Ontological Engineering Tourism Intelligence

Building an ontology from occupancy surveys

Andreas Koch

Asunción Gómez Pérez

5. November 2021

Introduction

As data and its representation become more important, knowledge bases and more specifically ontologies provide many advantages over traditional representations and storing data inside Excel spreadsheets. The goal of this work is to exemplify this by building an ontology from Spain's occupation statistics of tourism accommodations. First, the ontology and data will be specified, before giving an overview of the timeline and organisation of this project. Afterward, a conceptual model of the ontology and its implementation will be presented, which will be evaluated in the last section.

As far as the development of the ontology is concerned, the NeOn methodology[1] will be used as a guideline. It was introduced in 2009 to provide directions on how to develop an ontology or an ontology network. For this purpose, it presented 9 scenarios arising in the development of ontologies and the steps required to accomplish the respective goal. In this work, a new ontology is being developed and for this reason the first scenario is followed. In order to develop an ontology network from scratch, the ontology requirements need to be specified. They include a set of competency questions that the ontology should be capable of answering, as well as the purpose, scope, target group and implementation language[2]. This task is presented in the following section in the form of an ontology requirements specification document (ORSD). The most frequent terms in it are used to search for further resources.

[1] http://neon-project.org/nw/NeOn_Book.html

[2] Gómez-Pérez, Asunción, and Mari Carmen Suárez-Figueroa. "NeOn methodology for building ontology networks: a scenario-based methodology." (2009).

Ontology Specification

The first step in the development of an ontology is to specify all requirements. As the data for this work comprises tourism statistics from occupancy surveys in Spain only, the goal of this work is to test an ontology as a tool for exploring these statistics. Since tourists are not modelled individually but measured as a statistic in the occupancy surveys, any attempt to model tourist profiles requires more information. Consequently, this approach is left for future work. Apart from monthly tourist numbers, other variables from the surveys include monthly overnight stays, accommodations, capacity, rooms, occupied rooms, personal employed and the degree of occupied rooms on weekdays and on weekends. Statistics are recorded for both domestic and international tourists on a monthly basis from 1999 to 2021. Besides mathematical formulas, there is no connection between these data entries. Therefore, no meaningful connection can be made in the ontology.

The State Secretariat for Tourism does provide the results for several different subsurveys as in Figure 1: category of accommodation, country of origin of the tourists, province, region, touristic zone and touristic point. Although the API service only allows downloads of files of one of these classes with statistics of all mentioned variables, known relations between subsurvey classes can still be exploited for an ontology. To be more precise, the values of classes region, province, touristic zone and touristic point all entail a location in Spain. Hence, an ontology could incorporate “part-of” relations between these classes and all respective instances. Other than that, no relation between classes can be identified.

Lastly, there are 5 different surveys spanning statistics for camping sites, hostels, touristic apartments, hotels and rural accommodations. These are denoted as the type of accommodation in this work. Every survey includes

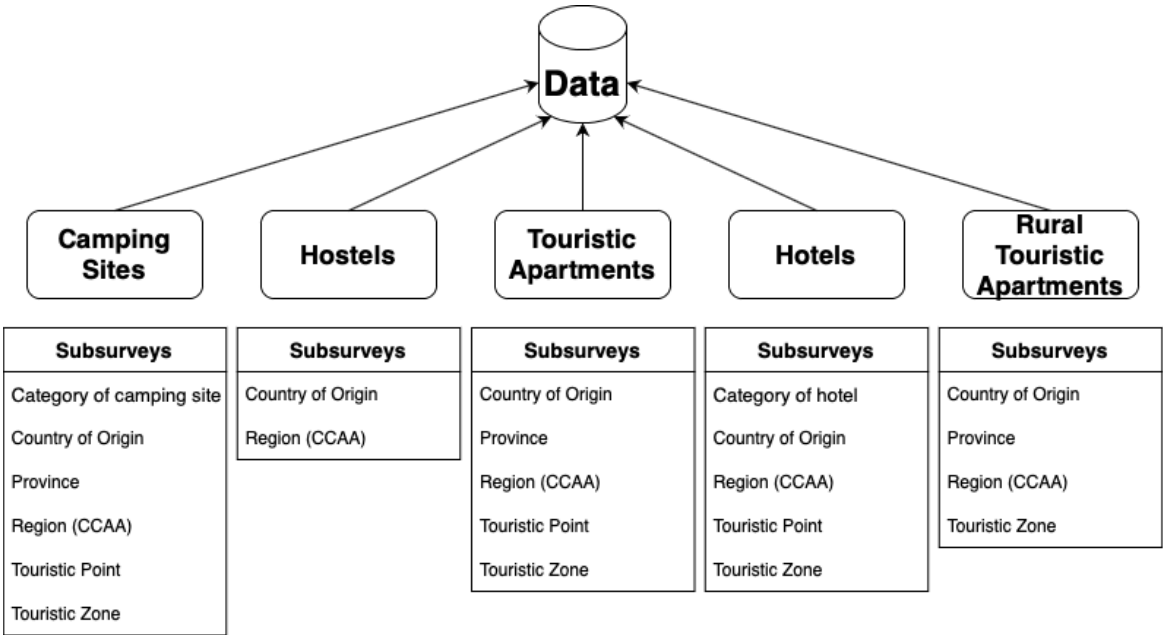


Figure 1: Subsurvey structure for the different surveys

statistics for several subsurveys, each distributed over multiple files. The first challenge is to download all files and organise them. Further, they will be used to create an ontology that should be able to answer all competency questions as outlined by the ORSD.

Competency questions have been classified by the questions' topics into 5 different groups. Since there are no agents or other entities besides the statistic of a specific location or region, most questions are formulated with the goal of reading numbers and processing them. Based on these findings, questions can also expect a time, location, the value of any corresponding variable or a list of values as an answer. Since it is impossible to provide answers to all of them by hand, the ORSD does not list any answers. Each competency question can be implemented via a SPARQL query. In order to validate the respective result, it can be computed via a different analytics tool such as custom python scripts. For this work, several competency questions are implemented and will be presented when evaluating the ontology.

Ontology Schedule

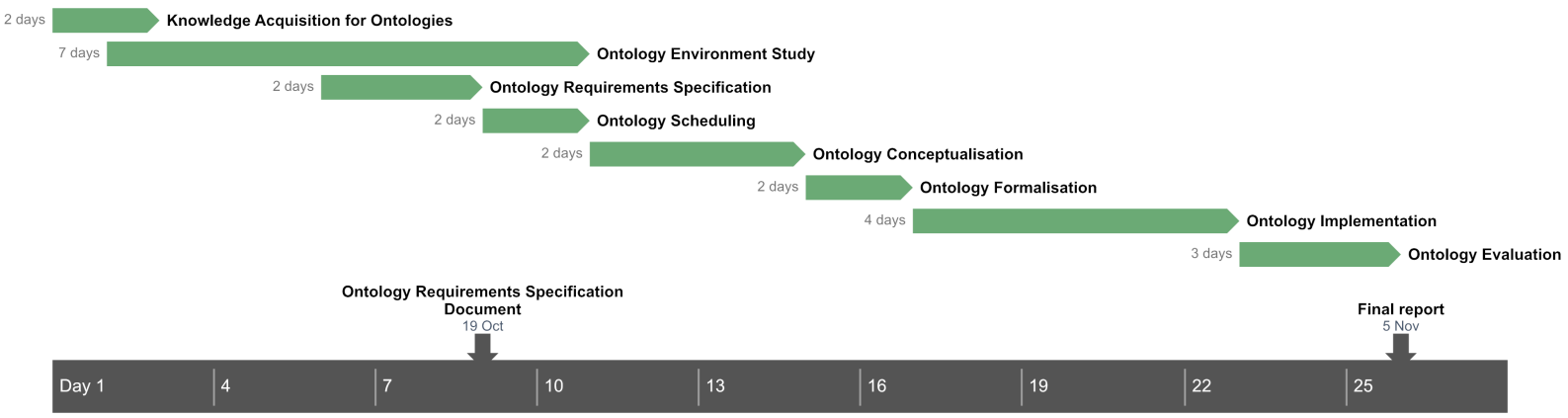
The most frequent terms appearing in the competency questions should be used to guide further research in the following section. Additionally, it might be useful to categorise in which way the competency questions inquire about its frequent terms. For this reason, frequent terms include not only topical keywords but also interrogative pronouns and modifiers. They are depicted in part 7 of the ORSD. From the results, it is apparent that questions concerning accommodations are formulated towards the number of accommodations rather than their type. Moreover, questions regarding the type of accommodations use camping sites, hostels and hotels as examples for types and really expect a list with results spanning all types of accommodations. Apart from accommodations, it is noteworthy that questions split roughly even among asking for an amount, for a specific data entry or for the result of a computation.

The goal of this section is to produce a Gantt chart with all activities to be carried out during the development of the tourism intelligence ontology. Up to this point, several activities have been completed. The knowledge acquisition phase was completed by getting familiar with the task and data at hand. Furthermore, the ontology environment study refers to the process of getting accustomed to working with the ontology development tool Protégé. With the ontology requirements specification and ontology scheduling phases now completed, further resources relevant to this written work will be presented in the following section.

Ontology Requirements Specification Document - Part 1

1	Purpose
	The purpose of building an ontology from occupancy surveys in Spain is to provide a better tool for exploring the statistics and making it possible to discover patterns in tourism behavior.
2	Scope
	The ontology focuses on official occupancy surveys of camping sites, hostels, touristic apartments, hotels and rural tourism lodges.
3	Implementation Language
	The ontology is implemented in Protégé with the OWL syntax.
4	Intended End-Users
	User 1. Managers in the tourism sector
5	Intended Users
	User 2. Employees of the ministry of tourism

Gantt Chart



Ontology Requirements Specification Document - Part 2

6	Ontology Requirements
	a. Non-Functional Requirements
	NFR 1. The ontology must have only English objects and properties while the underlying instances should keep their original names in Spanish.
	NFR 2. The ontology should be as simple as possible. (Is that a functional req?)
	b. Functional Requirements: Groups of Competency Questions
	CQG1. Accommodation
	CQ1. How many accommodations were available in any given month?
	CQ2. How many accommodations are available at a specific touristic point?
	CQ3. How many accommodations are available in a specific touristic zone?
	CQ4. How do accommodations divide into camping sites, hostels, etc.?
	CQ5. How does the accommodation distribution change over time?
	CQ6. When did the number of accommodations change drastically from one month to another?
	CQ7. Where did the number of accommodations change drastically from one month to another?
	CQ8. What is the most underserved touristic zone based on accommodations per tourist?
	CQG2. Tourists
	CQ9. What is the most popular touristic point by number of tourists?
	CQ10. What is the most popular touristic zone by number of tourists?
	CQ11. What is the most popular region by number of tourists?
	CQ12. How many tourists visit each touristic zone each month of summer on average?
	CQ13. What is the change of tourists in each touristic zone each month of summer on average?
	CQ14. How many tourists come from Germany in any given month?
	CQ15. What is the distribution of tourists over their countries of origins in any given month?
	CQ16. How many tourists visit each category of accommodation and how did this change over time?

	CQG3. Labor
	CQ17. How many people work in the tourism sector in any given month?
	CQ18. Which region sees the most growth in personal employed on average from one month to another?
	CQ19. What is the change of personal employed in each touristic zone each month of summer on average?
	CQ20. What is the rate of personal employed per tourist for each category of accommodation?
	CQ21. In which months do employment numbers decline significantly on average and which regions are hit the most? (Excluding covid)
	CQ22. How many jobs are provided by accommodations in total?
	CQ23. How did the number of jobs at accommodations change over time?
	CQ24. How does the rate of personal employed per tourist change over time?
	CQG4. Domestic / International Tourism
	CQ25. Which regions have the most domestic tourists?
	CQ26. Which touristic points have the highest international to domestic tourist rates?
	CQ27. What is the rate of international to domestic tourists for the different categories of accommodation?
	CQ28. How many international tourists come to Spain each year and how did that number change over time?
	CQ29. Which are the most popular touristic zones for domestic / international tourists?
	CQ30. For how many nights do domestic / international tourists stay on average?
	CQ31. What is the highest number of domestic / international tourists in Spain in a month and when did it occur?
	CQ32. Which touristic points had the most growth of international / domestic tourists in the last 5 years?
	CQG5. Degree of Occupancy
	CQ33. How does the average degree of occupancy on camping sites compare to hotels, etc.?
	CQ34. How much higher is the degree of occupancy for weekends compared to weekdays on average?
	CQ35. What is the average degree of occupancy each month for the different accommodations?
	CQ36. How did the degree of occupancy change over time for the different types of accommodations?
	CQ37. Which category of accommodation has the highest degree of occupancy?
	CQ38. Which touristic points have the highest degrees of occupancy?
	CQ39. In which touristic zones are the highest rates of degree of occupancy to number of establishments?
	CQ40. Which touristic zones have the lowest degrees of occupancy?

Ontology Requirements Specification Document - Part 4

7	Pre-Glossary of Terms	
	a. Terms from Competency Questions	
	Keywords	**88**
	Tourism	5
	Touristic Point	5
	Touristic Zone	9
	Number of Tourists	3
	Accommodation	16
	Types of Accommodation	1
	Category of Accommodation	3
	Camping	3
	Hostel	2
	Hotel	2
	Countries of origin	1
	Personal Employed	4
	Jobs	2
	Domestic	7
	International	7
	Occupancy	10
	Degree of Occupancy	6
	Degrees of Occupancy	2
	Interrogative Pronouns / Modifiers	**72**
	Month	13
	Year	2
	How many	10
	Number	10
	Which	11
	What	11
	Highest	5
	Lowest	1
	Rate	5
	Per	8
	Change over time	6

Ontology Requirements Specification Document - Part 5

	b. Objects
	Types: camping sites, hostels, touristic apartments, hotels and rural accommodations
	Classes: category of accommodation, country of origin of the tourists, region, province, touristic zone and touristic point

Resources

The purpose of this work is to develop an ontology for statistics on tourism in Spain provided by the State Secretariat of Tourism. The relevant files are obtained via their API service[3]. If differentiation of the statistics for the values of each class is desired, a GET method with a specific value needs to be evoked. In order to gain as much information as possible from the surveys, a call for every option is required. Since there exist many possible values, the process of downloading and organising files is automated[4]. This work spans surveys of occupation statistics in camping sites, hostels, touristic apartments, hotels and rural accommodations. Consequently, all files associated with these surveys are acquired.

To fulfil the requirements, the ontology for tourism data might need to connect and relate all information regarding regions and locations in Spain. Therefore, an ontology already mapping locations to regions would be useful. To this end, de León et al, 2010[5] published an RDF about Spanish geospatial data. It contains the latitude and longitude information of many individual named points, as well as lines and named areas. As the names used in the RDF do not necessarily overlap with the names in the occupation surveys, it could only partially solve the problem of matching regions and

[3] https://www.dataestur.es/en/apidata/

[4] https://github.com/theonlyandreas/tourism_intelligence_ontology

[5] de León, A., Saquicela, V., Vilches, L.M., Villazón-Terrazas, B., Priyatna, F. and Corcho, O., 2010, September. Geographical linked data: a Spanish use case. In *Proceedings of the 6th International Conference on Semantic Systems* (pp. 1-3).

locations. As this would still need a lot of work, it was chosen to be omitted altogether.

The Aragopedia Ontology[6] project is a vocabulary that extends DBpedia with concepts and properties about tourism data in Aragon, Spain among other things. It serves as an example of how to represent regions, municipalities and provinces in an ontology. Nevertheless, the Aragopedia ontology only includes classes and no instances. For this reason, it does not solve the problem of relating specific instances of regions and other locations.

There are several design patterns that could be applied for the development of the occupation ontology. Out of all submitted design patterns[7], the patterns for places, regions and time intervals are partly adapted within the ontological models presented in the following section.

[6] https://opendata.aragon.es

[7] http://ontologydesignpatterns.org/wiki/Submissions:ContentOPs

Conceptual Model

Since the data consists of statistics about tourism in Spain without any information about individual tourists or agents that perform actions, the simplest ontology possible will represent rows as individuals and columns as data properties. In order to provide better navigation, subsurveys were chosen to be represented by classes. Accordingly, they act as proxy classes and have all individuals be rows. This simple ontology will be referred to as the "minimal ontology" and is depicted in Figure 2 as well as in Figure 4.

As was mentioned, more complexity can be introduced by adding geographic and temporal information. The "complex ontology" in Figure 3 and Figure 5 has 5 proxy classes to represent the rows of each survey. Apart from these, it has a class to represent a region, touristic zone, touristic point, year and month respectively. Hereby, an individual of one of these classes will actually be a specific region, touristic point etc. All their relations can be modelled via object properties instead of data properties, including the relation to proxy classes. Furthermore, all relations between individuals can also be captured via object properties. Geographical and temporal relations are modelled by a "part-of" and a "contains" relation, e.g. a region contains a touristic zone and a touristic zone contains a touristic point. Similarly, a month is part of a year and a year contains a month.

To implement both ontologies, the RDFLib[8] python library was used, which provides a graph object to interact with an ontology. In this way, any subject-predicate-object triple can be added. Consequently, creating an ontology from the tourism statistics data involves iterating over all surveys, subsurveys, rows and columns[9]. Thereby, a row is the subject, a column the object and the data entry the value for the respective predicate.

---

[8] https://github.com/RDFLib/rdflib

[9] https://github.com/theonlyandreas/tourism_intelligence_ontology

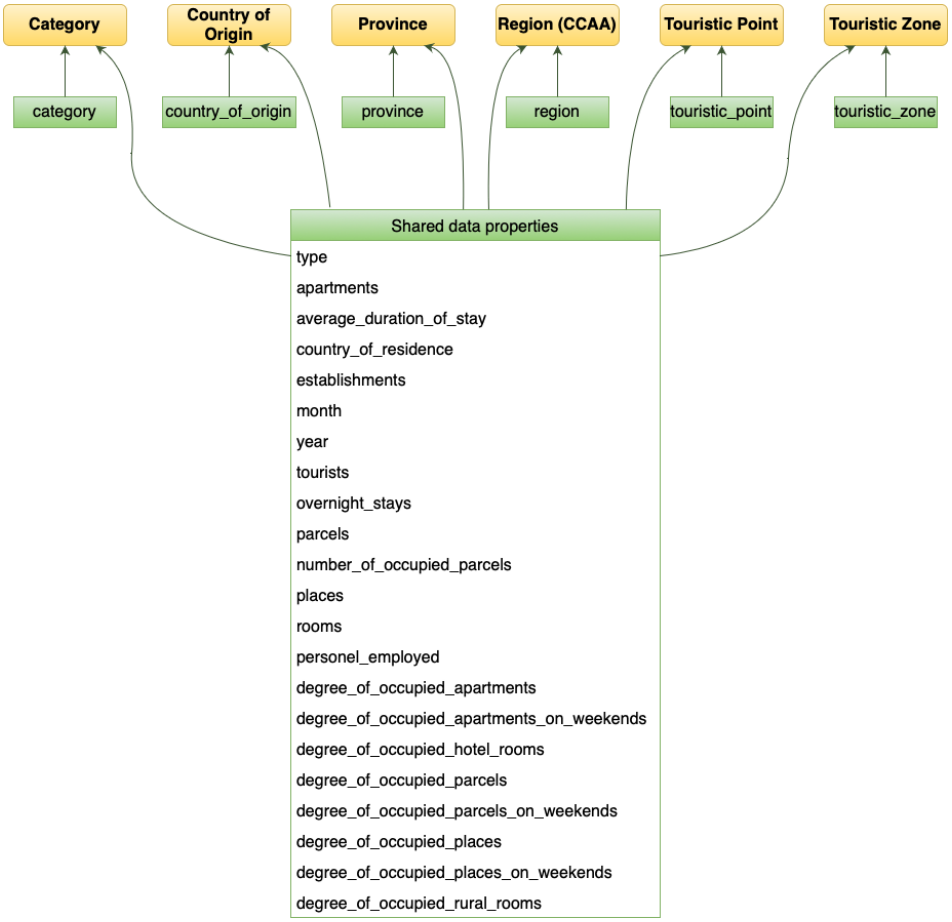


Figure 2: Minimal ontology with subsurveys as proxy classes (yellow) for rows. All columns are represented by data properties (green).

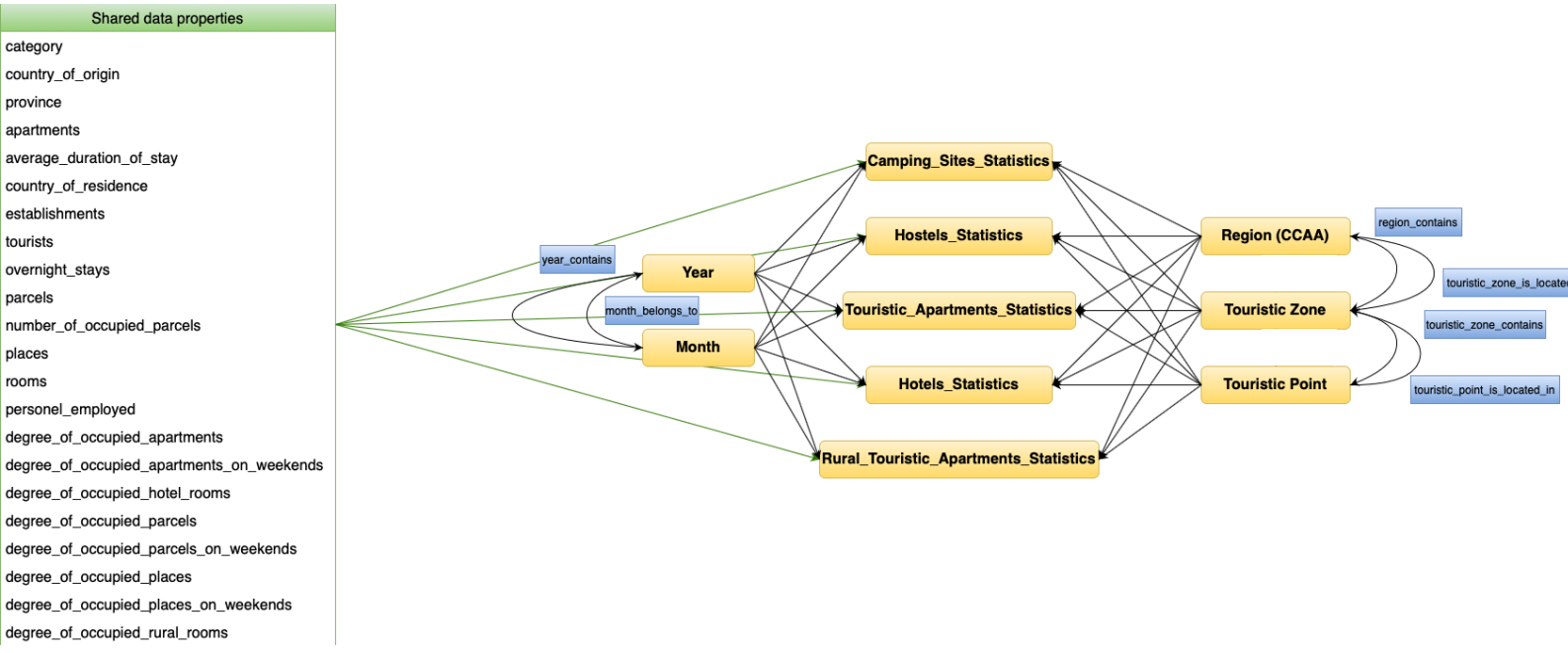


Figure 3: Complex Ontology with survey proxy classes (yellow) for rows. The region, touristic point, touristic zone, year and month are represented by their own classes (yellow) with object properties (blue) connecting them. Data properties (green) are shared for all proxy classes.

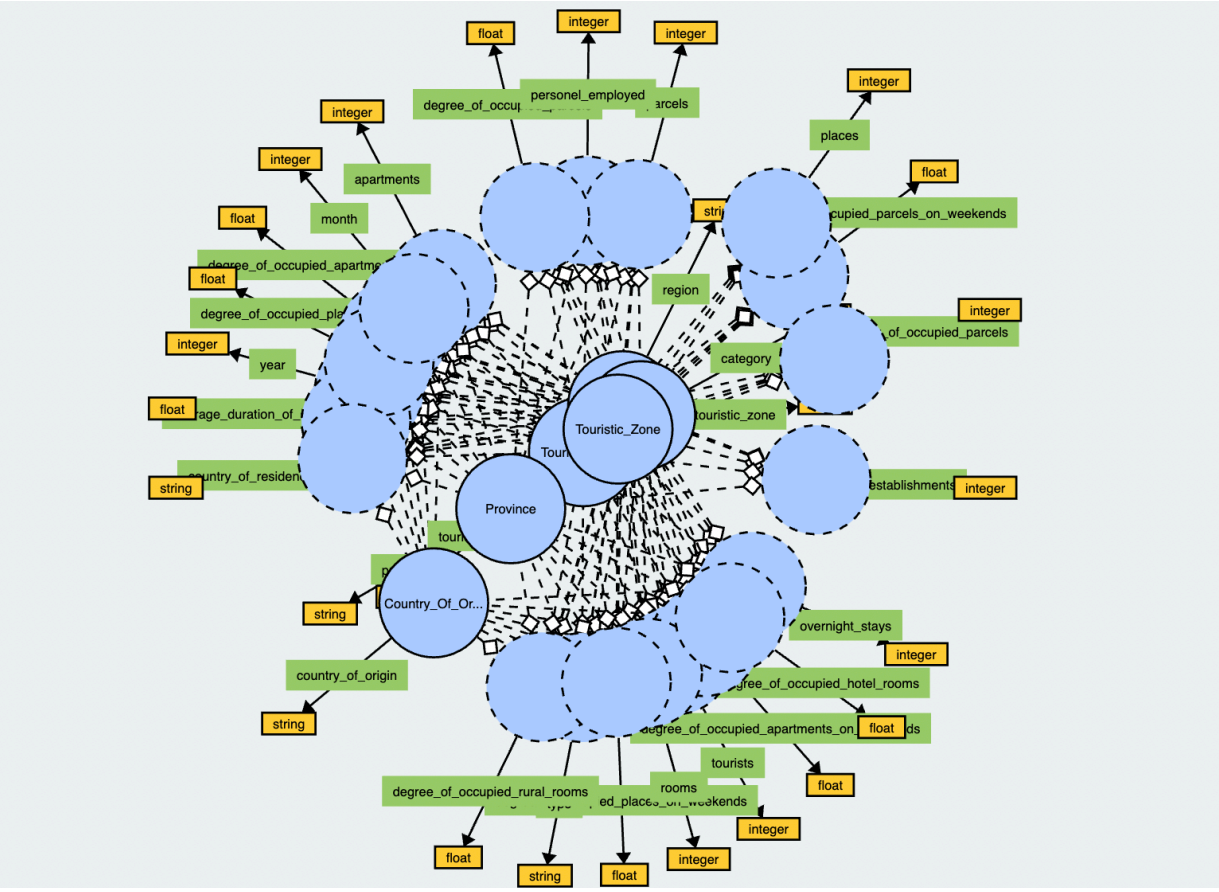


Figure 4: Minimal ontology visualised via WebVOWL. Subsurveys are used as proxy classes and thus have rows as individuals. Each has their own data property to hold the actual touristic zone etc. All other data properties are shared among all classes.

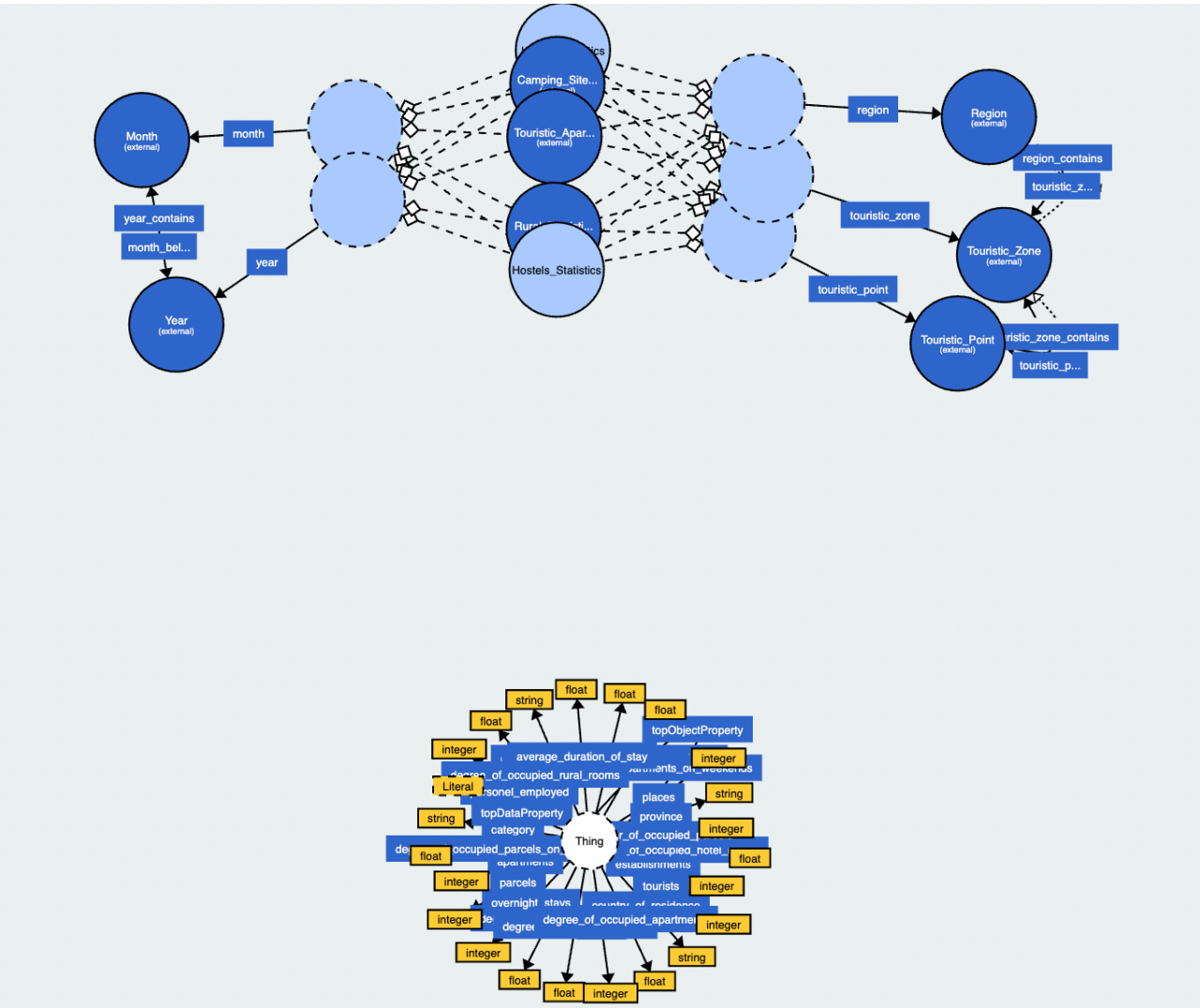


Figure 5: Complex ontology visualised via WebVOWL. Surveys are used as proxy classes for rows. Region, touristic zone, touristic point, year and month are modelled separately and incorporate geographical and temporal information with their object properties.

Results

Ontologies are best evaluated by validating their competency questions via SPARQL queries. For this purpose, only the complex ontology will be evaluated since it is more extensive. It is important to note that the validated competency questions presented here are merely examples and that all questions from the ORSD can be validated in a similar fashion. The full script for the evaluation is available at [10].

The first competency question is CQ2: How many accommodations are available at a specific touristic point? More specifically, the question is interested in the touristic point Benidorm and the number of accommodations available in hotels in August 2021.

```
In [6]: %run -i sparql_queries.py
# CQ2: How many accommodations are available at a specific touristic point? 242
SELECT DISTINCT ?acc_num
WHERE {
    ?s rdf:type tio:Hotels_Statistics .
    ?s tio:touristic_point tio:Benidorm .
    ?s tio:month tio:8 .
    ?s tio:year tio:2021 .
    ?s tio:establishments ?acc_num .
}
Answer: (rdflib.term.Literal('242', datatype=rdflib.term.URIRef('http://www.w3.org/2001/XMLSchema#integer')),)
```

Code example for CQ2: How many accommodations are available at a specific touristic point? The answer is 242 and is correctly returned by the ontology. Thereby, the namespace “tio” is used as an abbreviation for the tourism intelligence ontology.

The second competency question being validated is CQ14: How many tourists come from Germany in any given month? The validation is again performed for the most recent month, which is August 2021. Now the total number of tourists is needed and for this reason, the number of tourists is summed over all different surveys. Additionally, the country of origin is not modelled by its own class. Instead, it is represented by a data property.

[10] https://github.com/theonlyandreas/tourism_intelligence_ontology

Therefore, a filter needs to be applied to all possible countries to check for the correct string value.

```
In [8]: %run -i sparql_queries.py
# CQ14: How many tourists come from Germany in any given month? 547054 (for August,2021)
SELECT DISTINCT ?total_tourists (SUM(?t) AS ?total_tourists)
WHERE {
        ?s tio:country_of_origin ?country FILTER ( str(?country) = "Alemania" ) .
        ?s tio:month tio:8 .
        ?s tio:year tio:2021 .
        ?s tio:tourists ?t .
}
Answer: (rdflib.term.Literal('547054', datatype=rdflib.term.URIRef('http://www.w3.org/2001/XMLSchema#integer')), rdf
lib.term.Literal('547054', datatype=rdflib.term.URIRef('http://www.w3.org/2001/XMLSchema#integer')))
```

Code example for CQ14: How many tourists come from Germany in any given month? The answer is 547054 and it is correctly returned by the ontology. Again, the namespace “tio” is used as an abbreviation for the tourism intelligence ontology. As the country of origin is not being represented by its own class, but rather a data property, all possible values need to be compared to the correct string value.

As is apparent from these queries, modelling columns as classes instead of data properties has its advantages. To improve the ontology even further, all object property relations between individuals could be added as well. This would require matching regions with touristic zones and touristic zones with touristic points, but is not impossible. For the temporal information, 12 months would be needed for every year and they would have to be different individuals for each year. This would have the benefit of making it possible to use these object properties when relating two specific individuals of these classes. Consequently, forming sums over all months of a specific year or over all touristic points within a zone would become feasible.

Conclusion

In this work, it was shown that ontologies can be built for statistics data such as Spain's occupation statistics surveys. Thereby, the development process laid out in the NeOn methodology was followed, which included specifying all requirements in the ORSD as well as presenting further resources and a conceptual model for the ontology. In the last section, two validated competency questions were presented, being examples of how such an ontology might be used. Further improvements can still be made, but the ontology provides to be useful already. If queried correctly via SPARQL queries, it could provide better insight into all occupation statistics. Naturally, this would mean implementing a tool with many previously written queries, such that the user only needs to pick the right one and provide their values as input. Ideally, these queries would be abstracted and the user is only presented with a user interface. This would mean that the ontology would be used in the background without the user even noticing, which is the way they were intended to be used anyway. To conclude, ontologies are a good tool for representing data and should be considered as an option in the development of programs that involve a knowledge representation problem.