ABSTRACT: This paper focuses on introducing an ontological dataset for presenting tourist resources, which presently contains more than 100,000 places in big cities of Vietnam. This information is extracted and integrated from various relational datasets stored in many different social media or location-based websites such as Foody, Agoda, Mytour, etc. These enormous and heterogeneous data would be converted into a semantic knowledge base presented in RDF/XML format that is accessible by SPARQL query. Concentrating in studying the concept of linked open data with its publishing processes, this work adjusts and improves above repository to become a standardized linked open data, allows it connect with other famous ontology such as DBpedia.

Keywords: linked open data, ontology, semantic web, information extraction

I. INTRODUCTION

Since the concept of Semantic Web of Tim Berners Lee was introduced in 2001 [1], the Web has seen the growth of linked data websites that allow the creation of a global-scale interlinked data space, known as the Web of Data, by exposing datasets previously isolated as data graphs, which can be interlinked and integrated with other datasets. The purpose of Web of data is to connect object’s concepts and contents to each other, instead of simply connecting documents. Thus Web of data has led to the conversion of existing documents into linked data, and to the creation of new datasets. In addition, new data models have been implemented to represent data in a common standard way. The most famous data models are the Resource Description Framework (RDF) describing entities and the Web Ontology Model (OWL) (W3C, 2012) concepts.

The principles of Semantic Web have been applied to different fields of knowledge, spanning from cultural heritage to health. Tourism is one of many industries that has benefited enormously from the internet, and been explored in various recommender systems and mobile applications for instance Expedia, Yelp, etc. However, tourism datasets are mostly presented in relational databases. Moreover, these data contain many duplications from different sources so they are heterogeneous and difficult for many systems to search information about a tourist object in the relationship with the others. On the other hand, there are researches shown that semantic web brings many benefits if applied in tourism and heritage fields. Unfortunately, the semantic data about tourist are limited and it is a costly and time-consuming process of manually building a linked dataset even with skilled engineers and experts. There are two researches conducted by our Semantic Innovation Group focusing on this topic which are the Semantic Tourist informAtion Access and Recommending System called STAAR and the Vietnamese Tourism semantic data innovation system called ViSIS. While STAAR exploits the advantages of semantic web technology to help people find relevant information for their trips through Web and smart phone [2], ViSIS develops a semi-automatic environment for generating and enriching the tourism information presented in a semantic linked data repository [3].

Vietnam Tourism Innovation Ontology, VTIO, is a linked data which is the important base and used by both these two systems. However, even though this ontology has contained nearly 2 million triples, it still has been at an initial state, which is a local repository which has not been verified and published to contribute to the global data. Therefore it has been our motivation in this research to study about linked open data and also Vietnamese tourism semantic data innovation system with its ontology so that we could make some improvement for both the system and the tourism ontology VTIO to increase the semantic quality of the ViSIS to be compatible with international information sources and improve and publish the repository on the Web as a linked open data to be able to join the global semantic datasets with every famous linked open data such as DBpedia or ACCO.

II. RELATED WORKS

2.1. Dbpedia

The most well-known linked dataset, DBpedia [4] is available in different languages. Its English version contains about 4.0 million things, classified in different categories, including people, places, creative works, organizations, species and diseases. The DBpedia knowledge base has several advantages over existing knowledge bases: it covers many domains; it represents real community agreement; it automatically evolves as Wikipedia changes, and it is truly multilingual. However, DBpedia, as well as Wikipedia, contains only a small number of things related to
the tourism domain, such as accommodations and restaurants. In addition, to the best of our knowledge, only few linked datasets have been implemented in the field of tourism.

2.2. Tourpedia

There are also few initiatives of tourism linked data. A concrete example is Tourpedia, which combines and aggregates data extracted from four social media: Facebook, Foursquare, Google Places and Booking.com. Tourpedia contains almost half a million places, divided in four categories: accommodations, restaurants, points of interests (POIs) and attractions. Tourpedia was developed within the OpeNER Project [5]. The main objective of OpeNER is to provide a set of ready-to-use modules for the natural language processing. More specifically, OpeNER focuses on building linguistic pipelines in six languages (English, Spanish, German, French, Italian, and Dutch) that enable the identification and disambiguation of named entities and the analysis of sentiment in opinionated texts.

2.3. D2R Server

D2R Server is a system for converting content of relational databases to a linked data repository on the Semantic Web. Data from relational database would be modelled and stores in RDF format. D2R Server uses a customizable D2RQ mapping to map database content into this format, and allows the RDF data to be browsed and searched – the two main access paradigms to the Semantic Web.

III. VTIO ONTOLOGY MODEL

There are hundreds of enormous datasets from various sources about tourism on the Internet. It is a huge challenge to process and extract useful information from these data if we use them directly for our client applications. The semantic web allows heterogeneous data integration from multiple sources. By using of shared vocabulary and meanings for terms with respect to other terms, different tourism data could be standardized. Inspired by this concept, the ontology VTIO has been constructed using RDF format [6] so that data are stored and managed as semantic annotations, provide a good for reasoning and classifying the various information and ensure the flexible and interoperable system architecture. This ontology model plays a vital role on the semantic system as it stores and provide the information for both ViSIS and STAAR system. The VTIO ontology’s node is defined based on OWL class owl:Thing[12]. Every class is a subclass of owl:thing. This ontology also defined some our own custom vocabularies to store information.

<table>
<thead>
<tr>
<th>Open Data Field</th>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>hasLabel</td>
<td>Name of the place</td>
</tr>
<tr>
<td>address</td>
<td>hasLocation</td>
<td>The value is the URI of the location triple of the place which contains address number value and street URI</td>
</tr>
<tr>
<td>media</td>
<td>hasMedia</td>
<td>The value is the URI to the media triple of the place which contains image links or video links.</td>
</tr>
<tr>
<td>latitude</td>
<td>hasLatitude</td>
<td>The latitude value in double</td>
</tr>
<tr>
<td>longitude</td>
<td>hasLongitude</td>
<td>The longitude value in double</td>
</tr>
<tr>
<td>contact</td>
<td>hasContact</td>
<td>The value is the URI to the contact triple of the place which contains phone number or email or website</td>
</tr>
<tr>
<td>Open time</td>
<td>opentime</td>
<td>The value is the URI to the open time triple of the place which contains open time value</td>
</tr>
<tr>
<td>Open time</td>
<td>closetime</td>
<td>The value is the URI to the open time triple of the place which contains open time value</td>
</tr>
<tr>
<td>description</td>
<td>hasAbstract</td>
<td>The description of the place</td>
</tr>
<tr>
<td>famous</td>
<td>isWellKnown</td>
<td>Is this place famous or not?</td>
</tr>
<tr>
<td></td>
<td>isPartOf</td>
<td>Relation between streets, districts and cities</td>
</tr>
<tr>
<td></td>
<td>hasValue</td>
<td>Contain value in string, double or integer…</td>
</tr>
</tbody>
</table>
Figure 1. VTIO Ontology model

The diagram illustrates how an object is presented in the ontology model. Its information would be stored in nodes with reference URI. Every node is connected by the defined and custom vocabularies in an RDF/XML format file. The information could be queried by SPARQL syntax.

IV. REMODEL THE VIETNAMESE TOURISM ONTOLOGY AND PUBLISH AS A LINKED OPEN DATA

As mentioned above, the linked data is a web of data which connects objects and entities from all over the world. The current ontology is still restrained. The objects are only connected with themselves using the defined vocabulary by our team. The ontology is still only used for STAAR or “Diem Den Viet” applications instead of being published as an open linked data to connect with other linked data in the world. In this paper, we have proposed a strategy to re-model the VTIO ontology so it could become an open linked data as its initial motivation.

4.1. The case of Tourpedia

Tourpedia is a linked open data contains information about tourism of eight cities: Amsterdam, Barcelona, Berlin, Dubai, London, Paris, Rome and Tuscany. Therefore, they have a same purpose with our ontology VTIO. Consequently, we have studied them as a concrete example to have an accurate knowledge of linked open data motivation and its publishing process. Our main objective is to research Tourpedia model to use it for our VTIO situation.

Tourpedia contains four main classes: accommodation, event, tourist attraction and meteo (TBD). They have developed their ontology using these vocabularies:

- schema.org for describing entities such as people, places, or businesses (e.g. hotels);
- Basic Geo is sometimes used for expressing geographical coordinates;
- RDF Schema
- Custom Vocabulary (Tourpedia ontology)
4.2. Remodel the Vietnamese tourism ontology

There are many ontologies that are famous and widely accessed. Each ontology has its own namespace and properties that could be used. To connect VTIO ontology with these linked data, we must have studied these ontologies structure and its vocabularies so we could choose the suitable namespace and URI to link with our ontology. Here is some popular namespace that we would use in the new model of the VTIO ontology.

**dbpedia-owl**: The dataset from this ontology is generated from the manually created specifications in the DBpedia Mappings Wiki. Each release of this ontology corresponds to a new release of the DBpedia data set which contains instance data extracted from the different language versions of Wikipedia.

**Acco** [7]: This Accommodation Ontology provides the additional vocabulary elements for describing hotel rooms, hotels, camping sites, and other forms of accommodations, their features, and modeling compound prices as frequently found in the tourism sector, e.g. weekly cleaning fees or extra charges for electricity in vacation homes based on metered usages. In common hotel situations, this ontology separates between the legal agent operating the hotel (gr:BusinessEntity), the hotel as a whole (acco:Hotel), and the individual hotel rooms (acco:HotelRoom). This distinction is important, because we may want to represent hotel features and room features. For instance, a Wifi can be a room feature or a hotel feature. Hotel rooms and hotels or camping pitches and camping sites can be linked via the acco:partOf relationship type.

The current ontology structure contains a lot of custom vocabularies and some of them are not necessary because there are many common properties that are widely used in every open ontology could replace its purpose. we would propose some alternative vocabularies in the new model. The new model for VTIO ontology has been appended some common vocabularies and namespace from other famous LOD:

<table>
<thead>
<tr>
<th>Open Data Field</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td>rdfs: type</td>
</tr>
<tr>
<td></td>
<td>uri : <a href="http://purl.org/acco/">http://purl.org/acco/</a></td>
</tr>
<tr>
<td>city</td>
<td>dbpedia-owl :Location</td>
</tr>
<tr>
<td></td>
<td>uri : <a href="http://dbpedia.org/resource/">http://dbpedia.org/resource/</a></td>
</tr>
<tr>
<td>Services</td>
<td>acco : feature</td>
</tr>
<tr>
<td>latitude</td>
<td>geo : lat</td>
</tr>
<tr>
<td>longtitude</td>
<td>geo : long</td>
</tr>
<tr>
<td>telephone</td>
<td>vcard : hasTelephone</td>
</tr>
<tr>
<td>External link</td>
<td>Dbpedia-owl : wikiPageExternalLink</td>
</tr>
</tbody>
</table>
To use the other namespaces, we have to declare at the start of the ontology:

```xml
<rdf:RDF xmlns="http://hust.se.vtio.owl#"
    xml:base="http://hust.se.vtio.owl"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:acco="http://purl.org/acco/ns#"
    xmlns:dbpedia="http://dbpedia.org/resource/
    xmlns:dbpedia-owl="http://dbpedia.org/ontology/
    xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#" >
```

V. PUBLISH AN OPEN LINKED DATA

5.1. Publish an open linked data processes

Naming things with URIs

Before publishing the repository on the Web, any object in a domain of interest must be defined. As a LOD builds directly on Web architecture, the Web architecture term resource is used to refer to these things of interest, which are, in turn, identified by HTTP URIs.

As discussed above, a linked open data uses only HTTP URIs which is the URI Reference in the RDF format. The reason to use a HTTP URIs to present an item name are: first, they provide a simple way to create globally unique names in a decentralized fashion, as every owner of a domain name, or delegate of the domain name owner, may create new URI references; secondly, they serve not just as a name but also as a means of accessing information describing the identified entity.

Generating the base URI is one of the most important steps of the semantic enrichment module in the ViSIS system. It is the header of every URI belongs to places. Because each place is defined as an URI so there must not have any duplication. The former system generates the URI based on this formula: $URI = name + number + street + district + datatype$

For example: Sofitel Plaza is at 1 Thanh Nien Street, Ba Dinh District, Ha Noi, so its media uri would be: #sofitel-plaza1-thanhnien-ba-dinh-ha-noi-media.

However, after study other exited semantic web, we found out that the URI format need to be change so that the semantic web would be more highly organized and accessible. The new format of the URI would be set as below:

$URI = http://vtio.tech/city_name#name + number + street + district + datatype$
For example: Sofitel Plaza is at 1 Thanh Nien Street, Ba Dinh District, Ha Noi, so its media uri would be changed into: http://vtio.tech/ha-noi#sofitel-plaza-1-thanh-nien-ba-dinh-ha-noi-media.

Providing useful RDF information

In order to enable a wide range of different applications to process Web content, it is important to agree on standardized content formats. The VTIO ontology has been presented using RDF format, which is one of the most common format for presenting a semantic linked data. The RDF format in VTIO is using XML syntax and generated from the ViSIS system. The information is stored in RDF statements and connected using the both common and custom vocabularies defined in the ontology. The VTIO structure has been verified after uploading into Allegrograph Server and would be validated by W3C validation tools after publishing on the Internet.

Including links to other things

One of the most important properties of a LOD is to point into other data sources on the Internet. External RDF links are the elemental part in the Web of Data as they are the glue that connects data nodes in a global scaled, interconnected data space and as they enable applications to discover additional data sources in a follow-your-nose fashion. Technically, an external RDF link is an RDF triple in which the subject of the triple is a URI reference in the namespace of one data set, while the predicate and/or object of the triple are URI references pointing into the namespaces of other data sets. There are three important types of RDF links: Relationship Links point at related things in other data sources, for instance, other people, places or genes. In the modified VTIO ontology model, the relationship links would be the DBpedia Ontology, which each place in the VTIO repository will point to the corresponding city entities in the DBpedia by the vocabulary dbpedia-owl:location. For example: The Ho Guom Lake object in the VTIO RDF would point to Hanoi object in the DBpedia with the statement:

```xml
<owl:Thing rdf:about="http://vtio.tech/ha-noi#ho-guom-hoan-kiem-ha-noi ">
< dbpedia-owl:location rdf:resource="http://dbpedia.org/resource/Hanoi "/>
</owl:Thing>
```

Identity Links point at URI aliases used by other data sources to identify the same real-world object or abstract concept. In our proposed VTIO model, besides having the attribute type which is the defined category within the ontology, each place would also be pointed to the type defined by other linked open data namespace such as ACCO. For example: <rdf:type rdf:resource="http://purl.org/acco/ns#Hotel">

Vocabulary Links point from data to the definitions of the vocabulary terms that are used to represent the data, as well as from these definitions to the definitions of related terms in other vocabularies. The vocabularies of famous namespaces such as DBpedia, ACCO, VCARD or Geo would be used in the ontology VTIO to present the information. These namespaces would be declared in the header of each RDF file.

These new configurations are modified directly in the semantic enrichment module of the ViSIS system. The generating linked data would not only contain the former information but also the extended attributes which has been described in the previous part.

Serving linked data

Producing static RDF files and uploading them to a Web server is probably the simplest way to publish Linked Data and is a common approach when a person creates and maintains relatively small RDF files manually, e.g., when publishing RDFS vocabularies or personal profiles in RDF or a software tool or process generates or exports RDF data as static files. The VTIO ontology would be upload on a public hosting that allows any developer to access and use this Vietnamese tourism linked data. The repository would be stored in both static RDF file and a HTML web view.

a) Testing and debugging linked data

After publishing the Linked Open Data on the web, the data and its infrastructure should be checked to ensure it adheres to the Linked Data principles and best practices. There are many tools and website that provide these kinds of services. A useful starting point for testing Linked Data is to check that RDF data conveys the intended information. The W3C RDF Validator [8] can check RDF/XML for syntactic correctness and provides tabular N-Triples-like output of validated triples that is useful for visual inspection, as described above. To verify the LOD is able to be linked and used by other LOD, we could use a SPARQL Endpoints to try query the information from the ontology, if every query returns accurate results, the LOD is ready to serve.

b) Convert RDF to RDFA + HTML

The VTIO tourism data is present in the RDF/XML format. When the RDF file is uploaded into the host server. Even though the information from linked datasets could be queried using SPARQL language, the ontology is still invisible for user to observe. Consequently, W3C has propose RDFA format which is one of several ways of writing
down, or serializing, RDF data. Specifically, RDFa enables RDF data to be embedded in HTML documents, which makes it very useful for publishing RDF in contexts where Web publishing is limited to HTML, for example where a legacy content management system prevents publication in other formats. As simply another serialization of RDF, RDFa is ideally suited to publishing Linked Data.

To generate the RDFa+HTML format from RDF files, we have use the API from Parrot [9], which is a RIF and OWL documentation service. “The Semantic Web contains a number of knowledge artifacts, including OWL ontologies, RIF rule sets and RDF datasets. Effective exchange and management of these artifacts demands the use of metadata and prompt availability of accurate reference documentation. In this article, we analyze the current practices in metadata usage for OWL ontologies, and we propose a vocabulary for annotating RIF rules. We also introduce a software tool --Parrot-- that exploits these annotations and produces reference documentation for combinations of ontologies and rules.” - Carlos Tejo-Alonso, Diego Berrueta, Luis Polo and Sergio Fernández, Parrot developer.

5.2. VTIO LOD validation

After publishing the VTIO ontology, the linked dataset has to be verified to check whether it is well-formatted and connectable to truly become a linked open data. To validate the ontology, we have using some validation tools provided by the semantic community which have been used to verify other published linked open data:

**W3C RDF Validation Service:** This RDF validation service was created and is maintained by Jeremy Carroll at HP-Labs in Bristol. The service would parse the rdf file to verify whether it is well-formatted or not and list all the triples in the rdf file.

In this test we have input the uri: [http://vtio.tech/quangnam/quangnam.rdf](http://vtio.tech/quangnam/quangnam.rdf) to verify the RDF format. The result shows that the services have successfully extracted the list of triples in the rdf file presented as a table of triples of the data model which has id, subject, predicate, object of triples.

The result report could be read at this link with 19821 triples found:

```
```

**W3C RDFa validator [10]:** This is validator is meant to validate RDFa content in an HTML or an XML file.

In this test we have input the uri: [http://vtio.tech](http://vtio.tech) to verify the RDFa format. The result shows the URI is well-formatted and the RDF information could be extracted from this URI. The report contains a validator message “Congratulations, your RDFa source is valid” and the generated RDF content in Turtle format.

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
```
@prefix parrot: <http://vocab.citic.es/parrot#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xhv: <http://www.w3.org/1999/xhtml/vocab#>.

<http://vtio.tech/> a owl:Ontology.
<http://vtio.tech/#ATM> a owl:Class;
<http://vtio.tech/#Action-Movie> a owl:Class;
  rdfs:subClassOf <http://vtio.tech/#Movie>.
...
<http://vtio.tech/#Place> a owl:Class.

The result report could be read at this link:

SPARQL query: if a sparql query is success, it would prove that the linked open data is ready to be linked and queried by other systems. In this test, we have used the rdfa developer plugin of firefox browser with a sparql endpoint to verify the VTIO LOD website. For example: List all the triples in the RDF.

SPARQL query:
SELECT * WHERE { ?subject ?predicate ?object .}

This query has successfully returned 3081 triples in the base website http://vtio.tech/ and presents the results as a table contains subject, predicate and object of each triple.

![Figure 5. SPARQL result 1](image)

The results show that the VTIO ontology after published as a linked open data on a public hosting has successfully validated by three scenarios: RDF format verification, RDFa format verification and SPARQL query. The results conclude that the VTIO ontology is now a part of linked open data community and ready to be linked by other ontology. This would contribute the Vietnamese tourism information to the international web of semantic data.

VI. CONCLUSION

This work has studied the meaning and application of a semantic web as well as linked open data with the two semantic systems about Vietnamese tourism data, ViSIS and STAAR. The VTIO ontology used by both these systems has become our object of study to improve this repository as a linked open data with a better semantic quality that could connect with the Web of Data on the Internet as the first Vietnamese Tourism LOD. The procedure to update datasets is semi-automatically done by our system. This work has helped us to have a deep understanding about semantic web and applied it for practical applications. We have also gain some meaningful experience on how to manage a project and work with some international quality standard especially when publishing a linked open data. This would be our motivations for the future to pursue studying this topic and also continue to review and innovate the ontology model to contribute to the linked open datasets of the semantic web community.

VII. REFERENCES

TOWARD A LINKED OPEN DATA REPOSITORY ABOUT VIETNAMESE TOURISM

520


[9] Parrot, a RIF and OWL documentation service, ontorule-project.eu/parrot.


TỔM TÁT: Nghiên cứu này tập trung vào giới thiệu tập dữ liệu ontology biểu diễn các nguồn tài nguyên về du lịch với hơn 100.000 địa điểm từ các thành phố lớn ở Việt Nam. Thông tin được trích xuất và tích hợp từ những cơ sở dữ liệu quan hệ khác nhau của các trang mạng xã hội và website về địa điểm như Foody, Agoda, Mytour... Tập dữ liệu không hề thiếu dữ liệu quan trọng này sẽ được chuyển sang biểu diễn dữ liệu theo chuẩn RDF/XML và có thể được truy vấn bởi ngôn ngữ SPARQL. Bài báo tập trung vào tìm hiểu khái niệm về dữ liệu liên kết mở và quá trình xây dựng kho dữ liệu này để có thể xây dựng một kho dữ liệu liên kết mở được chuẩn hóa và có độ chính xác cao đặc biệt cho du lịch Việt Nam, góp phần làm tài nguyên cho các nghiên cứu khác và có thể để các cơ sở dữ liệu dùng ontology khác có thể liên kết đến, ví dụ như DBpedia.