Representing GeoData for Tourism with Schema.org

Oleksandra Panasiuk¹, Zaenal Akbar¹, Thibault Gerrier¹ and Dieter Fensel¹

¹University of Innsbruck, Technikerstrasse 21a, Innsbruck 6020, Austria
{firstname.lastname}@sti2.at

Keywords: semantic annotation, geographical data, schema.org, e-tourism

Abstract: A large amount of tourism data on the web, representing different touristic services, refers to information which is geographically located. With the intensive development of artificial intelligence, interest in the annotation of data is continuously increasing. It is therefore important to describe all tourist needs. To be understandable to search engines, chatbots or other personal assistant systems, content data should be structured, well-formed and semantically consistent. Schema.org is a de-facto standard for marking up structured data on the web. In this paper we show how to annotate geographical information related to different touristic services and activities (e.g., hotels, restaurants, events, hiking and climbing trails) available on an interactive map using the schema.org vocabulary.

1 Introduction

When visiting tourist destinations, especially new or big ones, tourists are typically challenged to decide the best plan on how to visit a number of points of interest with limited resources such as time and budget. This challenge can be seen as a tourist trip design problem (Vansteenwegen and Van Oudheusden, 2007), where the best route can be determined by combining user interests, tourist attraction, and trip information in respect to time and budget constraints. Different expert systems for trip planning have been developed to overcome the challenge, for example for a city trip (Vansteenwegen et al., 2011) and multi-modal tour (Gavalas et al., 2015). To make decisions, those systems built a knowledge base where points of interest’s data were obtained from tourists offices, tourist portals, or Wikipedia.

Various tourism related information can be found at numerous information sources including geographical data which are usually presented on maps. These data could be represented differently among systems, causing problems for data integration or data linking. A way to overcome this data representation disparity is through data annotation. Annotation can be seen as a process of connecting two pieces of information, typically with a purpose to give a better explanation of the data. Specifically for semantic annotation, the additional information is intended to connect data with its “meaning”, in a way so that it can be represented and accessed uniformly. Data that are available in distributed sources can be read and processed by machines automatically. Semantic annotation can be performed manually or (semi)-automatically for different purposes, for example to annotate digital music in such a way that a music item can be located easily according to users choice (Rahman and Siddiqi, 2012), improving search accuracy for web searches (Fuentes-Lorenzo et al., 2013), and including geographical data (e.g., points of interest) on a map (Ruta et al., 2012; Stadler et al., 2012).

Semantically annotated content impacts the tourism industry significantly. It feeds search engines structured data in a way that can be presented in a more interesting way visually, for example Google Search has rich features where events will be presented in a structured layout, or recipes in a carousel¹. It increases online visibility of a typical hotel’s website by up to 20% (Fensel et al., 2016). The annotated content can be consumed by intelligent applications, for example to guide a semi-automatic distribution of content to multiple online communication channels (Akbar et al., 2014). Latest technologies such as chatbots and personal digital assistants process annotated content to answer users requests in an automatic fashion.

In this paper, we present our work to annotate geodata (geographical data) which are related to the touristic domain in the region of Tyrol, Austria by using the Schema.org vocabulary. Our intention is clear,

¹https://developers.google.com/search/docs/guides/search-features
annotating geodata information will not only increase the region’s online visibility but also enable intelligent applications to consume it. We did not use any vocabulary that was developed specifically for geodata, because most of this vocabulary did not receive sufficient adoption. Instead, we use more common vocabulary that has been widely adopted where an extension will be proposed whenever necessary. To be precise, our contributions are: (i) domain specification for geodata, (ii) an approach to integrate annotation to a website, (iii) a significant amount of annotation of geodata of the Tyrol region. The paper will be organized as follow, Section 2 lists a few existing related works and discusses our contributions. Section 3 presents our analysis and defined specification for geodata, followed by our method to create annotations based on the defined specification in Section 4. Sections 5 and 6 discuss our results and their evaluation, and finally we conclude our work and list some potential future works in Section 7.

2 Related Works and Motivation

In this section, we outline our motivation for annotating geographical data and existing related works. The tourism industry cannot be separated from the development of the region of Tyrol, Austria. In 2015/2016, 11.5 million guests visited the region, producing 47.6 million overnight stays. One way to help the industry is by utilizing semantic web technologies, including semantic annotations (Akbar et al., 2017).

2.1 Motivation

Intelligent applications require a system that feeds them with knowledge in order to make decisions. A widely used system known as Knowledge Graph, popularized by Google, uses a graph database to provide structured and detailed information integrated from distributed sources. It has been used in various sectors, for example for sales assistance (Kim, 2017) and question answering system (Kumar et al., 2017). A Knowledge graph can be built from semantically annotated content, and therefore it is necessary to have a large amount of annotation to produce complete knowledge. According to the Web Data Com-

3https://en.wikipedia.org/wiki/LinkedData
4http://webdatacommons.org/structureddata/
5http://linkeddata.org/
6https://www.openstreetmap.org/
7https://en.wikipedia.org/wiki/LinkedData
mons in 2016 there were around 1.24 billion HTML pages containing structured data, which is about 38% of the total number of pages available on the Web. The number of Resource Description Format (RDF) triples has increased significantly: about 44.2 billion in 2016, compared to 24.4 billion in 2015.

Annotating geodata with specific vocabulary is not a new topic; there are a few existing works that were done with various purposes. The Linked-GeoData project transformed and published OpenStreetMap geographical data using Linked Data principles, enabling semantic-spatial search as well as geodata syndication (Stadler et al., 2012). Linking concepts detected in books to locations (spatial information) could produce a novel location-based recommendation system (De Meester et al., 2015). There are a few standards for digital geographic information, including ISO/TC 211.

2.2 Contribution

More than just structuring data in a uniform representation, the semantic annotations approach presents many advantages, including reasoning capability. This reasoning capability is vital when developing an intelligent application, for example to determine whether the description of a concept is more general than the description of another concept (subsumption of concepts), or to find all individuals that are instances of a concept (retrieval of individuals).

There are a few existing works that have been producing a large amount of annotation for touristic related information (Akbar et al., 2017), but none of them include geographical data yet. This work will enrich those existing annotations and provide a base for tourism related geographical data annotation.

Our work is taking advantage of the wide adoption of Schema.org, where classes in this vocabulary have been used more than 50% on average (Meusel et al., 2015). With the support of major search engines including Google and Bing, the rate of adoption is increasing. Therefore we are quite confident that our approach, annotations of geographical data with Schema.org, will receive the same rate of adoption.
3 Methodology of GeoData annotation

In this section we describe the methodology for annotating the touristic services based on their geographical data and discuss the main challenges that arise. As a pilot we use the maps of the Tourismusverband (TVB)\textsuperscript{9}, Mayrhofen-Hippach\textsuperscript{10} and Seefeld\textsuperscript{11}. The methodology consists of the following parts.

3.1 Data content and format analysis

Tourist Maps are built based on the information necessary for tourist needs and related activities. That’s why, for the annotation process, it is important to analyze the real world representation and to extract existing service types relevant in certain tourist areas according to tourist preferences (Gretzel et al., 2004). The next step is to analyze online representation of touristic services displayed on the maps, data content and its format. Very often data represent different types of information for the same touristic service, and are semi-structured or unstructured, which causes difficulties for the semi-automatic annotation process. The Interactive Maps of the TVB Mayrhofen-Hippach and Seefeld contain a variety of tourism-related information, including hiking or biking routes, accommodation, infrastructure and so on. Besides information about geodata objects, the maps also contain metadata information about points on the map, for example, a contact point for a store, the distance or altitude of a hiking route, or real-time bus schedules at a given bus stop. The content consists of the information from the Contwise Maps, an interactive map provided by General Solutions\textsuperscript{12}, and information from the external source Feratel\textsuperscript{13}.

There are two main data structures in the Contwise Maps: (i) Object Types (Forms) and (ii) Categories. The Object Types define the attributes of an object. Every resource (content entry) in the Contwise Maps can have only one object type. The Categories are the menu items and can be nested into a tree structure. They are completely independent from the object type. Resources of any object type can appear in one or more categories and subcategories.

All this information can be accessed through a web API which serves data in a JavaScript Object Notation (JSON) format.

3.2 Domain definition

In this step we construct the set of domains which we want to annotate. Each tourist object from the map belongs to an object type and categories as mentioned above and may have some subcategory. This is why our primary task is to define domains and select proper types from schema.org based on data structures in the Contwise Maps. We extract the list with object types, categories and subcategories and select a proper type from schema.org, as shown in Table 1. The main challenge here is that the data structures used by General Solutions and schema.org were built for different purposes, leading to discrepancies between their types and properties. Some categories and object types are too general or too detailed, some don’t have correct English matches, and some are not covered by the schema.org vocabulary.

3.3 Mapping to a domain specific subset of schema.org

From the domain definitions we get the set of different domains. The next step is to map the data content of these domains to schema.org in a way which suitably represents its types and properties. The main challenge here is to find the best and most suitable way to map the data, i.e. to choose the right class with properties from schema.org. For example, for a Cable Car we choose the type CivicStructure with the following properties: name, description, amenityFeature, contactPoint, image, address, geo, containsPlace, potentialAction, hasMap, and url. Each property has its range, e.g. Text, Url, DateTime, QuantitativeValue. Some elements can have external properties, and we considered them too. Geo in schema.org has

---

\textsuperscript{9}Destination Management Organization
\textsuperscript{10}https://maps.mayrhofen.at/
\textsuperscript{11}https://maps.seefeld.com/
\textsuperscript{12}https://general-solutions.eu
\textsuperscript{13}http://www.feratel.at/
the range GeoCoordinates, where GeoCoordinates is defined by longitude, latitude, elevation.

### 3.4 Domain specification

This section, based on previous steps, provides the common models for annotating different touristic domains presented on the map. For this purpose we analyze the results from sections 3.2 and 3.3 and choose the wide use domains. Then for each domain we model a domain specification, i.e. select or combine the most suitable defined properties and classes, including the range types. These domain specifications give us the patterns, that can help to automate the annotation process and make it easier (Panasiuk et al., 2018). In Table 2 an example of SportsActivityLocation domain specification is shown.

### 4 Implementation

In this section, we explain our implementation of the annotation for the content Interactive Maps of Contwise Maps from General Solution. The implementation consists of (i) getting data from the source, (ii) the mapping process from the input to structured data with the vocabulary from schema.org and (iii) annotation deployment. For implementation we use Node.js, as it is good for HTTP requests as well as manipulating data in the JSON format.

#### 4.1 Getting the data

We get the data via General Solution API. All the requests are simple HTTP GET requests and the results containing all of the data are in the JSON format, which has key value pairs. We obtain an array of JSON-objects for each category and the images for each individual resource. Figure 2 shows the process of getting data via General Solution API.

#### 4.2 Mapping process

The structure of JSON objects is as follows: a unique identifier, a name, a description, a field for geographic coordinates, an array of category ids and a "metadata" field containing different information such as address, phone number, and more specific information like the length of a trail, if the trail is suitable for children, the number of beds for a hotel and so on.

The first step of the mapping consists of finding out what schema type the resource should use based on its categories. For that, a list is made with every possible category id and its appropriate schema type. Those categories were often very vague and thus were followed by subcategories. The next step is to map those properties that can always be mapped the same way, those are: id, name, description, coordinates, fields concerning address (street address, postal code, ...), images and URLs. The structure of the program is shown in Figure 3 and can be summarized as follows: if the resource has this or that property then map it in this or that way. So, the program goes through all possible fields the data can have and maps it according to our previous mappings.

The result of this mapping is an array of JSON-LD objects using the schema.org vocabulary, where all meaningful data from the source was annotated as described by our mapping.
### Table 2: An example of the domain specification for SportsActivityLocation

<table>
<thead>
<tr>
<th>Domain Types</th>
<th>Set of properties</th>
<th>Range Type</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SportsActivityLocation</td>
<td>name</td>
<td>Text</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>description</td>
<td>Text</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>image</td>
<td>URL, ImageObject</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>hasMap</td>
<td>Map or URL</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>address</td>
<td>PostalAddress</td>
<td></td>
</tr>
<tr>
<td>PostalAddress</td>
<td>addressCountry</td>
<td>Text, Country</td>
<td></td>
</tr>
<tr>
<td></td>
<td>name</td>
<td>Text</td>
<td>optional</td>
</tr>
<tr>
<td></td>
<td>addressLocality</td>
<td>Text</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>addressRegion</td>
<td>Text</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>postalCode</td>
<td>Text</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>streetAddress</td>
<td>Text</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>geo</td>
<td>GeoCoordinates, GeoShape</td>
<td>required</td>
</tr>
<tr>
<td>GeoCoordinates</td>
<td>longitude</td>
<td>Number</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>latitude</td>
<td>Number</td>
<td>required</td>
</tr>
<tr>
<td></td>
<td>elevation</td>
<td>Number</td>
<td>optional</td>
</tr>
</tbody>
</table>

... 

---

4.3 Annotation deployment

The next step was getting those JSON-LD files into semantify.it. The platform has a nice API where you can upload a few hundred JSON files at a time to a given API-key (linked to an account or organization).

So, with just a few requests all the data is transmitted and ready for the customer. (Kärle et al., 2017)

5 Evaluation

We evaluate our annotation model using formal validation of the functional requirements of our domain specifications, including answering the competency questions (Uschold and Gruninger, 1996). We translate questions into queries in SPARQL, obtain the answers to them, and hence formally evaluate our ontology (Stolz et al., 2017). We use Ontotext GraphDB as a software product for storing and querying our annotations with SPARQL.

As explained in Section 2, the semantic annotations approach was utilized to not only represent data uniformly but also to obtain reasoning capability. Therefore, in the following SPARQL queries, we discuss the capabilities introduced by semantic annotations with Schema.org.

**Question 1.** Return all trails with average difficulty in Seefeld region.

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

SELECT ?s ?name WHERE {
  GRAPH ?g {
    ?s schema:name "Difficulty".
    ?feature schema:name ?name.
    FILTER(strafter(strbefore(?name,"Difficulty"),"avg") = "medium")
    FILTER(strafter(strbefore(?name,"Difficulty"),"avg") = "easy")
  }
}
```

14https://semantify.it/

15http://graphdb.ontotext.com/
According to the “SportsActivityLocation” specification of Schema.org\textsuperscript{16}, the type covers various more specific types such as “BowlingAlley”, “GolfCourse”, “PublicSwimmingPool”. By using this single query, instances of all those specific types will be also included in the result, in consequence of type and sub-type relationships.

**Question 2.** Return all cable cars and their length in ski resorts of Mayrhofen-Hippach map and with the altitude of mountain station more than 1815m.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX schema: <http://schema.org/>
PREFIX dbp: <http://dbpedia.org/resource/>
SELECT ?s ?name ?distance ?elevation
WHERE {
  GRAPH ?g {
    ?s rdf:type schema:CivicStructure.
    ?s schema:additionalType dbp:Cable_car.
    ?s schema:name ?name.
    ?action rdf:type schema:TravelAction.
    ?action schema:toLocation ?place.
    FILTER (?elevation >= 1815).
  }
  FILTER(str(?g)="http://maps.mayrhofen.at")
}
```

This second query requested instances of “CivicStructure” which have an additional type of “Cable_car” from DBpedia\textsuperscript{17}. This is another advantage of using semantic annotations with Schema.org, where data can be linked with specific types from external vocabularies, for instance to avoid ambiguity.

### 6 Results

In this section, we discuss results from our work of annotating geodata available in the maps of TVB Mayrhofen-Hippach and Seefeld. First, we describe the statistics of annotations produced, then discuss in detail entities that were annotated with multiple types.

#### 6.1 Annotation

In total, we produced about 58k annotations from both maps, where an annotation could have exactly one
\footnote{http://schema.org/SportsActivityLocation} type or a combination of several types, known as a multi-type entity (MTE). Table 3 shows the statistics of produced annotations.

As shown in this table, more than 93% of annotations were produced from geodata available in the map of TVB Seefeld, annotated with less than 100 types available in Schema.org. Type “Event” dominated the annotation (above 73%), followed by “LodgingBusiness” (5%), “SportsActivityLocation” (5%), “Restaurant” (2%), “Hotel” (2%), “BedAndBreakfast” (2%), and the other types were used in less than 1% of the annotations. About 22% of annotations for TVB Mayrhofen-Hippach were used MTE and only about 4% for TVB Seefeld.

We examined further the top 20 types used in the maps as shown in Figure 4. The most dominant types found are “LodgingBusiness” and “Event” in the maps of TVB Mayrhofen-Hippach and Seefeld respectively. Also, locations for doing sport activities (annotated with type “SportsActivityLocation”) dominated annotations in both maps. Mobility support (indicated with type “BusStation”) was dominated by geodata available in the map of TVB Mayrhofen-Hippach.

#### 6.2 Multi-types Entity Annotation

Other important results concern entities that were annotated with multiple types. A multi-type entity will be required whenever the entity can not be represented with a single type. This situation leads us to identify types that are not covered by Schema.org yet, as well as to find types that share similar properties.

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Maps of TVB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Annotations</td>
<td>3.986 53.997</td>
</tr>
<tr>
<td>2.</td>
<td>Used types</td>
<td>78 97</td>
</tr>
<tr>
<td>3.</td>
<td>Annotations with MTE</td>
<td>1.096 2.481</td>
</tr>
</tbody>
</table>
As shown in Figure 5, an entity could be annotated with more than one type. While most of the entities were annotated with 2 types, an entity could have up to 5 or 16 types for TVB Mayrhofen-Hippach or Seefeld respectively.

### 6.3 Discussion

We highlighted the usage of MTE in our annotations because a point of interest on a map could have a variety of organizational functions. Each function was reflected by a different type. For example, a point of interest\(^\text{18}\) has been annotated with 2 types, namely “ClothingStore” and “ShoeStore” to represent different business functions offered by the entity. These business functions were based on categories and sub-categories defined by the data provider, obtained through the mapping to the defined types from the selected vocabulary (as explained in Section 4).

Another possibility to represent multi-type entities, especially those related to e-commerce, is by defining the entity as “Shop” which offers various types of products such as “Clothing”, “Bike”, “Shoe”, and so on.

### 7 Conclusion and Future Work

In this paper, we presented a work on representing geographical data related to touristic information available on an interactive map using Schema.org. Our intention was to provide clear description to all identified points of interest such that can be consumed by intelligent application, such as search engines, personal digital assistants. Although many vocabularies for geographical data have been proposed, most of them have failed to become widely adopted. Since Schema.org vocabulary has been widely adopted, we believe that our work will gain wider acceptance from Internet users. In our work we analyzed data source and format, and identified types and categories (including sub-categories) of available objects. For every type and category, proper types from Schema.org were identified, and domain specifications were produced. The software tool was developed to consume the produced mapping, retrieve content from the API, produce relevant annotations and deploy them to a platform in a way ready to be consumed by customers. In the end, we produced about 58k annotations for data obtained from two maps, and utilized about 100 different types.

Unfortunately, schema.org doesn’t offer every possible type and property for the tourism and travel industry, so a lot of classes and features were just mapped into a new key-value pair. For example, there are no specific types for sport locations and activities, such as: hiking and climbing trails, running and cycling routes and ski slopes, which are so popular for the Tyrol region. We annotate such objects with the help of schema’s "SportsActivityLocation", "LocationFeatureSpecification", "TravelAction" and their properties (e.g. amenityFeature, potentialAction, distance) to describe the duration, length, start and end point of a trail and so on. There were also some mapping deficiencies, such as: language differences, different concepts and priorities for data types provided by Feratel, General Solution and schema.org. Also, it is worth mentioning that in some cases, a data instance might inherit properties from multiple types, known as a multi-type entity, as some objects appear in different categories or it is required to use two types to cover one object with schema.org.

Future work is to provide more specialized vocabulary for the touristic domain to build upon the core and start the process of extending the schema.org standard. For this purpose the Schema Tourism Working Group\(^\text{19}\) is running. Our ultimate goal is to have all tourism related information in the region of Tyrol, Austria semantically represented on the Web. More than just providing a high online visibility for the region on search engines, these data also contribute to opening new marketing channels such as chatbots and other intelligent personal digital assistants.

### Acknowledgements

This work was performed with cooperation and support from General Solutions, Tirol Werbung GmbH, TVB Mayrhofen and Seefeld. We would like to thank all the members of the Schema Tourism Working Group for their valuable feedback.

\(^{18}\)https://maps.seefeld.com/de#resourceDetail,1653626

\(^{19}\)https://schema-tourism.sti2.org/
REFERENCES


