ATONTE: towards a new methodology for seed ontology development from texts and experts

Helen Mair Rawsthorne[1,*], Nathalie Abadie[1], Eric Kergosien[2], Cécile Duchêne[3] and Éric Saux[4]

[1]LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-94165 Saint-Mandé, France
[2]GERiiCO, Université de Lille, F-59000 Villeneuve d'Ascq, France
[3]LASTIG, Univ Gustave Eiffel, IGN-ENSG, F-77420 Champs-sur-Marne, France
[4]IRENav, École navale, Lanvéoc-Poulmic, CC 600, F-29240 Brest Cedex 9, France

Abstract

ATONTE (ATlantis methodology for ONtology development from Texts and Experts) is a methodology for the manual development of low-level seed ontologies. It uses a combination of knowledge from non-fiction text corpora such as manuals, information guides or sets of instructions, and the knowledge of domain experts. Seed ontologies created with ATONTE can be used to develop and populate knowledge graphs for use in specific applications within given technical domains.

Keywords

application ontology, low-level ontology development, ontology development methodology, seed ontology, technical domain

1. Introduction

ATONTE (ATlantis methodology for ONtology development from Texts and Experts) is a methodology for the manual development of low-level seed ontologies based on non-fiction text corpora such as manuals, information guides or sets of instructions, as well as the knowledge of domain experts. Seed ontologies created with ATONTE can be used as to develop and populate knowledge graphs for use in specific applications within given technical domains.

Current methodologies for ontology construction from text primarily offer automatic and semi-automatic approaches [1]. They employ techniques in machine learning or natural language processing to perform statistical or linguistic analyses on the text. However, automatised approaches pose several problems. Firstly, they assume that the corpus contains precisely the necessary knowledge to satisfy the application without surrounding noise. Secondly, the type of knowledge found in technical corpora can be complex, which is difficult to model automatically. Thirdly, using an automatic approach leaves the door open to unexpected errors that could put the integrity of the content at risk, and errors in the modelling of texts that include information

EKAW'22: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, September 26–29, 2022, Bozen-Bolzano, IT
*Corresponding author.
helen.rawsthorne@ign.fr (H. M. Rawsthorne)
0000-0002-6540-8547 (H. M. Rawsthorne); 0000-0001-8741-2398 (N. Abadie)

relating to safety and security could put human lives at risk. On top of these three issues, automatised approaches do not allow for the easy integration of domain experts' knowledge. Such knowledge can be obtained by conducting targeted interviews with domain experts, and the knowledge they provide can supplement and clarify knowledge extracted from textual sources. A stable and dependable manually-created seed ontology could nevertheless still be enriched semi-automatically or automatically.

We thus present ATONTE as a reverse methodology that integrates domain experts' knowledge and manual modelling techniques during the early stages, and that depends on automation only towards the end in order to complete, instantiate and verify the ontology. To develop ATONTE, we took inspiration from SAMOD (Simplified Agile Methodology for Ontology Development) [2], MOMo (Modular Ontology Modeling) [3] and NeOn (Networked Ontologies) [4].

ATONTE is still under development and is being refined empirically. We are currently testing the methodology via the creation of the ATLANTIS (coAsTaL mAritime NavigaTion InstructionS) ontology [5]. ATLANTIS models the knowledge contained in the *Instructions nautiques*, a series of books published by the Shom (the French Naval Hydrographic and Oceanographic Service) that contain information on navigating in coastal waters.

2. Proposed methodology

In this section we present the five key steps of ATONTE, which are illustrated in Figure 1. Some elements from SAMOD are reused in steps 3 to 5, concepts from MOMo are incorporated into step 4, and elements from NeOn are integrated into step 5.

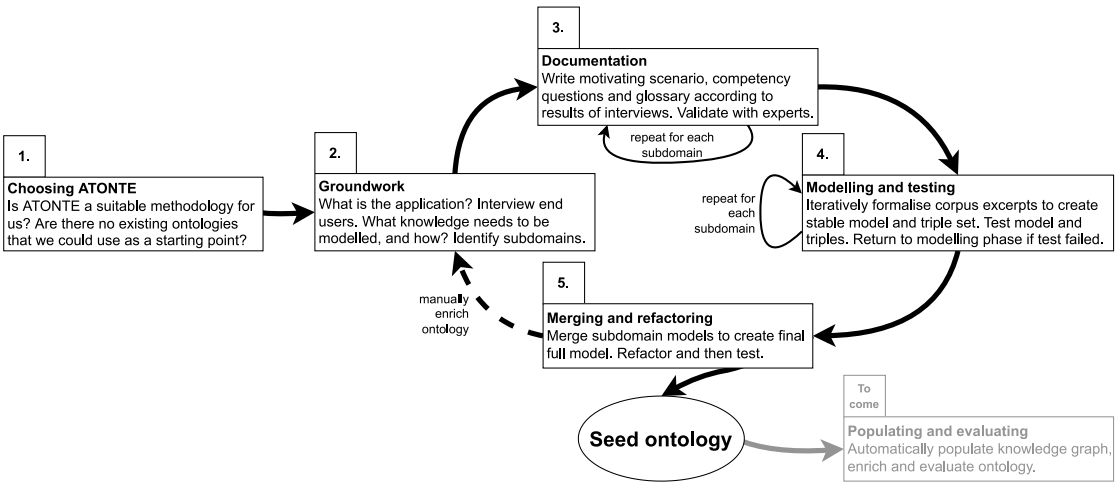

Figure 1: A flowchart showing the steps of the ATONTE process to produce a seed ontology.

2.1. Choosing ATONTE

Each of the three following requirements should be satisfied for it to be a good choice of ontology development methodology for a given task.

1. The aim of the ontology is to model some or all of the knowledge contained in a non-fiction text corpus, and combine it with the knowledge of domain experts, with a given application in mind.
2. The end users of the ontology application are known and can be consulted, along with domain experts (can be the same people), on an ad-hoc basis during the development process.
3. No other ontologies that could be used for the final application, or part of it, already exist and have been published on the Web.

2.2. Groundwork

Define the ontology application. Interview end users of the application to find out what knowledge from the corpus needs to be modelled in the ontology, what knowledge needs to be added (if any) and how it all needs to be structured in order to be useful to them. Unless the knowledge spans only a very small domain, divide it into coherent subdomains. This will facilitate the documentation and modelling phases.

2.3. Documentation

For each subdomain, write a motivating scenario, a list of competency questions and a glossary. Use the results of the interviews to guide the writing of this documentation.

In the motivating scenario, include:

- a name for the subdomain,
- an explanation of the motivation behind modelling the subdomain for the application in question and with the end users' needs in mind,
- a list of the principal characteristics of the concepts within the subdomain found in the corpus, and
- a set of excerpts from the corpus that demonstrate all the ways in which the subdomain knowledge is represented in the text.

Next, draw up a list of natural language competency questions that encapsulate all the uses for the knowledge mentioned by the end users during the interviews. The answers to the questions should figure in the excerpts in the motivating scenario. Finally, compile a glossary defining all the technical terms used in the subdomain. Enrich and validate the documentation with the help of domain experts.

2.4. Modelling and testing

For each subdomain, take the excerpts from the motivating scenario one at a time and semi-formalise the knowledge contained within them that is relevant to the subdomain. Do this by breaking down the relevant knowledge in the excerpt into finer-grained chunks, favouring a subject-predicate-object structure where possible. Once all the excerpts have been semi-formalised in this way, group together the subjects/objects and the predicates that serve the same purpose to create the first set of classes and properties for the subdomain: this is the first iteration of the subdomain model. Rewrite the semi-formal expressions formally as triples. This is the first iteration of the set of triples. Figure 2 shows an example of this process of semi-formalisation and then formalisation of two excerpts of text. Now search for another set of subdomain-relevant excerpts in the corpus and try to formalise their content by creating triples using the newly-created classes and properties. If this task cannot be performed satisfactorily, modify the model accordingly whilst ensuring that it still fits all the other excerpts. Continue in an iterative fashion with unseen sets of excerpts until the model has stabilised and you have a solid set of triples specific to the subdomain.

Submit each subdomain model to a series of three tests: a model test, a data test and a query test, in that order. For the model test, use a reasoner to verify the consistency of the model and then manually check that the model corresponds to the motivating scenario written for it. For the data test, verify the validity of the model by populating it with the triples created from the corpus. For the query test, translate the natural language competency questions into SPARQL queries and run them on the manually-created set of triples for that subdomain to check that the results match the answers specified in the documentation. Move on to the next test only once the previous test has been passed. If the subdomain model fails a test, return to the modelling phase to fix the issue before testing again.

2.5. Merging and refactoring

Now the subdomain models can be merged to create the final full model. Start with the largest subdomain model, merge the second largest into it and then perform the series of tests on this intermediate model. Repeat this process of merging the next-largest subdomain model into the intermediate model and then testing until all subdomain models have been integrated and the

Subdomain Geographical and Meteorological Features
Excerpt 1 "The current in the bay flows eastward, but the wind is northerly."
Excerpt 2 "The port features a lighthouse."
Semi-formalisation current 1 - is a - current / bay 1 - is a - bay / current 1 - is in - bay 1 / current 1 - flows - eastward / wind 1 - is a - wind / bay 1 - has - wind 1 / wind 1 - is - northerly / port 1 - is a - port / lighthouse 1 - is a - lighthouse / port 1 - features - lighthouse 1
Formalisation id:current_1 :hasType :Current / id:bay_1 :hasType :Bay / id:current_1 :isLocatedIn id:bay_1 / id:current_1 :hasDirection :east / id:wind_1 :hasType :Wind / id:bay_1 :contains id:wind_1 / id:wind_1 :hasDirection :north / id:port_1 :hasType :Port / id:lighthouse_1 :hasType :Lighthosue / id:port_1 :contains id:lighthouse_1

Figure 2: An example of the semi-formalisation and formalisation of two sample text excerpts belonging to a subdomain called "Geographical and Meteorological Features".

final full model has been created and tested. The merging process is done by exporting the .owl files of the models and manually combining them, removing all duplicate classes, properties and individuals in the process.

The refactoring process involves reusing existing knowledge in semantic resources, annotating the model and enriching it using the capabilities of the OWL language, for example to create inferences. The elements of the refactoring process are given in SAMOD [6, p. 11]. A detailed description of how to reuse existing semantic knowledge resources is given in NeOn [7]. Figure 3 shows some examples of this refactoring process. After the refactoring process has been carried out, the model should undergo a final testing cycle.

```
<owl:Class rdf:about="http://data.shom.fr/def/coastal_navigation#Wind">
        <owl:equivalentClass rdf:resource="http://sweetontology.net/phenAtmoWind/Wind"/>
        <owl:equivalentClass rdf:resource="https://bimerr.iot.linkeddata.es/def/weather#Wind"/>
</owl:Class>

<owl:ObjectProperty rdf:about="http://data.shom.fr/def/coastal_navigation#contains">
        <owl:equivalentProperty rdf:resource="http://www.opengis.net/ont/geosparql#sfContains"/>
        <rdfs:subPropertyOf rdf:resource="http://data.shom.fr/def/coastal_navigation#
            hasSpatialRelationshipWith"/>
        <owl:inverseOf rdf:resource="http://data.shom.fr/def/coastal_navigation#isLocatedIn"/>
</owl:ObjectProperty>
```

Figure 3: An example of refactoring part of a sample model: aligning classes and properties with existing semantic resources and exploiting the capabilities of the OWL language.

3. Conclusion and perspectives

The ontology development methodology ATONTE allows the creation of application seed ontologies from non-fiction text corpora and the knowledge of domain experts. One characterising feature of this methodology is the early creation of an exemplar dataset, upon which the modelling of the ontology is based. Another is the use and combination of knowledge from a text corpus with the knowledge of domain experts.

ATONTE is currently being used to create the ATLANTIS ontology, which models the contents of the Shom's *Instructions nautiques*. ATLANTIS will offer new possibilities for the access and the verification of the knowledge contained within the series of *Instructions nautiques*. For example, an online platform giving access to ATLANTIS could allow future navigators to search directly for the specific information they are looking for, by category or by geographic area, instead of having to read the full texts. Domain experts' knowledge has been vital in the construction of the ontology for this particular application. ATLANTIS currently contains 110 classes, 90 object properties, 90 data properties and 2190 axioms in total. A full evaluation of ATLANTIS is underway and will be the subject of a future publication.

The short-term perspectives of this project include improving ATONTE by adding a final ontology evaluation phase. It will entail automatically extracting the relevant information from

the corpus being modelled in order to populate a knowledge graph. The seed ontology can thus be enriched with new concepts and properties from other parts of the corpus that were not explored manually. Only once the ontological model is instantiated is it possible to carry out a rigorous evaluation of the ontology produced with ATONTE.

Acknowledgments

This work is co-financed by the Shom and the IGN and is being carried out at the LASTIG, a research unit at Université Gustave Eiffel.

References

[1] A. Al-Arfaj, A. Al-Salman, Ontology Construction from Text: Challenges and Trends, International Journal of Artificial Intelligence and Expert Systems 6 (2015) 15–26.

[2] S. Peroni, A Simplified Agile Methodology for Ontology Development, in: OWL: Experiences and Directions – Reasoner Evaluation, Bologna, Italy, 2016.

[3] C. Shimizu, K. Hammar, P. Hitzler, Modular Ontology Modeling, Semantic Web (2022) 1–31.

[4] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The NeOn Methodology for Ontology Engineering, in: M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, A. Gangemi (Eds.), Ontology Engineering in a Networked World, Springer, Berlin, Heidelberg, 2012, pp. 9–34.

[5] H. M. Rawsthorne, N. Abadie, E. Kergosien, E. Saux, ATLANTIS : Une ontologie pour représenter les Instructions nautiques, in: Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA 2022), Saint-Étienne, France, 2022, pp. 154–163. URL: https://hal.archives-ouvertes.fr/hal-03695242.

[6] S. Peroni, SAMOD: an agile methodology for the development of ontologies, 2016. URL: http://dx.doi.org/10.6084/M9.FIGSHARE.3189769.V2.

[7] M. d'Aquin, Modularizing Ontologies, in: M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, A. Gangemi (Eds.), Ontology Engineering in a Networked World, Springer, Berlin, Heidelberg, 2012, pp. 213–233.