federated data warehouses requires a large amount of resources and costs, including the cost of data storage, processing, and management. Therefore, it is necessary to have effective plans and strategies for managing costs and resources, such as optimizing processes and adopting advanced, effective technologies.

3.2.3 The potential of LOD applications in other fields

LOD can create diverse repositories of learning materials, linking resources from many universities, research institutes, and libraries around the globe [2]. This makes it easy for students and researchers to access and integrate resources for learning and research. By linking data from medical facilities, scientific research, and patient data, LOD can assist doctors and researchers in analyzing data, diagnosing diseases, and providing optimal treatments. LODs can assist in the creation of personalized advertising and marketing campaigns based on customer preferences and behaviors. Linking data from sources such as social media, commerce websites, and mobile applications helps businesses better understand customer needs and make relevant product recommendations. By harnessing the potential of linked open data warehouses in different fields, we can improve service quality, enhance collaboration, and create new value from rich and diverse data sources.

4. CONCLUSION

One of the key outcomes of the project was the development of an automated ontology system from Wikipedia and travel websites. Successful construction of travel ontology from Wikipedia and travel website data: Automatically extract and convert information from Wikipedia and travel websites into concepts, attributes, and structured relationships in ontology. The construction of a tourism open link data repository for the Mekong Delta through data extraction from websites and Wikipedia has created a valuable resource for both users and travel applications, forming a rich and diverse data store, provide detailed information about tourist attractions, infrastructure, and services in this area. This process not only provides a rich data platform but also illustrates the ability of LOD technology to integrate and enrich local tourism information. Despite the positive results, the process of building a LOD data warehouse still faces several challenges, including managing large volumes of data, ensuring synchronization and consistency, along with cost and resource issues. To overcome these challenges, further research and development should focus on optimizing data collection and processing and harnessing advanced technologies to improve efficiency. This LOD database has significant potential to support sustainable tourism development in the Mekong Delta, and provide a foundation for innovative research and applications based on open data. The expansion and maintenance of this database will not only improve access to tourism information but also contribute to the overall development of the region. The linked open data warehouse not only improves information retrieval but also opens up new opportunities for the development of smart travel services and applications in the future. The LOD address is located at: http://mekongcrm.com/mekonglod.owl. We can

connect to Ontology using the Protege app through the http://mekongcrm.com/mekongtour.owl link that has been uploaded to the internet.

REFERENCES

[1] Lê Trung Nghĩa (2017). *Dữ liệu Mở Liên kết 5 sao là gì?,* 11/04/2017. https://letrungnghia.mangvn.org/Education/du-lieu-mo-lien-ket-5-sao-la-gi-5680.html

[2] Nguyen Andie (2022). *Dữ liệu liên kết mở (LOD)*, saltlux, 13/05/2022. https://saltlux.vn/kien-thuc-cong-nghe/du-lieu-lien-ket-mo-lod

[3] Trần Công Án, Tống Thị Ngọc Mai, Lê Thị Thu Lan (2017). Xây dựng ontology tự động từ bảng chú giải. *Tạp chí Khoa học Đại học Cần Thơ*, 2017. https://sj.ctu.edu.vn/ql/docgia/tacgia-14731/baibao-50441/doi-ctu.jsi.2017.018.html

[4] Le Duc (2019). *Tìm hiểu về Semantic Annotation - Phần 3: Ontology và RDF, viblo,* 19/11/2019. https://viblo.asia/p/tim-hieu-ve-semantic-annotation-phan-3-ontology-va-rdf-GrLZDO1wKk0

[5] Nguyễn Hữu Dũng (2021). *Semantic Web là gì? Cấu trúc và lợi ích mà Semantic Web mang lại*, bizfly, 30/08/2021. https://bizfly.vn/techblog/semantic-web-la-gi.html

[6] To Tee Te (2023). *Dữ liệu mở (Open data) là gì? Mang lại những lợi ích gì?* Thegioididong, 21/02/2023. https://www.thegioididong.com/hoi-dap/du-lieu-mo-open-data-la-gi-mang-lai-nhung-loi-ich-gi-1327942

[7] Cao Lê Viết Tiến (2024). *Selenium là gì? Những điều cần biết về công cụ Selenium Automation Testing. V*ietnix, 25/04/2024. https://vietnix.vn/selenium-la-gi/

[8] Thiên Lam (2020). *Cách tạo trình thu thập dữ liệu web cơ bản với Scrapy. Q*uantrimang, 15/02/2020. https://quantrimang.com/cong-nghe/tao-trinh-thu-thap-du-lieu-web-voi-scrapy-169701

[9] W3C (2013). *SPARQL Query Language.* W3C, 2013. https://www.w3.org/TR/sparql11-query/

[10] Ontotext (2016). *What Are Linked Data and Linked Open Data?* Ontotext, 24/06/2016. https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data

[11] Tim Berners-Lee (2008). *Linked data planet, w3, 2008.* https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data

[12] Peng, Y., Alam, M., & Bonald, T. (2023). Ontology Matching using Textual Class Descriptions. In Proceedings of the 18th International Workshop on Ontology Matching (pp. 1-10). *CEUR Workshop Proceedings.* https://ceur-ws.org/Vol-3591/om2023_STpaper2.pdf?utm_source=chatgpt.com

[13] Jarrar, M. (2017). *SPARQL RDF Query Language.* https://www.jarrar.info/courses/WebData/Jarar.LectureNotes.SPARQL.pdf. [Accessed: Feb. 21, 2025].