A data-driven methodology to generate living ontologies

Miguel Á. Garrido and Paloma Cáceres
Higher Technical School of Computer Engineering, Rey Juan Carlos University, Madrid, Spain
E-mails: {miguel.garrido, paloma.caceres}@urjc.es

Abstract. At present, there are large volumes of data that, through appropriate processing, can be transformed into information that is valuable to users. In this sense, the Semantic Web is a promising initiative to enrich data in the digital world, through the creation of ontologies that provide data of higher quality and with an associated meaning. This paper describes MAOn, a data-driven methodology for creating living ontologies. The term ‘data-driven’ refers to obtaining the data through various multi-modal sources that provide knowledge about which data to annotate semantically. The term ‘living ontology’ is characterized by allowing the creation of domain ontologies that can be expanded through an iterative and incremental process. Additionally, MAOn is characterized by an exhaustive validation phase that in a complete way validates and verifies the ontology obtained. Finally, as a way of validating the methodology presented, a complete case study about public transport and accessibility issues is provided.

Keywords: methodology, linked data, ontology, public transport, accessibility

1. Introduction

The term Smart City is taking on great relevance in today’s world. This concept refers to a city that makes use of Information and Communication Technologies in order to create better infrastructures for citizens as extracted from [12]. In addition, it includes improvements that can be offered in aspects of urban life such as mobility, the economy, the environment, the quality of life for its inhabitants, and administration.

The research line that directs this work seeks to obtain datasets that include meaning in order to provide benefits to consumers in the field of mobility. The motivation is provided by the reality that we live in the information age. This promotes the generation of large amounts of data, which are created every minute that passes, and which have given rise to the increasingly popular term ‘Big Data.’ In [15] this excess of data is presented as a problem that concerns human beings because of the difficulties generated by capturing, storing, analysing and viewing such an enormous volume of data. On the other hand, this magnitude of data offers many opportunities to make great progress in many fields and for many different uses.

In addition to what is referred to about large volumes of data, the types of information obtained can be of a very diverse nature; therefore, its analysis must discriminate between what is important and what is not. Following this line, the nature of the information can come from diverse sources that provide different information of a similar quality or scope. For this reason, it is advisable to carry out a suitable combination process in order to obtain unified data that offer completeness to the information on the given domain. Several problems may arise from this unification due to the different formats in which the data are presented. In this sense, it will be necessary to make use of mechanisms that gather the data in a suitable format to be offered in an open way to data consumers.

As an example of large volumes of data in the field of mobility, observations can be made about information related to public transport and its accessibility. There are many sources of data relating to public transport in any industrialised city in the world today. This information is very diverse and is presented in very different forms through brochures, plans, guides, web portals, applications, etc. Each entity or company establishes which data is relevant to show its travelers
and with what degree of detail. The aim of this information is to make it easier for passengers to travel through the public transport network. In this sense, several questions arise: Is the information useful and comprehensible to all passengers? In principle, the answer is simple: yes, that is its purpose. However, is it really so? Well, it will depend on the circumstances faced by the travellers in question. That is why it is important to take into account the needs of people who have some kind of disability or limitation. In addition, there may be special circumstances for passengers, such as carrying a child’s trolley, carrying bulky and heavy luggage, having some kind of phobia, and similar situations that may arise in the daily lives of users of public transport. These circumstances, which can affect everyone, can prevent the completion of the trip and make it necessary to take other alternatives that are surely more expensive.

As is well known, today’s world is increasingly inclusive, in which everyone has a place without exceptions and where, through new technologies, it is feasible to offer improved facilities to citizens. Improving accessibility in public transport is a great challenge for this society, which is in the process of continuous improvement, all of this through infrastructures which incorporate more and more accessibility elements but which do not cover 100% of the needs. Therefore, it is very convenient to offer as much information as possible to public transport users in order to facilitate their trips and help them avoid encountering barriers that make travel difficult or impossible. In this sense, the case study proposed in this work as an evaluation method will be focused on.

In this context of research on the collection and presentation of information, this work aims to provide a mechanism for providing information that is not directly available or that one wishes to include in order facilitate its use. To this end, a method is presented based on a series of well-defined steps in a structured manner, which defines sets of semantic data that add value for consumers. All this is presented in the form of a data-driven methodology for the creation of living ontologies. Using the methodology, a simple, directed and direct method is promoted to create an ontology that provides enhancements to the data published on the Web in the form of enriched data. In conclusion, this methodology aims to provide a data-driven, multimodal information unification mechanism for the creation of living ontologies that promote open and linked datasets.

The paper is organised as follows. Section 2 presents the related investigations. Section 3 describes the steps that make up the methodology. Section 4 evaluates the proposed methodology through a case study and through different approaches aimed at validating and verifying the method followed. Finally, conclusions and future work are presented in section 5.

2. Related work

This work arises as a result of different investigations into the importance of information and the way in which it is presented on the Web. The Semantic Web is a discipline that aims to generate useful information derived from data by transforming it into information that can be processed by machines through a semantic theory [20]. The semantic aspect provides meaning to the actionable terms, thus facilitating interoperability between systems. Therefore, the semantic component facilitates the existence of computational data, understandable not only by humans but also by machines, in order to offer information enrichment. Through the inclusion of ontologies [10], a basis is provided for these pieces of information in the form of conceptual structures. The ontology will therefore provide the keys to result-processable data by machines in the Semantic Web.

The line of research that directs this paper is motivated by the key term public transport accessibility. It describes the background that guides this work and that has motivated the development of a domain-based methodology that is useful for its application in many different scenarios. In line with the key terms presented, different investigations have been carried out on the subject of accessibility in public transport.

The G3ict1 initiative promotes accessibility through e-inclusion using ICTs through policies, standards and technical resources developed with United Nations agencies. The Cities For All initiative promoted by World ENABLED2, a non-profit educational organization that promotes the rights and dignity of people with disabilities, promotes the inclusion of people with disabilities in all urban planning. Also promoted by the latter two entities is the initiative known as A Global Strategy for Digital Inclusion [19], which includes, as a method for increasing accessibility, action aimed at the use of open and accessible datasets that include information for and about people with disabilities. All

---

1 https://g3ict.org/

2 http://worldenabled.org/
these initiatives are focused on the field of accessibility with the aim of promoting the inclusion of all people through policies and regulations that defend and regulate it. In addition, they encourage the use of open data that is available, but all this through guidelines that are not reflected in practical cases or developments that provide them.

In the field of transport, [11] describes a system for automated generation of data linked from the transport domain. Another contribution presents a workflow for publishing and linking transport data on the web. Finally, [9] proposes the development of an ontology based on data models and standards linking public transport information for the creation of new services and applications for the benefit of passengers and suppliers. All these approaches promote the creation of linked data in the public transport domain, in order to provide information for systems that consume it. These proposals are useful, but do not take into account accessibility features.

Other approaches such as NaPTAN³, which provide a vocabulary with which to uniquely identify national public transport access nodes in the UK, do not incorporate any aspect of accessibility. For its part, the Ontology⁴ proposal makes it possible to represent whether a place is accessible to people with mobility problems through classes such as AccessFacilities and properties such as is_wheelchair_accessible, but it does not take into account other elements necessary to provide information to blind or deaf people. Similarly, DBPedia⁵ has a property called isHandicappedAccessible to indicate whether an accessible transit station is also mobility oriented. The Tube Facility⁶ ontology incorporates only free passage and elevation facilities as accessibility elements, that is, it addresses disabilities associated with mobility. Accessibility Ontology⁷ also focuses on concepts related to supporting the mobility problems of people with disabilities. TRANSIT⁸ is a specific ontology for transit, but does not incorporate any aspect of accessibility. Finally, the European Project OASIS⁹ (Open architecture for Accessible Services Integration and Standardisation) does not incorporate concepts relevant to the accessibility needs for public transport.

With regard to the mode of operation and the infrastructures incorporated in public transport, there are several standards that include an in-depth study of how public transport works and how it is structured, in specific terms. Through the study of the TRANSMODEL standard [21], in-depth knowledge is obtained of how public transport can be modelled in its entirety and through the different modes in which it can be given or provided to passengers. On the other hand, through the use of another standard — IFOPT [22], used by TRANSMODEL, which TRANSMODEL uses — data on accessibility in public transport is obtained. The information provided by these standards is remarkable because it provides great coverage when it allows description of public transport in all its extension.

In previous works, based on the expert knowledge obtained from the study of the TRANSMODEL and IFOPT standards, an ontology was elaborated that received the name of MAnTo: Mobility and Accessibility Ontology. This ontology was obtained as a method to provide open knowledge on accessibility in public transport to help passengers in their journeys through cities. MAnTo provides a comprehensive information mechanism for describing accessibility elements in the different modes in which public transport can be characterised. This ontology describes not only physical elements of transport, such as a station, a platform, an access point, etc., but also includes the logic that follows regarding its distribution in lines, the possible connections, the geographical location that represents a certain point on a map, etc. In addition, it should be noted that it includes, as a new contribution, information associated with accessibility that offers transport information for people with special motor and cognitive needs, as well as needs derived from health or physical conditions. Through this ontology, a mechanism was established to be able to provide linked and open information on public transport for any user who wishes to have it.

As stated above, MAnTo has great advantages that provide valuable information for public transport operations based on a standard, with all that this implies, which provides a fundamental feature such as information accessibility. In addition, it provides to any information system other desirable qualities, such as availability, completeness and openness, to anyone who wants to have it, or systems that want to consume

³ http://naptan.app.dft.gov.uk/datarequest/help
⁴ http://ip-kom.net/data/html_documentation_ontologien/ofi-ontology.html
⁵ http://dbpedia.org/ontology/isHandicappedAccessible
⁷ https://sourceforge.net/projects/accessibilitydomainowl
⁸ http://vocab.org/transit/
⁹ http://www.oasis-project.eu/
such information, with the aim of promoting open and linked data. In contrast, MAnto is a very extensive ontology that needs experts in the domain to understand its full scope. This ontology incorporates a multitude of terms that provide completeness, but make it unmanageable as it results in less agility in its use. In addition, this ontology is closed because it is based on what is specified by the IFOPT standard. For all these reasons, although it may in some aspects be superior, it can be impractical to use due to the great complexity introduced by its completeness. This assertion is based on quantitative terms through a study carried out in which, after noting transport data from a set of data obtained from the Empresa Municipal de Transportes de Madrid (EMT) [3]. As a result, the MAnto terms used to record the data obtained from EMT are around 5%. This data reinforces the hypothesis that MAnto, although it is very complete, because it is based on a standard, is gigantic relative to the data that are actually used; therefore, it is underused for a specific scenario, making it unmanageable due to its large dimensions.

The methodological object of this work is justified by the disadvantages found in MAnto with regard to the aspects that make it unusable. In addition, there are several methodologies presented by other authors such as UPON [14], which describes a very complete method based on the Unified Process of Software Development, but which is directed by use cases and has a more functional approach that is oriented to the ontology users' interactions rather than to the representation of a domain directed by data. In the same line, there is [13] that presents a lighter version of UPON. The NeOn methodological framework [23], although it presents a complete solution based on the creation of a glossary, nine scenarios for building ontologies and two life cycle models of ontology networks, in practice translates into a method that is not very agile when building ontologies. [8] presents a general framework for ontological engineering that offers very different paths to take for each problem, thus this generalist proposal is inappropriate for our purpose. [2] presents the SABiO methodology in version 2.0. This methodology offers a series of steps for the creation of ontologies based on the domain in order to provide the best possible description of the domain. These steps are of great interest; in fact, some of them will be picked up by the proposal presented in this paper. The main drawback of these proposals is that they are not based on data and do not provide living ontologies, as these aspects will be important in our proposal.

The methodology presented ways to increase agility through the introduction of two main features: the creation of living ontologies and a focus on being data-driven. These characteristics have led to the development of this methodology. Therefore, experts in the domain introduce qualities such as simplicity with this methodology, as it aims to follow a series of well-defined and structured steps that are easy to carry out, without it being necessary to incorporate ontology engineers in the initial phase of conceptualization of the domain to be annotated. The methods introduced are intended to allow the construction and maintenance of ontologies by those users that have a deep knowledge of the domain, such as domain experts, employees of entities related to the domain, or users who have a relationship or are familiar with the domain.

In summary, this methodology aims to follow guidelines well proven by other disciplines, such as Software Engineering, in which through several well-defined steps, as proposed by UPON [14], ontologies are obtained that are based on the domain, as proposed by SABiO [2]. As a new contribution, the methodology will be data-driven and will enable the creation of living ontologies. These characteristics are presented through a method structured in seven appropriately described steps, which direct the development of the ontology from obtaining the domain information to its evaluation. Following each one of the steps that compose this methodological framework is recommended in order to obtain a complete, high-ontology that covers all the presented domain, but not all the steps are obligatory, since there are multiple conditioning factors that establish the degree of complexity of each one of the steps according to the inherent complexity associated with the domain.

3. A data-driven methodology to generate living ontologies (MAOn)

The proposed methodology, called MAOn, is based on previous existing UPON [14] and SABiO [2] methodologies. All the phases of SABio have been incorporated to our proposal. With regard to UPON, only the phases applicable to the domain have been considered, because the proposal focuses on the creation of domain ontologies. In addition, new characteristics have been defined because they are needed to reinforce some aspects that are relevant to our purpose. Table 1 shows MAOn characteristics, both those related to UPON and SABiO and new ones, which are highlighted in grey.
The relevant contributions of our work are that MAOn proposes a data-driven process to develop living ontologies and a complete testing phase. Reference is made to data-driven in order to emphasize that the defined process incorporates the possibility of studying the data sources to include them as annotated data and, at this point, it could be possible to decide that new terminology needs to be defined. When reference is made to living ontology, that is intended to express the concept that the defined process facilitates the updating of the ontology, that is to say, to keep it active, in a simple way when necessary. To support both purposes, the process is defined in an iterative and incremental way, which facilitates incorporating new terminology by means of a new iteration of the process. MAOn moreover proposes a new and complete testing phase with respect to the other proposals. This testing phase is divided into five steps, which permit testing both the resulting ontology and the semantic annotated data by means of it.

<table>
<thead>
<tr>
<th>Characteristics of MAOn</th>
<th>UPON</th>
<th>SABiO</th>
<th>MAOn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterative and Incremental Process</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Domain study</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Knowledge of the domain using standards</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Identification of the objective</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Competency questions definition</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Obtaining competency requirements (functional requirements)</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
</tr>
<tr>
<td>Obtaining general aspects requirements (non-functional requirements)</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Search data sources</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Glossary</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Concepts and relationships identification</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>UML modelling</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ontological rules design</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ontology design</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ontological rules implementation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ontology implementation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Testing: Verification</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Testing: Validation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Testing: Semantic annotation of test data using ontology</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
</tr>
<tr>
<td>Testing: Generating a SPARQL Endpoint</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
</tr>
<tr>
<td>Testing: Use of annotated data by various applications</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
</tr>
</tbody>
</table>

The proposed methodology is structured in three different phases: (i) The conceptualization phase permits the determination of what specific information from the domain will be incorporated into the ontology; (ii) The construction phase defines design and implementation activities to carry out the ontology; (iii) The testing phase describes testing activities to validate the developed ontology and the resulting semantic dataset. Moreover, each phase is structured into a set of steps that are well defined and have specific objectives. Figure 1 shows the different phases and their corresponding steps.

The following subsections describe in depth MAOn phases and steps.
3.1. Conceptualization phase

This phase aims to determine what domain information has to be considered and incorporated into the ontology.

3.1.1. Step 1: Domain description

3.1.1.1. Step 1.a. Domain study

As mentioned before, the proposed methodology is focused on the development of the domain ontologies. For this objective, first of all, it is necessary to have an absolute knowledge of the domain, provided by the organization experts who participate in this development, or by the existing documentation related to this domain, such as standards, guides, handbooks, manuals, etc.

The domain knowledge is the input to this step (I.1.a). The description of the study domain as domain requirements is the output (O.1.a). It is recommended these requirements be enumerated as DRxx, where xx will be an ordinal number.

3.1.1.2. Step 1.b. Identification of the objective

In this step, the goal will be the identification of the specific objective of the ontology to be developed. This information will be collected in what will be called competency requirements. Each element of information should be considered as a different competency requirement. The recommendation is to enumerate each one by means of CRxx, where xx an ordinal number.

The entry receiving this step is the set of competency questions (O.1.c). The output will be the competency requirements (O.2.a).

3.1.2. Step 2: Requirements identification.

In the ontological engineering context, the requirements determine what information has to be represented by the ontology. Additionally, the ontology has qualities and general aspects that characterize it. Therefore, the proposal is to identify two different kind of requirements: those of competency and those of general aspects.

3.1.2.1. Step 2.a. Obtaining competency requirements.

This step identifies what information about the domain should take into account for the ontology to be developed. This information will be collected in what will be called competency requirements. Each element of information should be considered as a different competency requirement. The recommendation is to enumerate each one by means of CRxx, where xx an ordinal number.

The entry receiving this step is the set of competency questions (O.1.c). The output will be the competency requirements (O.2.a).

3.1.2.2. Step 2.b. Obtaining general aspect requirements.

To identify this kind of requirement, it is necessary to look at those characteristics, qualities and/or general aspects that are not related to the content of the ontology. In this sense, this type of requirement takes into account aspects related to availability, usability, extensibility, sources of knowledge or interoperability among others.

At this point, it is necessary to identify as requirements all those aspects collected by the objective of the ontology to be developed. These requirements are...
called general aspect requirements and it is recommended that each one be enumerated by means of GRyy, where yy is an ordinal number.

The entry receiving this step is the objective of the ontology identified above (O.1.b). The output will be the general aspect requirements (O.2.b).

3.1.3. Step 3: Search data sources

Based on competency requirements, a search is carried out for data sources that adequately cover the domain of study. These data will allow the identification of information that is of interest and that complements the competency requirements. These data must be analysed and incorporated, as an additional source of information, into the ontology. This information, together with the objective of the ontology, allows the scope of the ontology to be established with greater accuracy.

As an input to this step, the competency requirements (O.2.a) and the objective of the ontology (O.2.b) are considered. As an output, the actual data corresponding to real-world objects (O.3.1), from data sources, and new competency requirements are considered if new information has been discovered in the analysis of these data (O.3.2).

3.1.4. Step 4: Conceptualization of the domain.


This step proposes the elaboration of a glossary of terms where the relevant terms of the domain are described. It is important to eliminate any ambiguity or doubt in relation to the concepts you are going to work with. The recommendation is to create a list, alphabetically ordered, with an entry by each domain term included.

The entry in this step are the domain requirements (O.1.a) and the real-world objects data (O.3.1), obtained in the previous step. As an output, the glossary of domain terms is obtained (O.4.a).

3.1.4.2. Step 4.b. Conceptualization of the real world.

This step works with real-world objects that ontology is going to consider; that is, which glossary terms are real-world objects to consider or include. At this point, it is important to stop and define what the real world is: real-world objects are the terms or concepts that appear in the concrete data to be analysed and that are susceptible of semantic annotation. Once determined, the existing relations between them are identified in order to provide a set of interconnected concepts associated with the domain. This knowledge will be modelled in the form of a UML class diagram (Unified Modelling Language) [17]. This diagram will specify the real-world objects, their characteristics as attributes, and the relationships between these objects indicating their cardinality and roles, when necessary.

As part of this step, the real-world objects data (O.3) and the glossary of domain terms (O.4.a) are considered. The output is a UML conceptual class diagram (O.4.b), which supports the conceptualization of the real world within the domain.

3.2. Construction phase

In this phase the design and implementation of the ontology to be developed is carried out, in order to actualize the ontological scheme.

3.2.1. Step 5: Ontology design

This step carries out both the design of the ontological rules, where the restrictions are included, and the design of the ontology itself through the definition of a schema.

3.2.1.1. Step 5.a. Ontological rules design

Based on the information gathered by the competency requirements and by the relationships established in the UML conceptual class diagram, it is necessary to identify the set of restrictions established by the domain, which will determine the characteristics of the ontology. Once the constraints have been identified, it is necessary to design the rules that model them in terms of predicate logic.

As an input to this step, the competency requirements (O.2.a) and the UML conceptual class diagram (O.4.b) are considered. An output will be the design of ontological rules (O.5.a) are considered.

3.2.1.2. Step 5.b. Ontology schema design

This step aims at designing the schema that supports the ontology. In order to do so, the competence requirements provided by the information to be represented by the ontology will be taken into account, as well as the requirements of the general aspects that will characterise the ontology through the qualities and general aspects as defined.

As an input to this step, competency requirements (O.2.a), general aspects requirements (O.2.b) and UML conceptual class diagram (O.4.b) are considered. The output will be the ontology schema (O.5.b).
3.2.2. Step 6: Ontology implementation

This step is oriented to the implementation of the ontology and the rules associated with it. To do this, it is necessary to choose the appropriate languages that support both implementations.

3.2.2.1. Step 6.a. Ontological rules implementation

In this step, the ontological rules are implemented. Examples of rule languages could be Semantic Web Rule Language\(^\text{10}\) (SWRL) and Rule Markup Language\(^\text{11}\) (RuleML).

As an input to this step, the ontological rules (O.5.a) are considered. As an output, they are implemented in the chosen rule language (O.6.a).

3.2.2.2. Step 6.b. Ontology implementation

In this step, the implementation of the ontology is carried out, that is to say, the codification in the form of triples of the set of semantic sentences that will give coverage to the terms of the domain that are of interest for the ontology. Examples of languages that implement ontologies could be DARPA’s Agent Markup Language and Ontology Inference Layer\(^\text{12}\) (DAML+OIL), Ontology Web Language\(^\text{13}\) (OWL) or RDF Schema\(^\text{14}\) (RDFS).

It should be noted that it includes the recommendation to reuse terminology of existing ontological vocabularies that are widely extended. That is, if there is information previously disclosed, it is convenient to make use of existing semantic structures. This step proposes the validation of the coherence of the ontology with respect to the competency requirements. The proposal is to check whether each one of the competency requirements identified above is supported by the terminology contained in the ontology. For this, the recommendation is to use a table with two columns where each of the requirements is associated with the term of the corresponding ontology. These will be implemented through semantic structures generated specifically for the ontology.

As an input to this step, the ontology schema (O.5.b) is considered. The output will be the ontology code (O.6.b.1) and a list of the terms included in the ontology, as well as a set of terms from other vocabularies that support the study domain (O.6.b.2).

3.3. Testing phase

The objective of this phase is the validation of the methodology through the developed ontology. This type of validation (through a case study) is also used in other works (e.g., [2],[14] and [23]), considering that the methodological proposal is correct if the result obtained after its application is valid.

Based on this prior consideration, this paper proposes a set of validation steps that will determine the appropriateness and validity of the ontology developed.

3.3.1. Step 7: Ontology validation

3.3.1.1. Step 7.a. Verification

This first step proposes to carry out the verification that the ontology is constructing correctly through the syntactic correction of the implemented ontology. This step could be omitted if there has been the use an ontology development tool in the implementation — for example, Protégé\(^\text{15}\) — since in this case the syntactic correction is implicit. In the case of not using this type of tool, the proposal is to use a validator, such as the one provided by the W3C\(^\text{16}\), that verifies the syntactic correction of the semantic code implemented by the ontology.

As an input to this step, the ontology implemented (O.6.b.1) is considered. The output will be the result of this verification (O.7.a).

3.3.1.2. Step 7.b. Validation

This step proposes to validate the coherence of the ontology with respect to the competency requirements. The proposal is to check whether each one of the competency requirements identified above is supported by the terminology contained in the ontology. For this, the recommendation is to use a table with two columns where each of the requirements is associated with the term of the corresponding ontology. Table 2 shows an example.

As entries to this step, the competency requirements (O.2.a) and the list of ontology terms (O.6.b.2) are considered. The output produced by this step will be the result of this validation (O.7.b).

---

\(^{10}\) https://www.w3.org/Submission/SWRL/

\(^{11}\) http://wiki.ruleml.org/index.php/RuleML_Home

\(^{12}\) http://www.daml.org/

\(^{13}\) https://www.w3.org/OWL/

\(^{14}\) https://www.w3.org/TR/rdf-schema/

\(^{15}\) https://protege.stanford.edu/

\(^{16}\) https://www.w3.org/RDF/Validator/
3.3.1.3. Step 7.c. Data semantic annotation

This step proposes the semantic annotation of data from previously identified data sources in order to obtain a dataset in RDF format. Once this step has been carried out, the completeness of the ontology can be determined, since, if the terminology described by the ontology supports the annotation of the data, the ontology is complete.

Syntactic correction of semantically annotated data (RDF format) must also be ensured. For this objective, an RDF validator\(^\text{17}\) is recommended. This will allow the correction of the annotated data to be determined.

As entries, the data from the sources (O.3.1) and the list of ontology terms, including the necessary terms from other vocabularies for this domain, (O.6.b.2) are considered. As output, the dataset is in RDF format (O.7.e).

3.3.1.4. Step 7.d. Endpoint for SPARQL queries

This step proposes the creation of a mechanism of data queries through SPARQL to carry them out on the generated dataset. Protocol and RDF Query Language (SPARQL\(^\text{18}\)) is a standardized language for querying RDF graphs. To do this, the proposal is to create a SPARQL endpoint and define a set of test cases based on competency requirements that will be executed as a query through the endpoint and whose result will determine the correction of the ontology.

As entries in this step, the competency requirements (O.2.a) and the dataset (semantically annotated data) (O.7.e) are considered. The output is the endpoint (O.7.d.1) and the validation of the ontology in terms of its correction (O.7.d.2) are considered.

3.3.1.5. Step 7.e. Software Applications

The final step proposes real applications that offer services to meet the objective of the ontology (step 1.b) and support competency questions (step 1.c). These applications will make use of the semantically annotated data (or dataset) (step 7.c), making feasible the correction and completeness of the ontology as long as the dataset is useful and fulfils the expectations of the real developed applications.

As entries in this step, the objective of ontology (O.1.b), competency questions (O.1.c) and dataset (O.7.c) are considered. The output consists of real applications that make use of annotated data through ontology (O.7.e.1) and validation of ontology with respect to its objective and to competency questions (O.7.e.2).

In summary, the process described in the previous steps and as a reference script is designed to provide clarity and facilitate its use (see Table 3).

### 4. Evaluation

A study of the existing literature shows that there are different studies regarding the evaluation of methodologies or methods for the development of ontologies, but these are not very numerous. This conclusion is derived from what is expressed in [23], where it is explained that it is very difficult to evaluate a methodology objectively, since experimentation as a form of evaluation is not feasible because the development process is complex. Most of the proposed solutions, such as [2], [25], [16], [24], etc., are aimed at the evaluation of the product obtained, the ontology. This is why, in general, the criteria that guide the evaluation of methodologies do not follow a well-defined and structured process, but rather it is up to the ontology engineer to establish the criteria that evaluate the methodology.

The validation of the methodology developed in this article is based on verifying and validating the method carried out from various perspectives: (1) development of a case study with experts in the public transport domain. (2) Evaluation of the ontology based on a set of general characteristics of formal ontology validation. (3) Preparation of a questionnaire addressed to experts in the field of public transport and experts in ontological engineering as a method of evaluating the process proposed in the methodology.

#### 4.1. Case study

The main method of evaluation of this methodology is based on the development of a case study. The case studies provide a deep understanding of the developed methodology, allowing the researcher to understand how the methodology is adapted to the context [18].

---

\(^{17}\) https://www.w3.org/RDF/Validator/

\(^{18}\) https://www.w3.org/TR/rdf-sparql-query/
This case study presents the development of an ontology in the field of public transport and accessibility information, in particular for the specific scenario of the Metro network\(^{19}\) of the city of Madrid in Spain.

### Table 3 Summary schema of the steps described by the methodology with the inputs and outputs

<table>
<thead>
<tr>
<th>Conceptualization phase</th>
<th>Input</th>
<th>Step</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1. Domain description</td>
<td>Domain knowledge</td>
</tr>
<tr>
<td>Domain knowledge</td>
<td>- Domain study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domain requirements</td>
<td>- Identification of the objective</td>
<td>Objective of the ontology</td>
<td></td>
</tr>
<tr>
<td>Objective of the ontology</td>
<td>- Competency questions definition</td>
<td>Competency questions</td>
<td></td>
</tr>
<tr>
<td>2. Requirements identification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Competency questions</td>
<td>- Obtaining competency requirements</td>
<td>Competency requirements</td>
<td></td>
</tr>
<tr>
<td>Objective of the ontology</td>
<td>- Obtaining general aspects requirements</td>
<td>General aspects requirements</td>
<td></td>
</tr>
<tr>
<td>Competency requirements and objective of the ontology</td>
<td>3. Search data sources</td>
<td>Real-world objects data and new competency requirements</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Conceptualization of the domain</td>
<td></td>
</tr>
<tr>
<td>Domain knowledge and real-world objects data</td>
<td>- Glossary of domain terms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glossary of domain terms</td>
<td>- Conceptualization of the real world</td>
<td>UML conceptual class diagram</td>
<td></td>
</tr>
<tr>
<td>Construction phase</td>
<td></td>
<td>5. Ontology design</td>
<td></td>
</tr>
<tr>
<td>Competency requirements and UML conceptual class diagram</td>
<td>- Ontological rules design</td>
<td>Ontological rules</td>
<td></td>
</tr>
<tr>
<td>Competency requirements, general aspects requirements and UML conceptual class diagram</td>
<td>- Ontology schema design</td>
<td>Ontology schema</td>
<td></td>
</tr>
<tr>
<td>6. Ontology implementation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ontological rules</td>
<td>- Ontological rules implementation</td>
<td>Ontological rules implemented</td>
<td></td>
</tr>
<tr>
<td>Ontology schema</td>
<td>- Ontology implementation</td>
<td>Ontology implemented and ontology terms</td>
<td></td>
</tr>
<tr>
<td>Testing phase</td>
<td></td>
<td>7. Ontology validation</td>
<td></td>
</tr>
<tr>
<td>Ontology implemented and competency requirements</td>
<td>- Verification</td>
<td>Verification result</td>
<td></td>
</tr>
<tr>
<td>Ontology terms and real-world objects data</td>
<td>- Validation</td>
<td>Validation result</td>
<td></td>
</tr>
<tr>
<td>Dataset and competency requirements</td>
<td>- Endpoint for SPARQL queries</td>
<td>Dataset</td>
<td></td>
</tr>
<tr>
<td>Objective of the ontology, competency questions and dataset</td>
<td>- Software Applications</td>
<td>Software Applications and ontology validation</td>
<td></td>
</tr>
</tbody>
</table>

#### 4.1.1. Conceptualization phase

##### 4.1.1.1. Step 1. Domain description

This step focuses on describing in depth the domain in which the ontology is based. This step is divided into the following sub-steps:

**Step 1.a. Domain study.**

Previous work studied TRANSMODEL and IFOPT, and, as mentioned before, they are standards related to the public transport context. In this sense, they have provided information about the transport public domain and its accessibility characteristics.

For this step, the input elements include the domain knowledge, which is given by standards in this specific case study. As output, the following domain requirements are obtained:

- **DR1.** The public transport is usually structured on transport lines.
- **DR2.** Each line follows a specific route.
- **DR3.** A public transport line is structured as an ordered sequence of stops, which conform to the route of that line.
- **DR4.** A station belongs to one or more transport lines.
- **DR5.** A station serves one or more means of transport.

\(^{19}\) [https://www.metromadrid.es/en](https://www.metromadrid.es/en)
DR6. A station consists of one or more stops (also called quays) and has one or more entrances/ exits.

DR7. A station may have different accessibility features associated with transport elements.

DR8. A station has different accessibility elements such as stairs, elevators, ramps, escalators or travelators (moving walkways).

DR9. Passengers on public transport may have different accessibility needs.

DR10. Public transport may or may not be used by passengers with special needs, depending on the accessibility elements available.

Step 1.b. Identification of the objective

It is necessary to identify the objective of the ontology in order to determine the scope and the limits of the domain knowledge.

For this step, the input element is the domain description. As output, it obtains the following list of items that refer to the objective of the ontology:

PO01. Publish open linked data with information on the transport elements and their accessibility to the Metro de Madrid. To do this, a dataset of semantically annotated data will have to be generated from the different sources that offer Metro de Madrid data. This dataset will be the result of integrating and unifying the data from existing sources and, in this way, providing a single source of data that is unified and enriched with transport and accessibility data from Metro de Madrid.

PO02. Generate an open and linked dataset with information regarding the Metro de Madrid transport and its accessibility elements. The purpose of this dataset is to be able to use it for the development of information services for apps, applications and web pages that provide these data when required and enable the calculation of accessible routes in this area in the city of Madrid.

Step 1.c. Competency questions definition.

Bearing in mind the objective of the ontology to be developed, which is received as an entry to this step, it is necessary to raise a series of issues called competency questions, which establish those questions that must be answered by the ontology itself and which in this case study are the following:

CQ1. Is it possible to know the stations on a route between an origin station and a destination station on the Metro de Madrid? Is it possible to obtain all the accessibility information on these stations?

CQ2. Is it possible to know the connections of a station with other type of public transport?

CQ3. Is it possible to find out the connections between Metro de Madrid lines?

CQ4. Is it possible to know which Metro de Madrid line or lines pass through a specific station?

CQ5. Is it possible to know the location on a map of a specific Metro de Madrid station?

CQ6. Is it possible to know the tariff zone to which a specific Metro de Madrid station belongs?

CQ7. Is it possible to know the postal address of a Metro de Madrid location?

CQ8. Is it possible to know if a certain Metro de Madrid station is accessible to a person with a wheelchair?

CQ9. Is it possible to know if a certain Metro de Madrid station is accessible to a person carrying a twin baby carriage or carrying luggage?

CQ10. Is it possible to know whether a certain Metro de Madrid station is accessible to a visually or hearing-impaired person?

CQ11. Is it possible to know whether a certain Metro de Madrid station is accessible to a person who has certain phobias (fear of lifts, fear of escalators, etc.)?

After completing this step, the competency questions that cover the domain to be represented by the ontology are obtained as an output. Therefore, at the end of this first complete step, a complete description of the domain to be covered by the development of the ontology is obtained.

4.1.1.2. Step 2. Requirements identification

This step will allow us to discover everything that the ontology to be developed should collect. The process of requirements identification is divided into two steps:

Step 2.a. Obtaining competency requirements.

This step is aimed at obtaining the content that must be described by the ontology. In this sense, a list of requirements is obtained that includes all the content
to be modelled and that is of interest in the given domain. As an entry to this step, competency questions are received. As an exit from this step, the competency requirements are obtained, which will provide information on the competences established by the domain:

CR1. Information on types of public transport.
CR2. Information on line distribution.
CR3. Information on connections between different lines.
CR4. Information on connections between different types of public transport.
CR5. Information on the disposition of stations, stops and entrances/exits of stations.
CR6. Information on the association between stations, stops and entrances/exits.
CR7. Information on infrastructure related to accessibility: stairs, elevators, ramps, escalators or travelators.
CR8. Geographical information on stations, stops and entrances/exits.
CR9. Information on the locations of stations, stops and entrances/exits through postal addresses.
CR10. Information on the tariff zone associated with the stations.
CR11. Information on the time zone of application to public transport.
CR12. Information on wheelchair accessibility at the station and its stops and entrances/exits.

Step 2.b. Obtaining general aspects requirements

Following a somewhat different approach, this step has as its objective the compilation of all those general aspects or restrictions that are applicable to the ontology to be obtained in the form of requirements. Therefore, the entry to this step is established through the objective of the ontology.

GR01. The ontology must be available in open mode for free use by anyone.
GR02. The ontology must be understandable by systems and by people who know the notation used by the Semantic Web.
GR03. The ontology must be usable by any system that wishes to process it or any person that wishes to consult it.
GR04. The ontology must be easily maintainable, thus facilitating the extension of the ontology by means of new terms to be incorporated.

GR05. The ontology must be developed based on a deep knowledge of the domain or based on standards.
GR06. The ontology must allow interoperability with other existing reference ontologies in such a way that simple integration between them is possible.

The previous list includes the requirements of the general aspects of the ontology that are established as a way out of this step.


This step seeks to obtain data sources that collect information about the proposed scenario for this case study. Therefore, this step receives as input the competency requirements and the objective of the ontology. They are used to search for data sources that offer information about Metro de Madrid's infrastructure and the elements that promote accessibility in this means of transport. The data sources that provide the data files to be used in this case study are: (i) the open data portal of the Consorcio de Transportes de Madrid and (ii) the Metro de Madrid website. From these data sources (DS), the following have been obtained, respectively:

DS01. A GTFS (General Transit Feed Specification) open data file in CSV format.
DS02. Different XML files by means of web scraping techniques.

Based on the information gathered from the different data sources selected, the output of this step is offered in the form of domain data. In addition, optionally, if new knowledge is discovered through the data, new competency requirements will be obtained.

4.1.1.4. Step 4. Conceptualization of the domain

This step includes the modelling of the real-world objects associated with the domain. It should be noted that the work carried out in this conceptualization is supported by the information acquired from previous studies of the TRANSMODEL and IFOPT standards.

Step 4.a. Glossary of domain terms

In order to enhance comprehension of each of the terms described by the treated domain and to offer clarity in what it is desired to model, a list of terms with their associated definitions has been elaborated. These definitions are derived, on the one hand, from what is provided by the TRANSMODEL and IFOPT standards, in the form of knowledge about the domain, and, on the other hand, from the data obtained through the data sources found. These elements are received as

---

20 http://datos.crtm.es/

21 https://www.metromadrid.es/en
input to this step. This list is presented in the form of a glossary that includes all the terms that are relevant for this case study and that are named according to the TRANSMODEL and IFOPT standards already mentioned. As a result of this step, the glossary of domain terms (Domain Term -DT-) shown below will be obtained:

- **DT01. ACCESSIBILITY:** The ability of a user with special needs, such as a disability or impediment, to access fixed or mobile public transport services.
- **DT02. ACCESSIBILITY ASSESSMENT:** Accessibility characteristics of an entity used by passengers, such as a station or one of its components.
- **DT03. ACCESSIBILITY LIMITATION:** Categorization of accessibility characteristics associated with transport elements, such as a station or one of its components, which establishes whether it is usable by passengers with special needs.
- **DT04. ADDRESS:** Information that describes the situation of a site uniquely in a geographical context for identifying it.
- **DT05. CONNECTION LINK:** A passenger’s ability to change lines or to transfer on public transport to continue a route.
- **DT06. DIRECTION:** Classification of an orientation associated with a route.
- **DT07. ENTRANCE:** Point of entry or exit of travellers from a stop, station or point of interest. They may or may not have an access element in the form of a door, etc.
- **DT08. ESCALATORS:** A person-carrying device consisting of an inclined staircase whose steps move up or down.
- **DT09. LATITUDE:** Geographic coordinate that provides the location of a place north or south in relation to the equator.
- **DT10. LIFT:** Lifting device used to transport people or objects vertically from one level to another in a building or structure.
- **DT11. LINE:** Group of routes generally known to travellers by name or number.
- **DT12. LONGITUDE:** Geographical coordinate providing the location of a place east or west of the Greenwich meridian.
- **DT13. PUBLIC TRANSPORT:** Any means of transport for the general public; usually owned or managed by government agencies, public companies, private companies or both.
- **DT14. QUAY:** A place in the form of a platform, room or wharf where passengers have access to public transport vehicles.
- **DT15. RAMP:** Architectural element that has the function of communicating two planes of different level, so that different heights in a certain space are joined.
- **DT16. ROUTE:** An ordered list of points defining a simple path through a transport network.
- **DT17. SUITABILITY:** Indicates the possibility of using a particular infrastructure for a traveller with special needs.
- **DT18. STOP PLACE:** A place that comprises one or more locations where vehicles can stop and where travellers can board or alight from vehicles or prepare their routes. Normally, they will have one or more well-known names.
- **DT19. STOP POINT:** Point on a route where passengers can get in or out of the public transport vehicle.
- **DT20. TARIFF ZONE:** Zone used to define a tariff structure per zone.
- **DT21. TIME ZONE:** Areas into which the Earth is divided to establish appropriate local times.
- **DT22. TRANSFER:** The means for a traveller to change from one public transport to another in order to continue a route.
- **DT23. TRANSPORT MODE:** Characterisation of the mode of operation of a means of transport (metro, tram, train, bus, etc.).
- **DT24. TRAVELATOR:** A transport system in the form of a conveyor belt similar to an escalator, but moving horizontally and lacking steps.
- **DT25. USER NEED:** A traveller’s need for accessibility.
- **DT26. VEHICLE:** Public transport vehicle used to carry passengers.
- **DT27. WHEELCHAIR ACCESS:** Path or route free of architectural barriers that allows the transit of people in wheelchairs with various mobility needs.

**Step 4.b. Conceptualization of the real world**

This step represents those real-world objects previously identified through a UML conceptual class diagram. To do this, those terms from the glossary, received as input to this step, which are applicable to the scope of this case study of Metro de Madrid, are selected. Table 4 shows the list of terms.
These elements are interconnected themselves, as seen before in step 1.a. This step adds the specific characteristics of Metro de Madrid regarding to the data sources:

- Metro de Madrid is a public transport system that is structured along lines.
- Each line has a specific direction.
- The lines have two directions in terms of how to travel them.
- A station consists of one or more platforms (also called quays) and has one or more entrances/exits.
- A line of the Metro de Madrid consists of stations, which have a certain order, which form the route of the line.
- A station belongs to one or more Metro de Madrid lines.
- A quay belongs to a single line of the Metro de Madrid.
- A station is associated with one or more means of transport.
- Metro de Madrid stations have accessibility elements such as stairs, elevators, ramps, escalators or travelators.

<table>
<thead>
<tr>
<th>Glossary terms</th>
<th>Icon</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td></td>
<td>Tariff zone</td>
</tr>
<tr>
<td>Connection link</td>
<td></td>
<td>Longitudinal Directions</td>
</tr>
<tr>
<td>Entrance/exit</td>
<td></td>
<td>Public transport</td>
</tr>
<tr>
<td>Escalators</td>
<td></td>
<td>Quay</td>
</tr>
<tr>
<td>Latitude</td>
<td></td>
<td>Ramp</td>
</tr>
<tr>
<td>Lift</td>
<td></td>
<td>Step place</td>
</tr>
</tbody>
</table>

It should be noted that although Metro de Madrid only specifies stops within a line, sometimes the accessibility of this transport network is specified at the platform level to indicate that at that stop only one of the platforms is accessible. For this reason, the IFOPT quay term is chosen, which refers to the platform. With this, accessibility can be expressed at quay level and not at stop level. As an example, at the Avenida de América station, specifically for line 6, it has three quays for access to the station, but only one of them, the central one, has escalators, and they only go up, so it is not accessible for those going down, so it is listed as not accessible because it does not have an elevator.

As shown in Figure 2, the output established for this step is the list of real-world concepts and those relationships that are established between them, all in the form of a UML conceptual class diagram that describes the domain of study for ontology.

**4.1.2. Construction phase**

This phase has designated, in terms of logic, a series of rules that model those constraints that have not been reflected through the UML conceptual class diagram. Additionally, the bases of the ontology to be built are established and realized through the design work.

**4.1.2.1. Step 5. Ontology design**

In this step, design tasks related to the identification of rules and the design of the ontological schema are performed.

**Step 5.a. Ontological rules design**

In addition to what has been obtained so far, a series of restrictions arise that are not covered by the UML conceptual class diagram elaborated in the previous step. Based on this premise, the entry to this step are the competency requirements and the UML conceptual class diagram. Through these elements, a series of ontological rules are described, which are offered as an exit from this step. These rules allow giving a place to the restrictions inherent to the domain and that have not been possible to collect otherwise.

The rules obtained are shown below:

- **R1.** All Quay Qi belongs to one and only one Line Lj and in turn belongs to a Public Transport PTk. Then \( \forall i \quad \forall j \quad \forall k \quad \exists Q_i \in L_j \in P_Tk \).

- **R2.** A Line Lj belonging to a Public Transport PTk is characterised as an ordered list composed of a homogeneous sequence of Quays Qi. Then \( \forall i \exists Q_i \in L_j \) then Lj{Q1, Q2, Q3, Q4, …, Qn} where \( n \geq 2 \mid n \in \mathbb{Z} \).
R3. The elements of the list have the property of being ordered sequentially, according to the positions they occupy in the line. hasOrder (Lj) therefore: Qi precedes Qi+1 for i = 1, 2, 3, ..., n-1 and Qi succeeds Qi-1 for i = 2, 3, 4, ..., n. This means that each element has a single predecessor, except the first, and a single successor, except the last.

R4. For every Line Lj there is a list ordered in inverse order Lj’ which describes another way of crossing the Line Lj. Then ∀Lj ∃ Lj’ where Lj’ ≡ inverseDirection (Lj) ∧ sameAs (Lj’, inverseOf (Lj)).

Step 5.b. Ontology schema design
The conceptualization of the real world previously performed through the UML conceptual class diagram, as well as the requirements obtained in step 2, serve as an entry to this step. Through these entries, the guidelines for designing the ontology are established in terms of a schema. This schema will serve as the basis for the implementation of the ontology in a later step. The output offered by this step is therefore the ontology schema to be developed that will greatly facilitate the actualization of the ontology in the form of semantic sentences. Figure 3 shows a partial schema of the ontology to be developed.

Figure 3. Partial schema of the data properties of the ontology

4.1.2.2. Step 6. Ontology implementation
In a next step, the implementation is carried out, in which first the choice of ontology language and rules is made. Subsequently, a series of rules associated with the restrictions of the domain itself are implemented, enriching the ontology to be built. Finally, through the description of the different concepts acquired in the previous step, this step uses sentences to describe in semantic terms the conceptualization of the domain. To this end, another proposal is to encourage the reuse of terms that coincide in terms of the semantics of the concept described, through the reuse of terms that have already been defined in widely recognized vocabularies. For the rest of terms that do not coincide exactly, or have simply not been expressed in semantic terms, the creation of new semantic resources that represent these concepts is proposed.

Step 6.a. Ontological rules implementation
Starting from the design of ontological rules, as an entry to this step, the restrictions found in the domain are implemented, in such a way that the expressiveness of the ontology is increased through the inclusion of this type of rules. Therefore, based on the given design, the following rules implemented as an output of this step through the Semantic Web Rule Language (SWRL) are extracted (see Table 5).
Step 6.b. Ontology implementation

Analysing the different concepts or terms considered relevant to this case study, which have been obtained through the ontology schema, which is received as an entry to this step, a list of terms that have been described by well-known ontologies is obtained. This list of terms is expressed through the Ontology Web Language (OWL).

<table>
<thead>
<tr>
<th>Table 5 Ontological rules of the domain implemented through SWRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
</tr>
<tr>
<td>R1</td>
</tr>
<tr>
<td>R2</td>
</tr>
<tr>
<td>R3</td>
</tr>
<tr>
<td>R4</td>
</tr>
</tbody>
</table>

This practice of reusing terms enhances the integration and enrichment of ontology through existing knowledge. In this sense, three existing ontologies are used: (i) Schema.org offers a collection of vocabularies that describe generic items in order to enrich Web pages and facilitate the work of Internet search engines. (ii) WGS84 Geo Positioning offers a vocabulary to represent the latitude, longitude and altitude offered by the WGS84 global geographic coordinate system. (iii) Geonames makes it possible to add geospatial semantic information to the Web. From the terms analysed, it can be deduced that the following resources can be reused through Table 6.

<table>
<thead>
<tr>
<th>Table 6. List of reused terms from other ontologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology</td>
</tr>
<tr>
<td>Schema.org</td>
</tr>
<tr>
<td>WGS84 Geo Positioning</td>
</tr>
<tr>
<td>WGS84 Geo Positioning</td>
</tr>
<tr>
<td>Geonames</td>
</tr>
<tr>
<td>Geonames</td>
</tr>
</tbody>
</table>

On the other hand, for those terms that are intended to describe and that do not coincide with existing semantic terms on the Web, the proposal is to include them in the ontology through resources generated ad-hoc for this ontology. Making use of the mao alias and terms that are as self-descriptive as possible, the following list of ontological terms in the form of semantic resources is obtained through Table 7.

<table>
<thead>
<tr>
<th>Table 7. List of new ontology terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>URIs</td>
</tr>
<tr>
<td>mao:hasLift</td>
</tr>
<tr>
<td>mao:hasEscalator</td>
</tr>
<tr>
<td>mao:hasTravelator</td>
</tr>
<tr>
<td>mao:hasRamp</td>
</tr>
<tr>
<td>mao:transfer</td>
</tr>
<tr>
<td>mao:stopplaceCode</td>
</tr>
<tr>
<td>mao:address</td>
</tr>
<tr>
<td>mao:connectionLink</td>
</tr>
</tbody>
</table>

The following code shows a MANto ontology fragment, which is implemented by means of the Ontology Web Language (OWL). The complete definition of MANto can be downloaded from the GitHub repository.

```owl
This code has its corresponding partial schema, represented as a graph, which is obtained by means of the W3C RDF tool. Figure 4 shows the graph.

---

22 https://www.w3.org/OWL/
23 https://raw.githubusercontent.com/vortic3/LinkedUnifiedDataset/master/MAnto_Lite_ontology.rdf
24 https://www.w3.org/RDF/Validator/
4.1.3. Validation phase

Finally, the evaluation phase of the ontology obtained is carried out in order to verify and validate the ontology. In this step a series of alternatives are proposed that support these ontology tests.

4.1.3.1. Step 7. Ontology validation.

Step 7a. Verification

This step checks that there are no syntax errors and, therefore, that the generated code is correct according to the language used. The entry to this step, therefore, is the implemented ontology.

For this objective, an RDF document validator\(^\text{25}\) is used to verify whether the ontology created is syntactically correct. Figure 5 shows a capture of the result returned by the validator when the ontology was passed to him.

By means of the output obtained, it can be verified that the ontology is being constructed correctly. For all these reasons, the output established for this step is the result of the verification.

\(^\text{25}\) https://www.w3.org/RDF/Validator/
Step 7.b. Validation

This step is intended to validate the ontology developed, with regard to the requirements previously identified. Therefore, the entries to this step are the implemented ontology and the competency requirements. Then, the OWL code has been checked for compliance with the competency requirements identified in step 2.a. An example of those requirements is that the Metro de Madrid is public transport that offers a set of lines, stations, stops and entrances/exits. The following fragment of MAOn OWL code represents the classes, which support those requirements:

```
<owl:Class rdf:about="http://com.vortic3.MANTO#Line"/>
<owl:Class rdf:about="http://com.vortic3.MANTO#StopPlace"/>
<owl:Class rdf:about="http://com.vortic3.MANTO#Entrance"/>
<owl:Class rdf:about="http://com.vortic3.MANTO#Quay"/>
```

Table 8 summarizes the check between the competency requirements and the MAOn terms on which they are based.
Figure 5. Screenshot with the result of the ontology by means of a validator

Table 8. Association between competency requirements and ontology terms

<table>
<thead>
<tr>
<th>Competency requirement</th>
<th>MAOn’s terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR1. Types of public transport</td>
<td>mao:PublicTransport</td>
</tr>
<tr>
<td>CR2. Line distribution</td>
<td>mao:Line</td>
</tr>
<tr>
<td>CR3. Connections between different lines</td>
<td>mao:hasConnectionLink</td>
</tr>
<tr>
<td>CR4. Connections between different types of public transport</td>
<td>mao:hasTransfer</td>
</tr>
<tr>
<td>CR5. The disposition of stations, stops and entrances/exists</td>
<td>mao:StopPlace</td>
</tr>
<tr>
<td>CR6. Associations between stations, stops and entrances/exists</td>
<td>mao:Entrance</td>
</tr>
<tr>
<td>CR7. Infrastructures related to accessibility</td>
<td>mao:hasLift</td>
</tr>
<tr>
<td>CR8. Geographical information on stations, stops and entrances/exists</td>
<td>mao:Quay</td>
</tr>
<tr>
<td>CR9. Postal address of stations, stops and entrances/exists</td>
<td>mao:Ramp</td>
</tr>
<tr>
<td>CR10. The tariff zone associated with the stations</td>
<td>mao:hasTravelator</td>
</tr>
<tr>
<td>CR11. The time zone of the application to public transport</td>
<td>mao:Escalator</td>
</tr>
<tr>
<td>CR12. Wheelchair accessibility</td>
<td>mao:wheelchairAccess</td>
</tr>
</tbody>
</table>

As can be seen in the table above, there is a direct correspondence between the competence requirements and the semantic terms expressed in the ontology; this process validates the ontology relative to what it wanted to obtain. As an exit to this step, the results, in this case correct, of the validation are established.

Step 7.c. Data semantic annotation
This step receives as input the list of terms of the implemented ontology and the data obtained from the studied data sources. The MAOn ontology, obtained as a result of applying the methodology proposed in this work, has been used to annotate and integrate transport and accessibility data about the Metro de Madrid from different sources. The resulting dataset is named Dataset_v2 and it can be downloaded from our semantic repository. The dataset, which will later also be used as a validation object in this phase, is an output artefact of this step.

Step 7.d. Endpoint for SPARQL queries
In order to be able to consult information on the Metro de Madrid and its accessibility elements, use is
made of the dataset generated and the competency requirements, which are the entrance to this step. To this end, a SPARQL endpoint\(^{27}\) has been developed. The endpoint offered as an output of this step has been generated through an Apache Jena Fuseki\(^{28}\) server. Figure 6. Query SPARQL supporting the CR2 shows a partial capture of a SPARQL query, which requests the lines of public transport from the dataset, and supports the competency requirement 2 (CR2).

The results of this SPARQL query show the corresponding lines of Metro de Madrid: line 1, line 2, line 3, line 4, line 5, line 6, line 7, line 8, line 9, line 10, line 11, line 12, line R, line ML1, line ML2, line ML3 and line ML4.

In order to ensure the correctness of the dataset, two different activities have been carried out. The first one, carried out by people familiar with the dataset, consisted of carrying out a consultation for each term of the ontology. The objective of this activity was two-fold: on the one hand, the aim was to determine possible data annotation errors (for example, the postal address of a station must be annotated with the term of the ontology mao:address. Therefore, a query is made to determine the postal address of the stations by means of that term and it is verified that the content is a postal address). On the other hand, it was to determine if the correct annotation, the data coming from the source and following the specification of the same one, was not fulfilled. (For example, that the specification of the source indicates that this particular data is a postal address, but that the data associated with it are not). In fact, this last aspect was decisive in correcting some errors coming from the data source and is reflected in [6]).

The second activity also consists of carrying out a set of queries on such data, which, as in the previous

\(^{27}\) http://coruscant.my.to:3030/dataset.html?tab=query&ds=dataset_metro

\(^{28}\) https://jena.apache.org/documentation/fuseki2/
activity, was intended to determine whether the result was correct; but in this case was to be carried out with people who were not familiar with the MAOn ontology or the dataset. For this purpose, a set of five test cases based on the CR2, CR3, CR4, CR5, CR6 and CR7 competency requirements were defined. Next, the different tests are described, which are ordered from the lower to the higher degree of difficulty, and which indicate the corresponding competency requirements supported by them:

- T1. Determine whether a particular station has a stair, escalator, lift, ramp or travelator: CR7.
- T3. Obtain stations for a specific line that correspond to other lines: CR3, CR5 and CR6.
- T4. Know if a given station (or any station) corresponds to a means of transport: CR4.
- T5. Obtain the lines that pass through a given station: CR3, CR5 and CR6.

This activity has been carried out with 16 people who had higher academic training in various branches such as mathematics, statistics, computer science, engineering, etc., and matriculated in a master’s program of the Rey Juan Carlos University. The students did not have any previous experience in the use of SPARQL, although they did use other query languages such as SQL (Structured Query Language). Therefore, they were given a 30-minute training on the use of SPARQL, then they carried out the queries. Some students could not carry out any tasks, a result which has been called uncompleted tasks. However, the tasks that were completed, which are called completed tasks, gave correct results in all cases. This result proves that the dataset contains correct data and that the semantic annotation of data has been carried out in a correct way. Figure 7 summarizes the percentage of students who completed the corresponding test cases compared to the percentage that did not.

It should be noted that the dataset, which contains the semantic annotated data by means of MAOn ontology, is correct from two different perspectives: first, the semantic terminology is correct; second, data included into the dataset have been semantically annotated in a correct way. Therefore, it can be concluded that the ontology developed using the proposed methodology covers the information domain and is sufficient to annotate real data of public transport and its corresponding accessibility information from Metro de Madrid.

Step 7.e. Software Applications

Three different Android applications have been developed to offer different services to the Metro de Madrid users. These apps work with the generated dataset, Dataset_v2, and are part of a complete architecture, which enables both communication and requesting and sending information to and from the semantic repository. This architecture supports this work and other previous work. In addition, the objective of the ontology and competency questions are taken as a reference. As an output, software applications are obtained that make use of the data annotated by the ontology. The applications developed are described below.

(a) The first of the developments, called Metro Madrid Accessible, is capable of calculating Metro de Madrid routes, taking into account the special needs of users and their preferences for that route. To this end, the app offers the possibility of indicating the origin and destination of the desired route. It also offers the ability to select what special needs must be met during this trip: mobility restriction (for example, movement with a baby carriage, wheelchair, etc.), and/or visual and/or auditory restriction. And as regards the route preferences, criteria can be included for the route to be calculated in relation to the lowest possible number of transfers or the lowest number of stops. In addition, the app will allow users to visualize the route obtained from a map and to navigate it, and to consult information on stops and the location of the user. Figure 8 shows an example of the user interface and a route obtained.

(b) The next app, named MMA4A, is an improved and full version of the previous version, presented in [4]. The aim of this app is to offer the user two different functionalities. The first is to calculate an accessible route and show it on a map, as in the previous app, but in this case the app has expanded the possible needs that a user may have when planning a route (specifically, phobias about different elements of accessibility such as an escalator and/or an lift, situations that
in the previous version were not considered). The second is to record the incidents that the user encounters during his journey in the public transport infrastructure, in relation to the accessibility elements of the network (for example, a lift that has stopped working, an escalator under repair, etc.) and which limit accessibility at that point. Figure 9 shows some screenshots of this app.

The users (who are, in fact, players) can then capture the ACEs, thus achieving stations and lines in the public transport system, although the app also has other functionalities. A player can capture an ACE when s/he is close to it and can confirm its working state. Figure 10 depicts the screen employed to capture ACEs. (The player’s current position is a red pin, the ACEs already captured are in grey, while the ACEs that have not been captured are green; the scores (points, levels and badges) are shown at the top of screen).

Figure 8. Screenshots of the app Metro Madrid Appcensible

Figure 9. MMA4A screenshots

(c) The last app developed is based on gamification and crowdsourcing strategies, which is denominated as Access ‘n’ Go! [5]. This app obtains the public transport ACcessibility Elements (ACEs) from Dataset_v2 and represents them on a map. ACEs are the entrances to a station (lifts, escalators, stairs or ramps) and have geographical coordinates (latitude and longitude), thus enabling their representation on the map.

Figure 10. Capture an ACE! Screen of Access ‘n’ Go!

4.2. Ontology evaluation

Another way of evaluating the methodology proposed in this work is through the use of a framework based on generic characteristics that every ontology must possess, for which it has been followed as what was proposed in [16]. In this way, the ontology obtained through the case study is evaluated in a general way. The characteristics used will be the following:

– Correctness: this feature is intended to ensure that the developed ontology does not contain ambiguities or inconsistencies. In our case, as already mentioned in the specification of step 7.a., in order to verify this aspect, the Protégé tool has been used, which makes use of a reasoner that checks the ontology, in addition to its integration and the SWRL rules created. Therefore, the output produced by this tool will ensure that the ontology is correct.

– Feasibility: this characteristic is intended to ensure that the developed ontology gathers all the knowledge about the domain of interest. In this
case, the domain of interest is centred on the Metro de Madrid, where the elements associated with transport are shown, as well as the associated accessibility information. As an example, the ontology collects information through the terms station, stop and entrance/exit associated with the Metro de Madrid infrastructure. On the other hand, it provides information about accessibility such as that given by the wheelchairAccess property, which establishes whether a station, stop or entrance/exit is accessible by a passenger using a wheelchair. In addition, the fact of being able to create annotated data using the ontology, as seen in step 7.e. of the case study, provides evidence that the ontology is capable of collecting the domain of interest.

- **Usefulness**: this characteristic must ensure that the ontology obtained is useful for its potential users or consumers. In our case, the ontology developed is intended to be processed both by machines and by people in order to obtain information about the Metro de Madrid and its associated accessibility. This information may be processed by applications developed ad hoc in order to obtain accessible routes that help Metro de Madrid passengers on their journeys. This feature is well proven by the different apps presented in step 7.e., which verify the real usefulness of the ontology developed through the information and calculation functionality of accessible routes provided by the following.

Other aspects taken into account for this evaluation are described in [1] and [25], among which are: clarity, conciseness, precision and adaptability. There are more aspects that could be assessed, but for this specific case, the aspects considered are the most relevant, as they are identifiable in the process described in this study. They are described in more detail below:

- **Clarity**: the established domain over which the ontology to be developed will deal needs to be adequately and objectively expressed. It can be said that clarity measures the effectiveness with which the ontology presents the desired meaning of the defined terms. To make this aspect possible, the ontology developed is based on the TRANS-MODEL and IFOPT standards that guarantee adequate information. On this basis, a glossary of terms is presented, which is described through step 4.a., which presents a list of definitions of terms of the domain to be represented. With this, a method of representing the domain in a simple and direct language is obtained through definitions that provide clarity to its readers.

- **Conciseness**: the developed ontology acquires this characteristic if it omits unnecessary or irrelevant information, that is to say, that it does not incorporate definitions or terms that do not contribute any value to the domain of study. In our case, the origin of the terms or definitions comes from standards, as already mentioned, so that everything defined by the ontology is related to the treated domain. Furthermore, since ontology is directed by data, this characteristic becomes more credible, implying that the terms to be taken into account about the domain will not be mere descriptive literature, but will be limited to the essential, thus producing a direct ontology with no terms other than those strictly necessary given by the real data.

- **Precision**: this aspect establishes whether the semantic terms provided by the ontology comply with the information expected by the consumers of the ontology. As extracted from step 1.b. of the case study, the objective of the ontology to be developed is twofold. First, the aim is to unify and enrich Metro de Madrid information regarding accessibility. In addition, it will allow the calculation of accessible routes for Metro de Madrid passengers. This feature is achieved through step 7.e., where real developments are obtained in the form of Apps that make it possible to obtain this information and calculate routes.

- **Adaptability**: The intention of this characteristic is to know or measure to what extent the ontology’s uses can be anticipated. Therefore, it is a question of knowing whether the ontology allows possible extensions, specializations or refinements. In short, this aspect implies that the ontology includes the capacity to enlarge the vocabulary without altering what is already pre-established. The methodology presented in this work has an important characteristic, which is that it allows the creation of living ontologies, that is to say, that they are open to the incorporation of new terms, all through an iterative and incremental process. Therefore, this characteristic is fulfilled if the created ontology can be considered living.

4.3. Evaluation of the methodology by experts

Finally, a way is launched to evaluate the methodology through the contributions that can offer the experts in the domain and experts in the development of
ontologies. Some works propose to use a questionnaire as an evaluation and information-gathering method [7]. Following the same method, a questionnaire has been designed and it is composed of: (i) questions with which evaluate the conformity degree based on a five-point scale; (ii) open-ended questions to extract users’ opinions.

This questionnaire aims to evaluate aspects related to the specification proposed in the methodology, in terms of practicality and efficiency, ease of comprehension, correction, completeness and usefulness. Based on these criteria, the following questions were listed:

1. Is the specification of each of the steps that make up the methodology clear?
2. Do you consider that the steps presented by the methodology are balanced and necessary?
3. Overall, is the process described by the methodology complete?
4. Do you think that any of the steps presented should be removed from the methodology?
5. Do you think it is necessary to include some additional task to the methodology?
6. Do you think that the terms ‘live and data-driven’ are the most prominent features of this methodology?
7. Can you evaluate whether you would use the methodology for the construction of ontologies for future projects in which you participate?
8. What weaknesses would you highlight in the methodology presented?
9. What strengths would you highlight in the methodology presented?

5. Conclusions and future work

The methodology proposed in this paper, which is called MAOn, describes a process defined in three phases: conceptualization, construction, and testing. These three phases are further subdivided into seven steps that cover the entire process for domain-based ontology creation. MAOn takes the UPON and SABiO methodologies as a reference for the definition of the proposed method, but introduces three new features that are important to highlight. The first of these is that it is a data-driven methodology. This characteristic provides adherence to reality, since it promotes the obtaining of ontologies that use semantic terms that correspond to real-world objects. The second is by the development of living ontologies, which, through an iterative and incremental process, boost the enrichment or extensibility of the ontology developed following what was established by MAOn, maintaining everything previously represented. The third and last main new feature is a complete testing phase that validates the ontology from different perspectives.

This intension of this proposal is to facilitate the creation of ontologies in a systematic way that provides complete products that adequately cover the domain of interest. Through the ontologies obtained, it is possible to enrich the Web with data that provide associated semantics that promote Web searches and the availability of higher quality information. More specifically, it provides more complete and adequate information that is useful for a greater number of consumers of the same. In this way, the developed ontologies will be valid for the domain of interest, taking into account reusability, as it promotes the use of those well-known ontologies published on the Web with the aim of reusing the knowledge of the Web.

The domain of interest used to validate this work is that of public transport and the accessibility of such transport. As has been shown in the section on related work, transport operators offer complete information on lines, stops, connections, etc., but do not always incorporate information on accessibility and, if they do show it, it is not sufficiently complete for users with special needs. From the review of the existing literature, it can be concluded that, there are proposals that provide open linked public transport data and, on the other hand, there are proposals that provide open accessibility data, but there are no proposals that jointly provide open public transport and accessibility data, which prevents the provision of services to users, such as the calculation of accessible routes. Therefore, in this work, the validation aims to confirm, among other questions, if the set of open and linked data, generated from the ontology, is valid to develop services for all in the public transport domain and its accessibility.

In summary, the proposed methodology aims to be a formal method adequately described through a series of steps for the development of ontologies. These steps, based on information about the domain, produce a series of items that are actualized in inputs and outputs that provide feedback in the process. The product resulting from carrying out each of the steps will be an ontology based on the domain that is correct and complete according to the established purpose.

With regard to future work, the proposal is to use the methodology to generate more case studies applied to different modes of public transport and their associated accessibility in different cities around the world. The aim of these case studies is to obtain a greater
overview of the methodology in the form of practical cases that validate the correction and completeness of the ontologies obtained by the methodology. On the other hand, as a future line of study, the aim is to make use of data transformation languages in different formats that can be translated into RDF making use of existing vocabularies without establishing strong dependencies.

Acknowledgements

This work is supported by the Multiply@City and Access@City projects (TIN2016-78103-C2-1-R), funded by the Spanish Ministry of Science, Innovation and Universities.

References