Analysis of Schema.org Usage in the Tourism Domain

Boran Taylan Balcı, Umutcan Şimşek, Elias Kärle and Dieter Fensel

STI Innsbruck, Department of Computer Science, University of Innsbruck

Technikerstrasse 21a 6020 Innsbruck
{boran.balci, umutcan.simsek, elias.kaerle, dieter.fensel}@sti2.at

Abstract

Schema.org is an initiative founded in 2011 by the four-big search engine Bing, Google, Yahoo!, and Yandex. The goal of the initiative is to publish and maintain the schema.org vocabulary, in order to facilitate the publication of structured data on the web which can enable the implementation of automated agents like intelligent personal assistants and chatbots. In this paper, the usage of schema.org in tourism domain between years 2013 and 2016 is analysed. The analysis shows the adoption of schema.org, which indicates how well the tourism sector is prepared for the web that targets automated agents. The results have shown that the adoption of schema.org type and properties is grown over the years. While the US is dominating the annotation numbers, a drastic drop is observed for the proportion of the US in 2016. Poorly rated businesses are encountered more in 2016 results in comparison to previous years.

Keywords: schema.org; e-tourism; annotation.

1 Introduction

The schema.org vocabulary has been enabling publication of structured data on the web since 2011. Many businesses have included structured data publication to their online communication strategy, mainly for improving their visibility on search engines that utilize schema.org annotations for bringing better search results.

In this paper, the analysis is conducted to see the development of the schema.org usage in the tourism domain over the years. The analysis will give us a better understanding of the most demanded types and properties as well as the change of the adoption level within a given time span. The analysis is conducted on the Web Data Commons (WDC) datasets (Meusel, Petrovski, & Bizer, 2014), extracted between 2013 and 2016, on 12 different tourism related types and their properties. Additionally, a preliminary analysis has been made to see the adoption of the types and properties introduced by the Hotel Extension (Kärle, Simsek, Hepp, & Fensel, 2017). This extension was published in August 2016; therefore, the analysis is only made for the timespan between August and October 2016.

The relevant data published as N-Quads [1] by WDC is loaded to a triple store. Afterwards, a set of SPARQL[2] queries are run against the triple store for aggregation of the data. Additional post-processing techniques such as Entity Reconciliation[3] (ER) and Reverse Geocoding(RG)[4] applied for the country analysis. The remainder of this paper is structured as follows: In Section 2 literature review is explained. Section 3

[1] https://www.w3.org/TR/n-quads/
[2] https://www.w3.org/TR/rdf-sparql-query/
[3] https://en.wikipedia.org/wiki/Record_linkage
[4] https://developers.google.com/maps/documentation/geocoding/start

describes the analysis framework whereas Section 4 discusses the results. The summary and the future work are explained in Section 5.

2 Related Work

Several analyses of schema.org usage on the web exist in the literature. The study in (Toma, Stanciu, Fensel, Stavrakantonakis, & Fensel, 2014) describes an implementation of structured data as a CMS extension. The work states that 8.63% increase in visitors of DMO (Destination Management Organization) Innsbruck website is seen after schema.org implementation in a specific time interval which shows the importance of the structured data in tourism. The analysis made by (Stavrakantonakis, Toma, Fensel, & Fensel, 2013) indicates that only 5% of Austrian hotels implement annotations on their web pages whereas our work compares country annotations in a quantitative manner.

The work in (Guha, Brickley, & Macbeth, 2016) gives some statistics about the most frequently used types between 2011 and 2015. The statistics obtained from 10 billion web pages show that schema.org usage in 2015 increased from 22% to 31.3%. In our study, the analysis focuses the usage in a qualitative manner for tourism related types.

(Meusel, Bizer, & Paulheim, 2015) makes an analysis which investigates the evolution of the vocabulary alongside the adoption. They show that the evolution works two-ways, (1) users adapt to the changes in the vocabulary well (2) users use non-existent types and properties in their annotations, which influences the evolution of the vocabulary. Additionally, they point out that around half of the types in schema.org vocabulary is not used at all. The study used the statistics to show the development of schema.org vocabulary while our study provides more specific statistics based on tourism related types in both qualitative and quantitative way.

The closest related work to our work is (Kärle, Fensel, Toma, & Fensel, 2016). Our analysis is an extension to which only covered the Hotel annotations. This paper extends the previous work in the following ways: (a) our study covers several tourism related types and their properties (b) the country based analysis is more accurate with the post-process methods (c) naturally, this analysis covers a longer time span.

3 Analysis Framework

The data between 2013 and 2016 is analysed based on the classes that are relevant to the tourism domain. These classes can be listed as follows: Airport, Event, Hotel, LakeBodyOfWater, LandmarksOrHistoricalBuildings, LocalBusiness, Mountain, Museum, Park, Restaurant, RiverBodyOfWater and SkiResort. For each year, the total amount of triples vary; in 2013 1.2 billion, in 2014 622 million, in 2015 1.1 billion, in 2016 2.1 billion of triples are processed[5]. The workflow of the framework is shown in Figure 1.

[5] More than 65% of 2013 annotations come from citysearch.com. This domain was not crawled abundantly in other years which caused the drop on 2014.

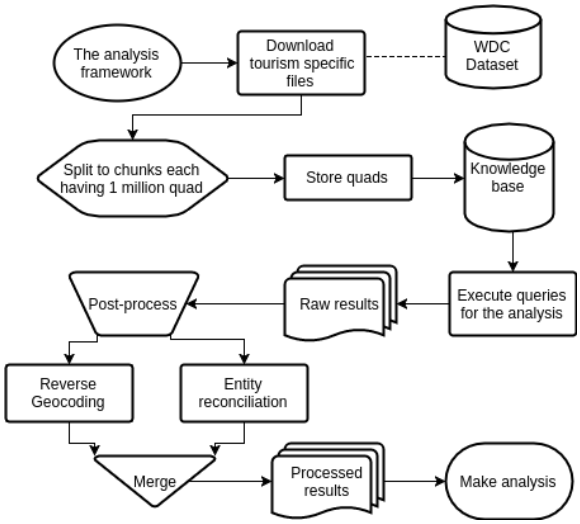

Fig. 1. The flow diagram of the framework

ER is used to reconcile country names in different languages to unified one in order to make an accurate aggregation. OpenRefine[6]'s kNN (k-Nearest Neighbour)[7] method is adopted in the framework with the similarity metric as Levenshtein distance[8]. The results are combined with the output of RG which is used to fetch country names.

The analysis is conducted regarding several aspects which are considered essential to understand the schema.org adoption from the tourism point of view. Therefore, a set of queries are prepared and executed, in order to supply results for further deduction. The queries are chosen with respect to following perspectives:

- *Count of all types:* All tourism related type instances are fetched along with other types e.g., PostalAddress, ImageObject.
- *Count of tourism related types' properties:* Tourism related type instances and their properties are fetched to observe the change in frequencies over the years. The wellness of the classes is determined by Mean Squared Error (MSE)[9].
- *Adoption of Hotel extension:* Additional instances of the classes from the Hotel Extension[10] are fetched which are introduced in August 2016 as schema.org 3.1.
- *Count of aggregate ratings:* The values of aggregateRating[11] property are fetched and normalized to 1-5 range.
- *Count of Pay-Level Domains(PLDs):* Pay-Level Domains(PLDs)[12] are obtained by querying the context of quads.

---

[6] http://openrefine.org/
[7] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[8] https://en.wikipedia.org/wiki/Edit distance
[9] https://en.wikipedia.org/wiki/Mean_squared_error
[10] http://schema.org/docs/hotels.html
[11] http://schema.org/aggregateRating

- *Count of countries:* addressCountry[13] values are processed with ER and RG.

In the following section 4, the results will be discussed.

4 Results

The observation shows that the data distribution among 12 tourism related types is disproportional. The annotations of natural beauties are found very rare in comparison to business related schema.org types such as Restaurant, Hotel, and LocalBusiness. In order to assess how well the annotation of these types are implemented, MSE values are used. The result shows that the quality of annotations generally increased over the years since MSE values decrease or remain same.

It is observed that majority of the entities are located in the US. However, over the years, the distribution of annotations is slightly shifting to other countries. For instance, in Hotel annotations, for 2013 and 2014 the ratio of the US are 79%, for 2015 72% and for 2016 the number decreases drastically to 28.4% which evidently shows that the schema.org adoption is increased by other countries over time.

One other important aspect is how well the implementation of geo and address property is made over the years. Figure 2 shows that Hotel annotations contain address or geolocation information to extract the country names whereas the success ratio of other classes could not exceed 70%. As a result, the implementation of geo and address properties remain sparse in other classes.

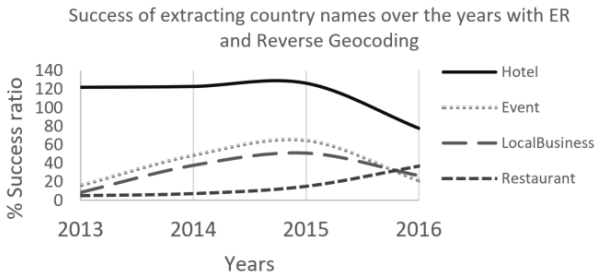

Fig. 2. Reconciliation ratio over address property frequency

Rating values showed an increase for the low values from 2015 to 2016. The study in (Park & Nicolau, 2015) analyses 5090 Restaurant reviews to find out the potential effect of ratings. The result shows that consumers appreciate extreme ratings especially bad ratings more than medium ratings. Our findings support this result.

The analysis of the Hotel Extension is made for the 3 months period between August and October 2016. The newly introduced types Campground and HotelRoom have 716 and 117 annotations respectively. Type Room has been used 3339 times. The adopters seem to use LocationFeatureSpecification for amenities almost 7000 times. Significant usage of the hasAmenity property is observed for Room type, 17000 times

[12] http://webdatacommons.org/structureddata/vocabulary-usage-analysis/
[13] http://schema.org/addressCountry

to be precise. A quick search we made has shown that hasAmenity property was proposed by an early extension attempt[14] in 2013. The analysis shows that, within the 3 months period, some classes introduced by the extension started to be used. However, most of the classes have no occurrence in the WDC datasets.

5 Conclusion

In this paper, an analysis conducted with a tourism related subset of schema.org from different perspectives (i.e., types, properties, geo data, PLD, and aggregated rating) has been presented. It is observed that the adoption has become better over the years. However, the values of geo and addressCountry properties were insufficient to determine country names for the instances, except for the Hotel type. The newly introduced classes in the hotel extension are used in a small amount of annotations. A significant increase of low ratings is seen from 2015 to 2016. Even though, the quality of schema.org annotations is improved, there are still issues with important properties such as address. Nevertheless, touristic service providers not only in the US, but also in other countries is increasingly getting ready for the automated agents on the web.

As a future work, PLDs can be analysed in more detail to observe the improvement in annotation pattern. The analysis of 2017 WDC data dump would show better results for the hotel extension since 2016 data dump is extracted in an early stage of the extension. The extended results are under the link http://btbalci.sti2.at/enter2018/.

References

Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema.org: Evolution of Structured Data on the Web. *Communications of the ACM*, 59(2):44–51.

Kärle, E., Fensel, A., Toma, I., and Fensel, D. (2016). Why Are There More Hotels in Tyrol than in Austria? Analyzing Schema.org Usage in the Hotel Domain. In Inversini, A. and Schegg, R., editors, *Information and Communication Technologies in Tourism 2016: Proceedings of the International Conference in Bilbao, Spain, February 2-5, 2016*, pages 99–112. Springer International Publishing, Cham.

Kärle, E., Simsek, U., Hepp, M., and Fensel, D. (2017). Extending the schema.org vocabulary for richer hotel annotations. In *Information and Communication Technologies in Tourism 2017*, volume 26, pages 31–41. Springer.

Meusel, R., Bizer, C., and Paulheim, H. (2015). A Web-scale Study of the Adoption and Evolution of the Schema.Org Vocabulary over Time. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, WIMS '15, pages 15:1—-15:11, New York, NY, USA. ACM.

Meusel, R., Petrovski, P., and Bizer, C. (2014). The WebDataCommons Microdata, RDFa and Microformat Dataset Series. pages 277–292. Springer, Cham.

Park, S. and Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. Annals of Tourism Research, 50:67–83.

Stavrakantonakis, I., Toma, I., Fensel, A., & Fensel, D. (2013). Hotel websites, Web 2.0, Web 3.0 and online direct marketing: The case of Austria. In *Information and communication technologies in tourism 2014* (pp. 665-677). Springer, Cham.

Toma, I., Stanciu, C., Fensel, A., Stavrakantonakis, I., & Fensel, D. (2014). Improving the online visibility of touristic service providers by using semantic annotations. In *European Semantic Web Conference* (pp. 259-262). Springer, Cham.

[14] https://www.w3.org/wiki/WebSchemas/LodgingExtensions