Creation of geospatial knowledge graphs from heterogeneous sources

Helen Mair Rawsthorne

Université Gustave Eiffel

Creation of Geospatial Knowledge Graphs From Heterogeneous Sources

Création de graphes de connaissances géospatiaux à partir de sources hétérogènes

Thèse de doctorat de l'Université Gustave Eiffel

École doctorale n° 532, Mathématiques et Sciences et Technologies de l'Information et de la Communication (MSTIC)
Unité de recherche : UMR LASTIG
Conseil national des universités : 27 - Informatique
Domaine scientifique : 9 - Sciences et technologies de l'information et de la communication
Spécialité de doctorat : Sciences et Technologies de l'Information Géographique

Thèse présentée et soutenue à l'Université Gustave Eiffel, le 15/01/2024, par :

Helen Mair RAWSTHORNE

Composition du jury

Nathalie HERNANDEZ Professeure, Université de Toulouse - Jean Jaurès	Rapportrice
Ian GREGORY Professeur, Lancaster University	Rapporteur
Thierry JOLIVEAU Professeur, Université Jean Monnet	Président du jury
Antoine ZIMMERMANN Professeur, Mines Saint-Étienne	Examinateur
Nathalie LEIDINGER Ingénieure, Pilote innovation, Shom	Invitée

Encadrement de la thèse

Cécile DUCHÊNE Enseignante-Chercheure HDR, LASTIG, Univ Gustave Eiffel, IGN-ENSG	Directrice de thèse
Éric SAUX Maître de conférences HDR, IRENav, École navale	Co-directeur de thèse
Nathalie ABADIE Chargée de recherche, LASTIG, Univ Gustave Eiffel, IGN-ENSG	Co-encadrante de thèse
Eric KERGOSIEN Maître de conférences, GERiiCO, Université de Lille	Co-encadrant de thèse

Funding and Computing Resources

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

This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013699 made by GENCI.

Acknowledgments

I would like to say a huge thank you to all of the following people who contributed to making the last three and a bit years such a productive and enjoyable experience.

My four supervisors. Nathalie Abadie for her guidance and excellent advice throughout. Cécile Duchêne for her constant support and encouragement. Éric Saux for his valuable insights and for welcoming me to the École navale on multiple occasions. Eric Kergosien for his practical recommendations and constructive comments.

Nathalie Leindinger, my point of contact at the Shom, who made things happen and was unfailingly supportive of my work.

Coralie Monpert, Marc Fromentin, Geoffroy Scrive and team, Henri Kajjaj and everyone else at the Shom who took part in the meetings about my work for their interest and enthusiastic engagement.

All the interviewees from the École navale and the École nationale supérieure maritime who gave their time and offered valuable insights into how they use the *Instructions nautiques*, and also those who set up the interviews.

Léa Lamotte, Romain Ruiz and Mayeul de Loynes for the research they carried out prior to my arrival at LASTIG, which provided a sound foundation for my work.

Jean-Guillaume Benoit and Corentin Kergus for the research they produced during their internship at LASTIG in 2022, which they undertook with curiosity, enthusiasm and high spirits.

Claire-Marie Alla, Maxime Charzat, Théo Hermann, Lucie Jeannest and Paul Miancien who produced a brilliant prototype platform during their group project in 2022 and who showed great professionalism throughout.

The TextMine group who gave me the opportunity to promote awareness of my dataset to a wide audience and encourage its further use.

The fantastig lastig crew. In particular Solenn Tual, Emile Blettery, Grégoire Grzeczkowicz, Florent Geniet, Romain Loiseau and Melvin Hersent for always finding the time to help me with my work and teach me what I needed to know to move forward.

Those colleagues from the IGN who made Tuesday lunchtimes full of badminton fun.

Yanis Marchand, for his tireless encouragement and dependable advice, for believing in me, for always being able to make me smile, and whose determination will always be an inspiration to me. Caru ti, cariad, I love you to pieces.

My mum, my dad and all the rest of my family in the UK and in France for their unconditional love, support and kindness. Cariad mawr i bawb.

Abstract

Some spatial knowledge, current or historical, exists only in the form of text. Examples of such sources of unstructured spatial knowledge include travel guides, historical documents and social media posts. Textual sources contain naturally heterogeneous spatial knowledge: they can be written by different authors, using different vocabulary, from different points of view, they can cover large and diverse geographic areas, and they can contain varied levels of detail. These are some of the reasons why it is difficult to integrate geographic information from textual sources into GIS models, which require highly-structured complete data with direct spatial referencing. The open-world assumption of semantic Web technologies makes knowledge graphs a better solution for modelling and storing geographic information extracted from heterogeneous, incomplete and imperfect natural language text. Structured as a geospatial knowledge graph, what was once ambiguous spatial knowledge can be disambiguated and formally linked to reference geographic resources, thereby enriching it with direct spatial referencing where possible and significantly facilitating its accessibility and reuse.

The objective of this thesis is to develop an operational approach for the creation of knowledge graphs from text and geographic reference data that is adapted to the special case of constructing geospatial knowledge graphs that include both direct and indirect spatial referencing.

We apply our research to a French text corpus, which allows us to empirically identify and validate a functional methodology for creating geospatial knowledge graphs from text. The corpus is composed of the *Instructions nautiques*, a series of books published by the Shom that describe the maritime environment and give coastal navigation instructions.

The main contribution of this thesis is the ATONTE Methodology for the semi-automatic construction and population of knowledge graphs, geospatial or not, from heterogeneous textual sources, expert knowledge and reference data. We present the ATONTE Methodology in detail and demonstrate how we implemented it to construct a geospatial knowledge graph of the content of the *Instructions nautiques*.

The first of the three components that make up ATONTE is a novel

methodology for the manual development of domain ontologies from text and the knowledge of domain experts. We apply this methodology to our corpus, integrating our findings from interviews carried out with expert users of the corpus, to develop the ATLANTIS Ontology: a geospatial seed ontology of the domain of the *Instructions nautiques*.

The second component consists of a baseline approach for automatic nested entity and binary relation extraction from text using a deep neural network. It requires training two existing pretrained deep language models, one for the task of entity extraction and the other for relation extraction, on a domain-specific manually-annotated textual dataset. We implement the approach to extract the spatial entities and relations from our corpus, creating a French-language annotated training dataset in the process. We provide benchmark results for this dataset for three tasks: nested spatial entity extraction, binary spatial relation extraction, and end-to-end spatial entity and relation extraction.

The third and final component is dedicated to automatically structuring the information extracted during the previous stage as a knowledge graph according to the ontology developed during the first stage, and disambiguating the entities via entity linking to a reference resource. We present a proof of concept of this stage, using off-the-shelf tools to first structure the spatial entities and relations extracted from the *Instructions nautiques* according to the ATLANTIS Ontology and then link the entities to their corresponding entries in the BD TOPO®. The result is an operational basis for the geospatial ATLANTIS Knowledge Graph of the *Instructions nautiques*.

Keywords: natural language processing, ontology, deep learning, geospatial data

Contents

List of Publications, Posters and Presentations

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2021). "Création de bases de connaissances topographiques à partir de sources hétérogènes". **Invited poster and short presentation**. 30es Journées de la Recherche de l'IGN-ENSG : L'Information géographique encore +. Online. URL: https://hal.science/hal-03239957 (poster), https://youtu.be/2fL4EsLRSZM?si=b7X6AvHlRJQMMlQr (presentation).

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2022). "Création de bases de connaissances à partir de sources hétérogènes : application aux Instructions nautiques". **Invited full presentation**. Journée de l'information scientifiue et technique du Shom : Les sciences océaniques au service de la navigation maritime. Brest, France. URL: https://hal.science/hal-04400549.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2022). "ATLANTIS : Une ontologie pour représenter les Instructions nautiques". In: *33es Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA 2022)*. **Long article and full presentation**. Ingénierie des Connaissances. Saint-Étienne, France, pp. 154–163. URL: https://hal.science/hal-03771117. Nominated for best paper award and selected to appear in the following publication.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2022). "ATLANTIS : Une ontologie pour représenter les Instructions nautiques". In: *Actes CNIA 2022 : Conférence Nationale d'Intelligence Artificielle*. **Invited article**. Conférence Nationale d'Intelligence Artificielle. Saint-Étienne, France, pp. 155–162. URL: https://hal.science/hal-03777860.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2022). "ATONTE: Towards A New Methodology for Seed Ontology Development from Texts and Experts". In: *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management.* **Poster and short article**. 23rd International Conference on Knowledge Engineering and Knowledge Management. Vol. 3256. Bozen-Bolzano, Italy: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-3256/paper4.pdf.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2022). "ATLANTIS : l'ontologie qui modélise les connaissances contenues dans les Instructions nautiques". **Full presentation**. SemWeb.Pro 2022. Paris, France. URL: https://peertube.semweb.pro/w/8kC5ZC4zii9mLgr5jB5jkF.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2023). "Création d'une base de connaissance topographiques à partir des "Instructions nautiques"". **Invited full presentation**. 32es Journées de la Recherche de l'IGN-ENSG : Jumeaux numériques et Anthropocène. Champs-sur-Marne, France. URL: https://youtu.be/5iGMmasUFVo?si=3VGMU5MZsNtozCSB.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2023). "Automatic Nested Spatial Entity and Spatial Relation Extraction From Text for Knowledge Graph Creation: A Baseline Approach and a Benchmark Dataset". In: *7th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities '23), November 13, 2023, Hamburg, Germany.* **Long article and full presentation**. 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities. GeoHumanities'23. Hamburg, Germany: Association for Computing Machinery, pp. 21-30. DOI: 10.1145/3615887.3627754.

Rawsthorne, Helen Mair, Nathalie Abadie, Adrien Guille, Pascal Cuxac, and Cédric Lopez (2024). "Défi TextMine'24 : Reconnaissance d'entités géographiques dans un corpus des Instructions nautiques". In: *TextMine '24 : Atelier sur la Fouille de Textes.* **Invited article**. Atelier TextMine'24, 24ème conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC'24). Dijon, France, pp. 87–92. URL: https://cnrs.hal.science/hal-04434981.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2024). "Extraction automatique d'entités spatiales imbriquées et de relations spatiales à partir de texte pour la création de graphes de connaissances : Une approche et deux jeux de données". In: *TextMine '24 : Atelier sur la Fouille de Textes*. **Invited article**. Atelier TextMine'24, 24ème conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC'24). Dijon, France, pp. 75–86. URL: https://hal.science/hal-04444358.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2024). "De l'ontologie ATLANTIS vers un graphe de connaissances des *Instructions nautiques*". Invited for submission to: *Revue Ouverte d'Intelligence Artificielle*. **Long article**.

List of Published Resources

ATLANTIS Ontology with the documentation and RDF datasets produced during its development. Released under the Etalab Open License Version 2.0. URL: https://github.com/umrlastig/atlantis-ontology/.

ATLANTIS Dataset for nested spatial entity and binary spatial relation extraction with our benchmark results. Released under the Etalab Open License Version 2.0. URL: https://github.com/umrlastig/atlantis-dataset/.

TextMine'24 Dataset for two-level nested spatial entity extraction. Released under the Etalab Open License Version 2.0. URL: https://www.kaggle.com/competitions/defi-textmine-2024.

Scripts produced for the structuring of information using SPARQL-Generate and the disambiguation of spatial entities using Elasticsearch. Released under the Etalab Open License Version 2.0. URL: https://github.com/umrlastig/atonte-structure-and-disambiguate/.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Some spatial knowledge, current or historical, exists only in the form of text. Examples of such sources of unstructured spatial knowledge include travel guides, historical documents and social media posts. These sources can hold information about individual spatial entities that is absent from reference geographic resources [1] such as alternative names in the same or in different languages (Jiménez–Badillo et al. 2020; Beall 2010), and can even mention spatial entities that are missing entirely from reference geographic resources despite being present in the local culture or being part of shared community knowledge (Berragan et al. 2023; Moncla 2015). They often rely on indirect spatial referencing via place names, hierarchical systems [2] and relative locations, and lack direct spatial referencing in the form of geographic coordinates or geometries (Bucher et al. 2019; Keller et al. 2018; Southall et al. 2011; Sallaberry et al. 2007). Their real geographic locations can therefore be vague or even unknown (Elliott and Gillies 2011; Hart and Dolbear 2007). Such texts can also harbour spatial knowledge about the environment at a larger scale that does not exist elsewhere, for example how it is perceived, how it behaves and how it can be navigated (Y. Hu et al. 2019; Kim et al. 2015). Text-based sources contain naturally heterogeneous spatial knowledge: they can be written by different authors, using different vocabulary, from different points of view, they can cover large and diverse geographic areas, and crucially they can contain varied levels of detail (Feliachi et al. 2014; Kuhn 2005; Rodríguez and Egenhofer 2004).

For all these reasons, the knowledge contained within natural language

1. DBpedia (https://www.dbpedia.org/) is a frequently-used global reference geographic resource. The BD TOPO® (https://geoservices.ign.fr/bdtopo) is a reference geographic resource for the French territory.

2. Referencing a given town only as being in a given county is an example use of a hierarchical system. It is common for historical places to be located only hierarchically (Southall et al. 2011).

texts is difficult to access and exploit without first being structured. However, it is difficult to integrate geographic information from text-based sources into geographic information system (GIS) models, which require highly-structured complete data with direct spatial referencing (Keller et al. 2018; Elliott and Gillies 2011). The open-world assumption of semantic Web technologies makes knowledge graphs a better solution for modelling and storing geographic information extracted from heterogeneous, incomplete and imperfect natural language text, and thus making it accessible and reusable (Janowicz et al. 2022; H. Chen et al. 2018; Melo and Martins 2017; Kuhn et al. 2014; Stadler et al. 2012). Structured as a geospatial knowledge graph, what was once ambiguous spatial knowledge can be disambiguated and formally linked to reference geographic resources, thereby enriching it with direct spatial referencing where possible (Keller et al. 2018).

Harnessing the spatial knowledge contained within text by extracting it and structuring it as a geospatial knowledge graph opens up a vast range of possibilities. Structured geographic information can be queried or processed in order to provide access to it in other forms and it can be enhanced by linking it to other sources of information (Janowicz et al. 2022; Melo and Martins 2017). It also makes it possible to verify its coherence and infer new facts thanks to reasoning (Hogan et al. 2021; Paulheim 2017; F. M. Suchanek 2014). The process of structuring the spatial knowledge contained within text can also lead to the creation of reference geographic resources, the enrichment of existing ones, error detection and the identification of necessary updates. By applying linked data standards, such resources can be published to conform with the Findable, Accessible, Interoperable, Reusable (FAIR) principles to ensure their reusability by machines and humans (Wilkinson et al. 2016).

1.2 Objective and Challenges

The objective of this thesis is to provide a functional approach to constructing geospatial knowledge graphs from heterogeneous text-based sources and geographic reference data that is suited to applications that require performing direct and indirect spatial reasoning. To achieve this objective, three main scientific challenges must be overcome:

1. How can we acquire a domain ontology suited to the text corpus to serve as a structure for the geospatial knowledge graph?
2. What techniques for spatial information extraction can we apply to heterogeneous text-based sources and do they need to be refined for

this use case?

3. How can we automatically structure the extracted spatial information according to the domain ontology to populate it and therefore construct the geospatial knowledge graph, and how can we disambiguate the spatial entities and link them to a reference geographic resource?

1.3 Application Context

We apply our research to a text corpus as part of a case study in collaboration with the *Service hydrographique et océanographique de la Marine* (Shom) [3], the French Naval Hydrographic and Oceanographic Service.

The maritime industry is currently being transformed by trend towards digitisation (DNV 2023; Murat 2023). The objective is to tend towards intelligent GIS for maritime products and navigation. In particular, the S-100 Universal Hydrographic Data Model developed by the International Hydrographic Organization (IHO) [4] is a new standard for digital products and services for hydrographic, maritime and GIS communities. Its implementation period being between 2020 and 2030 shows that this industry is becoming more aware of the value of structured and linked data (Contarinis et al. 2020). With this in mind, the Shom is looking to modernise its nautical publications.

One such publication is its series of *Instructions nautiques*. Available to buy as PDF publications, each volume contains essential information for navigating safely in the coastal waters of a specific geographic area, including descriptions of the coastal maritime environment and instructions for entering ports. The geographic areas covered by the *Instructions nautiques* range all over the world. Written in French, the *Instructions nautiques* have several categories of users including francophone military, civilian, professional and amateur navigators. During itinerary planning navigators consult the *Instructions nautiques* alongside nautical charts, which cannot display all the information required by navigators to plot a safe and suitable route for their voyage.

Structuring the content of the *Instructions nautiques* as a geospatial knowledge graph would make it machine-readable and would allow processing the spatial information contained within it. This in turn could offer new possibilities to improve the production chain, maintenance process and user experience of the *Instructions nautiques*. A detailed analysis of the application context is given in chapter 2.

3. https://www.shom.fr/
4. https://iho.int/

1.4 Global Approach

The global approach that we adopt to fulfil the objective described in section 1.2 consists of the practical application of the three challenges to a corpus of the *Instructions nautiques*. This allows us to empirically identify and validate a functional methodology for constructing geospatial knowledge graphs from text. Although we use a corpus based on the maritime domain, our methodology is equally applicable to the terrestrial domain.

The first challenge is to acquire an ontology of the domain of the *Instructions nautiques*. With existing ontologies unable to represent most of the technical vocabulary and the navigation instructions that make up a large part of the *Instructions nautiques*, we develop a dedicated seed ontology of the domain. We choose to develop the ontology manually instead of using an ontology learning approach given the delicate nature of the corpus and the need to integrate knowledge contributed by domain experts. Integrating experts' knowledge facilitates the correct understanding of a vast and complex domain and helps to ensure that the final knowledge graph will fit its multiple purposes: improving the production chain, the maintenance process and the user experience of the current *Instructions nautiques*. In the absence of a suitable methodology to develop ontologies to represent knowledge originating from textual sources and from experts, we create and implement a new methodology that reuses elements from Simple Agile Methodology for Ontology Development (SAMOD) (Peroni 2016a), Modular Ontology Modeling (MOMo) (Shimizu et al. 2022) and Networked Ontologies (NeOn) (Suárez-Figueroa et al. 2012).

The second challenge is to extract spatial information from the *Instructions nautiques* according to the ontology. The fundamental elements of spatial knowledge being spatial entities, their types and the spatial relations between them, these are the elements on which we focus during the extraction process. We present a baseline supervised deep learning approach for automatic nested [5] spatial entity and binary spatial relation extraction from text. Our approach involves applying the Princeton University Relation Extraction system (PURE) (Zhong and D. Chen 2021), made for *flat*, *generic* entity extraction and *generic* binary relation extraction, to the extraction of *nested*, *spatial* entities and *spatial* binary relations. The advantage of extracting *nested* spatial entities and the *spatial* relations between them is that it captures more information that can aid entity disambiguation. We carry out experiments to compare the per-

5. *Flat* entity extraction involves the identification of the word or set of words that refer to an entity. *Nested* entity extraction also involves the identification of the word or set of words that refer to an entity, but additionally aims to identify finer-grained information about the entity name within the set of words. A more detailed explanation of these concepts is given in section 5.1 of chapter 5, on page 107 onwards.

formance of a pretrained monolingual French BERT language model with that of a pretrained multilingual BERT language model for these tasks, and study the effect of including cross-sentence context.

The third challenge is to structure the extracted spatial information according to the ontology and then disambiguate the spatial entities by linking them to their corresponding entries in reference geographic resources. The result is a geospatial knowledge graph. We present a proof of concept that uses SPARQL-Generate [6] (Lefrançois et al. 2017) to structure the nested spatial entities and relations extracted from text using our extraction approach as Resource Description Framework (RDF) triples. Then, we build upon the work of Loynes and Ruiz (2020) to disambiguate the named spatial entities by linking them to their corresponding entries in the BD TOPO® using algorithms based on Elasticsearch [7] queries. We improve the overall accuracy of the algorithms by taking advantage of the extra information available thanks to the nested extraction of spatial entities.

Based on our empirical experiences, we formalise our global approach as the reproducible three-step ATlantis Ontology and kNowledge graph development from Texts and Experts (ATONTE) Methodology. ATONTE is not only suited to the construction of geospatial knowledge graphs from text containing spatial knowledge: we present it as a generic methodology that can be used for the creation of knowledge graphs from texts on any domain. A workflow diagram that illustrates the ATONTE Methodology is presented in figure 1.4.1.

1.5 Outline of the Manuscript

In the next chapter we introduce the case study to which we apply our research during the remainder of the thesis: the *Instructions nautiques*. We discuss how they are produced and by whom, their uses and users, and how our work can benefit their future evolution. In chapter 3 we first present related work in knowledge graph creation from text. Then we give an overview of the ATONTE Methodology for constructing (geospatial) knowledge graphs from heterogeneous textual sources and experts. The three following chapters are dedicated to the three main components of the ATONTE Methodology, each one starting with related work followed by the implementation of the component within our case study. In chapter 4 we present our methodology for the manual development of (geospatial) domain ontologies from text and experts (Rawsthorne et al. 2022b),

6. http://w3id.org/sparql-generate
7. https://www.elastic.co/

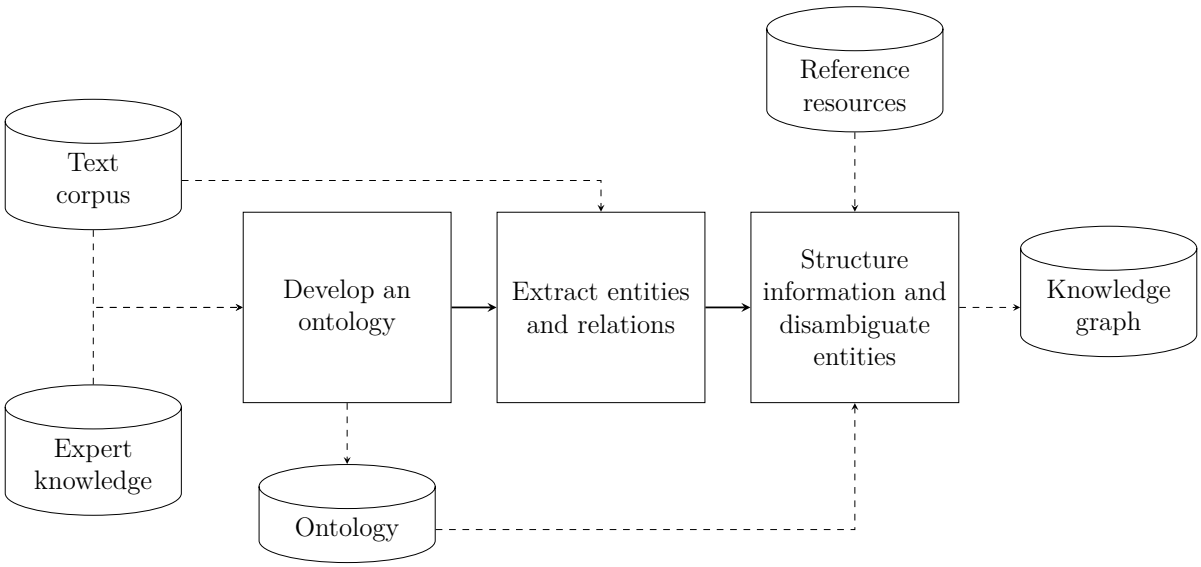


Figure 1.4.1 – Diagram of the global approach adopted in this thesis. The main processes are represented by squares whilst the input and output resources are represented by cylinders. The solid arrows represent the task workflow and the dashed lines represent flows of information or knowledge.

and show how we applied it to create the coAsTaL mAritime NavigaTion InstructionS (ATLANTIS) Ontology [8], a geospatial seed ontology of the domain of the *Instructions nautiques* (Rawsthorne et al. 2022a). In chapter 5 we present our baseline approach for nested (spatial) entity and binary (spatial) relation extraction from text, and show how we applied it to the spatial entities and relations contained within the text of the *Instructions nautiques* (Rawsthorne et al. 2023). In chapter 6 we discuss the structuring and linking of the extracted (spatial) entities as a (geospatial) knowledge graph, and show how we applied it to the spatial entities and relations extracted from the *Instructions nautiques*. Finally, in chapter 7 we conclude our work on the ATONTE Methodology and its application to our case study and discuss associated future work.

To show examples of the textual content of our corpus of *Instructions nautiques*, we provide extracts of the text in shaded mauve-coloured boxes throughout the thesis. We have translated the extracts into English, and give the original French text in the caption. Where possible, we give definitions for technical maritime vocabulary from the *S-32 IHO Hydrographic Dictionary* (Hydrographic Dictionary Working Group 2019). The most important terms are defined directly in the text whilst definitions for terms of secondary importance are given in footnotes.

8. https://github.com/umrlastig/atlantis-ontology/

Chapter 2

Analysis of Application Context

2.1 Introduction

In this chapter we introduce and examine the corpus to which we apply our research, the *Instructions nautiques*, in detail.

We first present the organisation that produces the *Instructions nautiques* in section 2.2. Then, we discuss the way in which the *Instructions nautiques* are produced, their content, their typical users and the way in which they are used in section 2.3. Finally, in section 2.4 we explain the advantages of applying our research to the *Instructions nautiques*, in particular the possibilities it presents to their producers in terms of production and maintenance, and to their users in terms of access. Written extracts and figures from various volumes of the *Instructions nautiques* are inserted throughout the chapter to illustrate their textual content, writing style and visual content.

2.2 The Shom

The *Service hydrographique et océanographique de la Marine* (Shom) is the French Naval Hydrographic and Oceanographic Service, a public administrative establishment under the supervision of the *ministère des Armées*, the French Ministry of Armed Forces. The Shom is the official representative to the International Hydrographic Organization (IHO) as designated by the French government (International Hydrographic Organization 2023). Its mission is to study and describe the physical marine environment in relation to the atmosphere, the seabed and coastal zones, predict its evolution and distribute relevant information. It produces reference coastal and maritime geographic information. The three main aims of the Shom are:

1. To provide for all national hydrography [1] needs

1. "Hydrography is the branch of applied sciences which deals with the measurement and description

2. To provide defence support
3. To support public maritime and coastal policy

The Shom produces nautical charts and a range of nautical publications [2] that are necessary to ensure safe navigation in maritime spaces under French sovereignty or jurisdiction [3] and in other selected regions (see figure 2.3.1). Nautical charts are charts that have been specifically designed to meet the requirements of maritime navigation. They show values for the depth of the water at various points, the nature of the seabed, the configuration and characteristics of the coast, dangers and aids to navigation [4] (Hydrographic Dictionary Working Group 2019). The Shom produces nautical charts in three different formats: paper charts, raster navigational chart (RNC) and electronic navigational charts (ENC). A RNC is a digital image of a paper chart. An ENC is the subset of the Electronic Chart Data Base (ECDB) that is held on a vessel. It contains information on features that are useful for navigation such as the coastline, obstructions and beacons.

2.3 The *Instructions nautiques* and other Sailing Directions

2.3.1 What are the *Instructions nautiques*?

One of the nautical publications produced and distributed by the Shom is the series of *Instructions nautiques*. The series is composed of 16 PDF volumes of around 100 to 800 pages written in French, each one dedicated to a different region around the world. The *Instructions nautiques* cover coastal zones in Africa, Europe, North and South America, as well as in the Indian and Pacific Oceans. Figure 2.3.1 shows the coverage of each volume of the *Instructions nautiques*. Each volume contains three main types of information (Shom 2020) that are essential to planning a safe and suitable itinerary in the coastal waters of the region in question:

1. Information that is complementary to that which is displayed on nautical charts, such as descriptions of the coastal maritime environment

of the physical features of oceans, seas, coastal areas, lakes and rivers, as well as with the prediction of their change over time, for the primary purpose of safety of navigation and in support of all other marine activities, including economic development, security and defence, scientific research, and environmental protection." (Hydrographic Dictionary Working Group 2019)

2. https://diffusion.shom.fr/

3. https://maritimelimits.gouv.fr/themes/french-maritime-areas

4. An aid to navigation is defined as "A visual, acoustical, or RADIO device, external to a ship, designed to assist in determining a safe COURSE or a vessel's POSITION, or to warn of dangers and/or OBSTRUCTIONS." (Hydrographic Dictionary Working Group 2019).

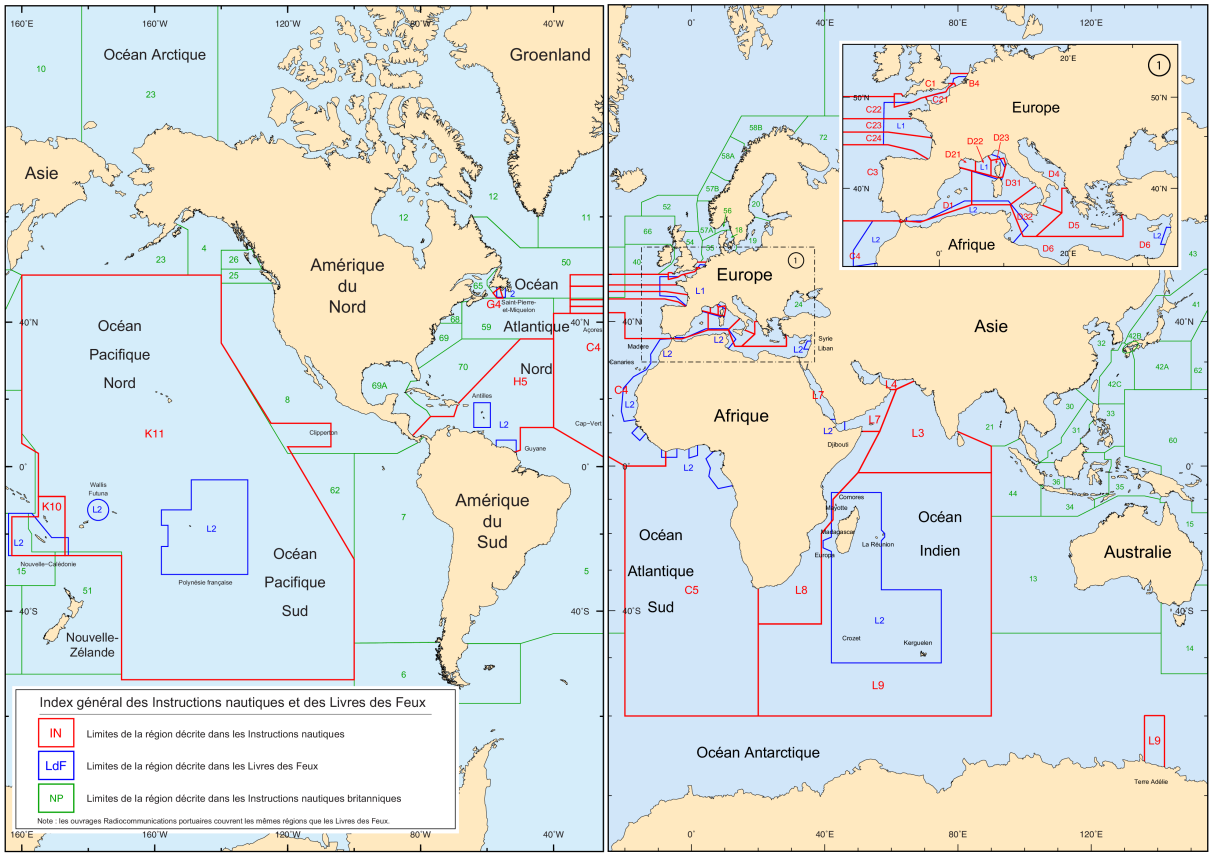


Figure 2.3.1 – Maps showing the boundaries of the regions covered by volumes of the *Instructions nautiques* in red, volumes of the *Livre des Feux* in blue and volumes of the *ADMIRALTY Sailing Directions* in green. These maps can be found at the beginning of every volume of the *Instructions nautiques*.

from the point of view of a vessel on the water, including the physical characteristics of landmarks [5] (colour, shape, size, etc.) and the spatial relations between them, as in extract 2.3.1

2. Information that is absent from nautical charts such as the typical currents and climate of a region, as in extract 2.3.2
3. Navigation instructions, as in extract 2.3.3, and other information about coastal navigation including rules and regulations, recommended routes, port access conditions and dangers, as in extract 2.3.4

The region covered by one volume is defined either as the section of coastline between two points on the coast of a given landmass, or as the entire coastline of an island or a group of islands. Figure 2.3.2a shows the section of coast covered by one volume of the *Instructions nautiques*.

Each volume begins with a chapter of general information that applies to the entire region covered by the volume, including geography, meteorology, hydrography, oceanography and maritime radio services. The remaining chapters follow the coastline in a linear fashion as illustrated in

5. A landmark is defined as "Any PROMINENT OBJECT at a fixed location on LAND which can be used in determining a location or a DIRECTION." (Hydrographic Dictionary Working Group 2019).

"Barn Hill Point (23° 33.3' S — 43° 44.6' E) is the extremity of a narrow craggy peninsula that reaches 1 M SSW of Taliokoaka, a headland 60 m tall. This peninsula, also known as Ny Andrea (Nosy Andrea), is lined with white limestone cliffs that stand out when illuminated by the sun."

Extract 2.3.1 – Translated from the original French text: "La pointe Barn Hill (23° 33,3' S — 43° 44,6' E) est l'extrémité d'une étroite péninsule escarpée qui s'avance à 1 M au SSW de Taliokoaka, promontoire haut de 60 m. Cette péninsule, connue sous le nom de Ny Andrea (Nosy Andrea), est bordée de falaises calcaires de couleur blanche, très apparentes lorsqu'elles sont éclairées par le soleil." (Shom 2021g, p. 309)

"The climate is cold, humid and very windy. On the coastal plains, snow can fall at any time of year but rarely lasts more than a few days."

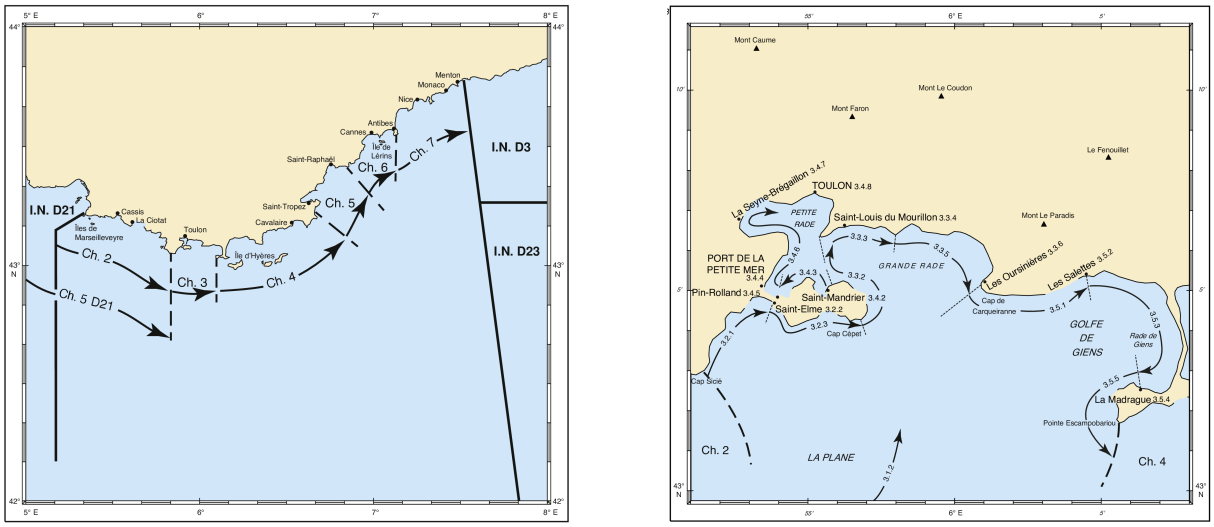
Extract 2.3.2 – Translated from the original French text: "Le climat est froid, humide et très venteux. Sur les plaines côtières, la neige peut tomber à toute époque de l'année mais subsiste rarement plus de quelques jours." (Shom 2021g, p. 458)

"The Great Western Pass (12° 47.90' S — 44° 58.00' E) is unsafe and unmarked. It is not recommended to take this channel."

Extract 2.3.3 – Translated from the original French text: "La grande passe de l'Ouest (12° 47,90' S — 44° 58,00' E) est malsaine et non balisée. Il est déconseillé d'emprunter cette passe." (Shom 2021g, p. 231)

"INSTRUCTIONS. — During the day, the channel entry access route is oriented at approximately 114° towards the southern extremity of the summit of Mont Mahinia, or towards the northern slope of Mont de la Selle (§ 5.5.2.). As soon as the beacons have been identified, follow the leading line (114.5°) indicated on the chart."

Extract 2.3.4 – Translated from the original French text: "INSTRUCTIONS. — De jour, la route d'approche de l'entrée de la passe est orientée à environ 114° vers l'extrémité Sud du sommet du mont Mahinia, ou vers le versant Nord du mont de la Selle (§ 5.5.2.). Dès que les balises sont identifiées, suivre l'alignement (114,5°) indiqué par la carte." (Shom 2021g, p. 309)



(a) Volume D22 **(b)** Chapter 3, volume D22

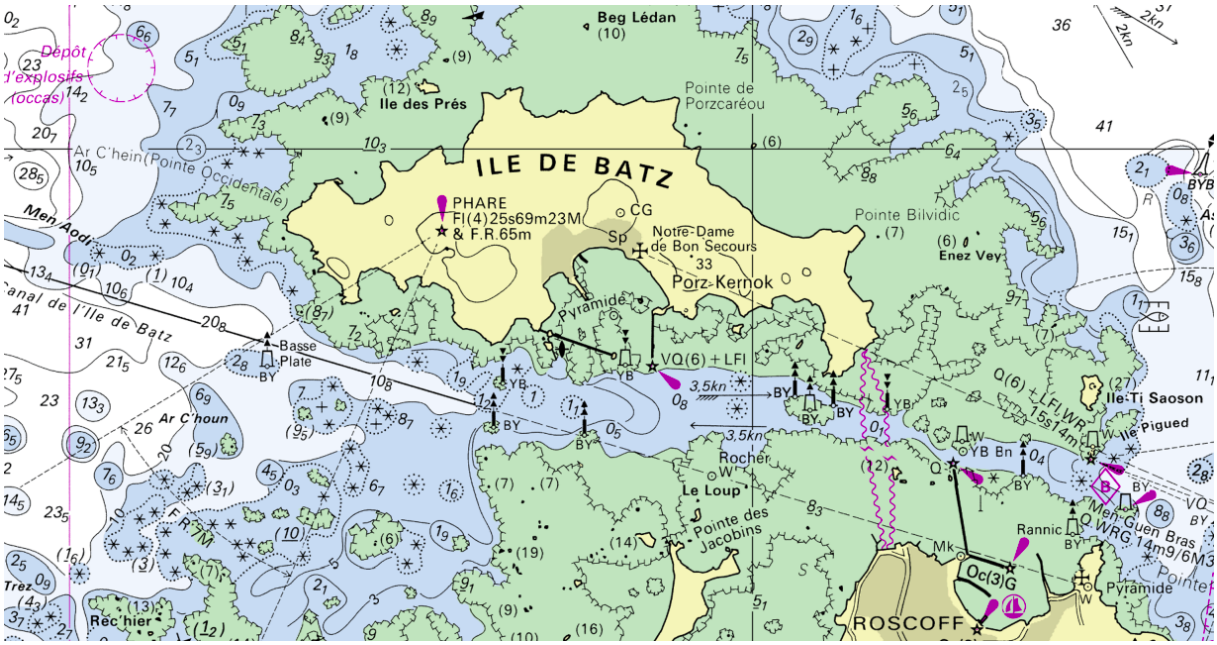
Figure 2.3.2 – Maps showing the boundaries of the regions covered by one volume and one chapter of that volume of the *Instructions nautiques* (Shom 2021c, p. 2 and p. 127).

figure 2.3.2a, each chapter dedicated to one subsection of coast. Within a chapter, descriptions of waterways [6], routes, ports and the surrounding coast are given. Descriptions are written from the point of view of a vessel that is progressively approaching the coast, a port or an anchorage. When reading a chapter we have the impression of being led along the coast by the author: every landmark, danger and other specificity of the environment is described in order, and each anchorage, port and waterway entry is indicated. The local meteorology, currentology and regulations are also presented. Figure 2.3.2b shows the division into subsections of the region covered by one chapter of a volume of the *Instructions nautiques*. Photographs dispersed throughout the text, such as the one in figure 2.3.3b, show important landmarks and ports, and also show the relative position of different spatial entities to help the reader imagine an accurate representation of the coastal environment before being at sea.

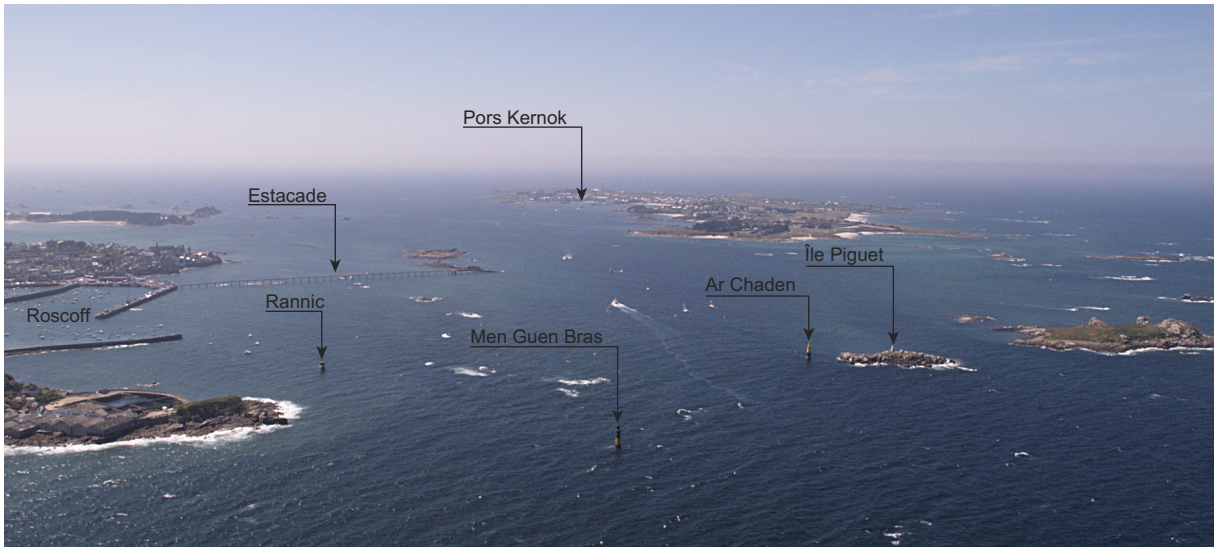
2.3.2 Spatial Definitions

As mentioned in section 1.4, in this thesis we focus on extracting specific elements of spatial information from text: *spatial entities*, their *types*, and the *spatial relations* between them. Here we define these terms and related ones in relation to our corpus of *Instructions nautiques*.

6. A waterway is defined as "A line of water (RIVER, CHANNEL, etc.) which can be utilized for communication or transport." (Hydrographic Dictionary Working Group 2019).



(a) Extract of the *Assemblage des cartes marines (RasterMarine)* RNC product published by the Shom showing the Île de Batz channel (Shom 2023). Yellow areas correspond to land, green areas correspond to intertidal zones (exposed at low tide and underwater at high tide) and blue areas correspond to areas that are permanently underwater.



(b) Annotated photograph of the Île de Batz channel from volume C22 of the *Instructions nautiques* (Shom 2021a, p. 399).

Figure 2.3.3 – The nautical chart in (a) and the annotated photograph in (b) show different representations of the Île de Batz channel.

2.3.2.1 Spatial Entity

Kuhn (2005) notes that natural language expressions are used to express concepts that exist in our minds and refer to *entities* in the real world. According to Casati and Varzi (1997), spatial reasoning involves reasoning about *spatial entities*: things located in space. They distinguish spatial entities from purely spatial items, examples of which are points, lines and regions. Worboys and Hornsby (2004) indicate that the uniqueness of *geospatial entities* lies in the fact that they have a setting, which can be either spatial, temporal or both. A spatial setting can be a point or a region. They also note that the setting of a geospatial entity has an appropriate, macroscopic scale.

We adopt the definition of Casati and Varzi (1997) for spatial entities as "things located in space", acknowledging that they are necessarily associated with a location [7] (International Organization for Standardization 2019) or setting that is macroscopic (Worboys and Hornsby 2004).

We extend this definition by specifying that spatial entities can be physical or virtual [8] whilst still occupying a location in space. The location of a spatial entity can be expressed via a spatial reference [9] (International Organization for Standardization 2019), which can either be direct (see section 2.3.2.2) or indirect (see section 2.3.2.3). When referenced in natural language text, a spatial entity can either be named or unnamed. In extract 2.3.1, 'Barn Hill Point' refers to a named spatial entity whilst 'peninsula' refers to an unnamed spatial entity.

2.3.2.2 Direct Spatial Referencing

Direct spatial referencing is a way of expressing a geographic location using geographic coordinates or geometries (Bunel 2021; Keller et al. 2018). In extract 2.3.1, the geographic location of 'Barn Hill Point' is expressed via direct spatial referencing using geographic coordinates.

2.3.2.3 Indirect Spatial Referencing

Indirect spatial referencing is a way of expressing a geographic location using place names or relative locations, without using geographic coordinates or geometries (Bunel 2021; Bucher et al. 2019). In extract 2.3.1, the geographic location of 'Barn Hill Point' is expressed via indirect spa-

7. https://www.iso.org/obp/ui/en/#iso:std:iso:19112:ed-2:v1:en:term:3.1.3

8. An example of a virtual spatial entity is a *leading line*, which is defined as a "line passing through two or more clearly defined charted objects, and along which a vessel can approach safely" (Hydrographic Dictionary Working Group 2019).

9. https://www.iso.org/obp/ui/en/#iso:std:iso:19112:ed-2:v1:en:term:3.1.4

tial referencing as being 1 M south-south-west of another spatial entity, 'Taliokoaka'.

2.3.2.4 Spatial Entity Type

Mirroring the terms *feature*[10] and *feature type*[11] defined in ISO 19101-1:2014 (International Organization for Standardization 2014), we define the term *spatial entity type* as a class of spatial entities that have common characteristics, such as their physical and functional nature. These types correspond to common nouns (Rodríguez and Egenhofer 2004). Examples of spatial entity types are *beacon*, *channel*, *lighthouse* and *peninsula*. For named spatial entities, the spatial entity type can optionally be contained within the name of the spatial entity. In extract 2.3.3, 'Great Western Pass' refers to a named spatial entity of type *pass*, which we know thanks to the noun being present in its name, whilst 'Taliokoaka' in extract 2.3.1 refers to a named spatial entity of type *headland*, which we know only as the noun is mentioned in the text. Unnamed spatial entities are always referred to by the noun that describes their type, such as 'peninsula' in extract 2.3.1.

2.3.2.5 Spatial Relation

A spatial relation is defined in ISO 24617-7:2020 as a "segment or series of segments of a text that rebounds to qualitative spatial relations or orientational relations, or to movement relations indirectly through the specification of the bounds of paths or event-paths" and can involve topological, orientational or metric values (International Organization for Standardization 2020). We apply this definition of spatial relations to spatial entities. Qualitative relations express connectedness or continuity using topological values. Orientational relations express the spatial disposition or direction of a spatial entity within a frame of reference using orientational values. The magnitude of a spatial relation can be measured using metric values. A spatial relation is an example of indirect spatial referencing. In extract 2.3.2, it is indicated that the thing 'snow' can be present at the location of the unnamed spatial entity 'coastal plains'. The 'snow' therefore has a spatial relation with the 'coastal plains'. In extract 2.3.1, 'Barn Hill Point' is indicated as being part of the 'peninsula', meaning that a spatial relation exists between the two spatial entities. These are both examples of *binary* spatial relations, which involve exactly two things. Higher order spatial relations, known as n-ary relations, involve three or more things.

10. https://www.iso.org/obp/ui/en/#iso:std:iso:19101:-1:ed-1:v1:en:term:4.1.11
11. https://www.iso.org/obp/ui/en/#iso:std:iso:19101:-1:ed-1:v1:en:term:4.1.16

"The port of Sidi Ifni is located between the towns of Tiznit and Guelmim, 160 km south of Agadir."

Extract 2.3.5 – Translated from the original French text: "Le port de Sidi Ifni est établi entre les villes de Tiznit et de Guelmim, à 160 km au sud d'Agadir." (Shom 2021b, p. 349)

For example, in extract 2.3.5 it is indicated that 'Sidi Ifni' is located between two other spatial entities: 'Tiznit' and 'Guelmim'. It is impossible to represent this ternary relation in the form of triples.

2.3.3 Production of the *Instructions nautiques*

Apart from the transition to digital formats, the production method of the *Instructions nautiques* has little changed since the first editions were published in the middle of the 19$^{\text{th}}$ century. The volumes are updated manually by Shom personnel, when changes in the environment are discovered or when new information becomes available. If the updates are urgent, they are published in an online *Groupe d'Avis aux Navigateurs* (GAN) notice that summarises all urgent modifications, additions and deletions that should be applied the Shom's charts and publications every week (Shom 2022). If the updates are non-urgent, they are published in the subsequent versions of the relevant volumes.

The Shom has developed two in-house tools specifically for the writing, maintenance and publication of the *Instructions nautiques* and their associated GAN notices: *Correction des Instructions nautiques* (CorIN), used by the production team for volumes in production and for the corrections published in a GAN notice, and *Édition des Instructions nautiques* (EditIN), used by the team of writers for writing brand new volumes and for volumes or chapters undergoing a full rewrite.

The main steps involved in the maintenance of the *Instructions nautiques* are as follows:

1. When a new piece of information is received or uncovered by the Shom, it is analysed manually to determine:
   a. Its consequences on Shom charts and other publications (*Livres des feux*, *Instructions nautiques*, *Radiosignaux*, *Guide du Navigateur*, *Album des pavillons nationaux et marques distinctives*, etc.)
   b. Its urgency, and therefore how it should be circulated
2. If the piece of information is determined to have a consequence on the *Instructions nautiques*:
   a. If it is determined to be urgent, the *Instructions nautiques* are updated via a GAN notice (see step number 3)

b. If it is determined to be non-urgent, it is transferred to the member of the writing team who is in charge of the relevant volume of the *Instructions nautiques*, the current text is manually modified accordingly and is published in the next version of that volume (see step number 4)

3. If the piece of information is determined to be urgent:
   a. It is analysed and supporting elements are searched for manually (more precise information, relevant regulations, photographs, etc.) to improve the quality of the publication and therefore safety standards
   b. The lines concerned by the update in the relevant volume or volumes of the *Instructions nautiques* are searched for manually (lines or paragraphs or sections or chapters: anything is possible)
   c. A document containing a summary of the actions to be carried out, a list of the references to be cited and the new version of the *Instructions nautiques* text is drafted
   d. The CorIN tool is used to manually correct the *Instructions nautiques* line by line
   e. The corrections are attributed to a GAN notice
   f. The GAN notice is produced by the production team
4. If the piece of information is determined to be non-urgent:
   a. It is analysed and supporting elements are searched for manually (more precise information, relevant regulations, photographs, etc.) to improve the quality of the publication and therefore safety standards
   b. The lines concerned by the update in the relevant volume or volumes of the *Instructions nautiques* are searched for manually (lines or paragraphs or sections or chapters: anything is possible)
   c. If it requires a big change or insertion, a new outline is determined for the corresponding paragraph, section or chapter
   d. The EditIN tool is used to manually write the new lines by modifying, replacing or removing the current text
   e. The new version is transferred to the CorIN tool and passed to the production team

2.3.4 Users and Uses of the *Instructions nautiques*

Military, civilian, professional and amateur navigators use the *Instructions nautiques* during the preparation process of a voyage, before leaving

land. They are used to help identify an itinerary that is suitable for the vessel, the experience of the navigators and the weather, and gather the information needed to navigate it. They are consulted alongside ENC, RNC or paper nautical charts, on which it is impossible to graphically display all the information that is necessary to identify a safe and appropriate route. Figure 2.3.3 shows an extract of a RNC showing the Île de Batz channel and a photograph from a volume of the *Instructions nautiques* that shows the same region. The RNC shows large intertidal zones [12] on either side of the channel and suggests that the part of the channel that is navigable is very narrow with visible dangers. The photograph shows a large channel, wide enough to safely perform manoeuvres without visible dangers. The RNC and the photograph show considerably different representations of the region. Whilst both are accurate, neither is precise enough to plan a safe and efficient itinerary without the other nor without the accompanying written instructions in the *Instructions nautiques*, a part of which are shown in extract 2.3.6. More specialised complementary information can be gathered from other nautical publications produced by the Shom such as the *Livres des feux*[13] which gives light and fog signal information, *Radiosignaux*[14] which gives maritime radio communication information, *Courants de marée*[15] which gives tidal streams and tidal heights, and *Annuaire des marées*[16] which gives tidal predictions. A fuller description of how the *Instructions nautiques* are used is given in section 4.4.3.1, which summarises the results of the interviews that we carried out with users of the *Instructions nautiques*.

2.3.5 Sailing Directions around the World

The *Instructions nautiques* cover maritime regions all around the world but are only published in French, limiting their potential users to mariners who have a good understanding of the French language. Many other national hydrographic services produce their own versions of the *Instructions nautiques*, which are commonly known as Sailing Directions. They are generally organised in a similar way to the *Instructions nautiques* and contain the same types of information, but are written in English and/or the main language or languages of the nation.

12. An intertidal zone is defined as "The zone generally considered to be between MEAN HIGH WATER and MEAN LOW WATER levels." (Hydrographic Dictionary Working Group 2019). In other words, an intertidal zone is exposed at low tide and underwater at high tide.
13. https://diffusion.shom.fr/ouvrages/livres-des-feux.html
14. https://diffusion.shom.fr/ouvrages/radiosignaux.html
15. https://diffusion.shom.fr/marees/courants-de-maree.html
16. https://diffusion.shom.fr/marees/annuaires-de-marees.html

"INSTRUCTIONS. — When coming from the east, the channel is approached with Île de Batz bell tower (Notre-Dame de Bon Secours chapel) [48° 44.65' N — 4° 00.58' W] and the white pyramid of Île Pigued (48° 43.98' N — 3° 58.22' W) in range bearing 293.3°. This alignment is visible to small vessels up to around 0.6 M to the east of 'Le Menk' turret (at half tide) and, for vessels with higher bridges, up to the north of the turret. This alignment is situated in the white sector (289.5° – 293°) of 'Ar Chaden' turret light. The route at 293.3° passes south of the plateau des Duons and north of 'Le Menk' turret (48° 43.29' N — 3° 56.70' W), lighted west cardinal, and the Basse de Bloscon."

Extract 2.3.6 – Translated from the original French text: "INSTRUCTIONS. — En venant de l'Est, on prend le chenal en suivant l'alignement à 293,3° du clocher de l'Île de Batz (chapelle Notre-Dame de Bon Secours) [48° 44,65' N — 4° 00,58' W], sur la côte Sud de l'île, par la pyramide blanche de l'Île Pigued (48° 43,98' N — 3° 58,22' W). Cet alignement n'est visible par les petits navires que jusqu'à environ 0,6 M à l'Est de la tourelle « Le Menk » (à mi-marée) et, par les navires à passerelle plus haute, jusqu'au Nord de la tourelle. Cet alignement se situe dans le secteur blanc (289,5° – 293°) du feu de la tourelle « Ar Chaden ». La route à 293,3° laisse au Nord le plateau des Duons et au Sud la tourelle « Le Menk » (48° 43,29' N — 3° 56,70' W), cardinale Ouest lumineuse, et la Basse de Bloscon." (Shom 2021a, p. 399)

The *United States Coast Pilot* series[17] is published by the National Oceanic and Atmospheric Administration (NOAA) and covers all coasts and some inland waters of the US. Written in English and updated weekly, they contain "supplemental information that is difficult to portray on a nautical chart" and are freely available to download as PDF publications.

The US *Sailing Directions (Enroute)*[18] are published by the National Geospatial-Intelligence Agency (NGA) and cover all non-domestic coastlines in the world. Like the *United States Coast Pilots* they are written in English and updated weekly, although the descriptions given are briefer than those in the *United States Coast Pilots* and the *Instructions nautiques*. They provide "detailed coastal and port approach information, supplementing the largest scale chart of the area" and are freely available to download as PDF publications.

The *Canadian Sailing Directions*[19] or *Instructions nautiques du Canada*[20] are published by the Canadian Hydrographic Service (CHS) and cover all coasts and some inland waters of Canada. They are described as an ideal tool for "planning and assisting in navigation because they provide information that cannot be shown on a chart", and are freely available to download as PDF publications that are updated monthly and written in English or in French.

The UK *ADMIRALTY Sailing Directions*[21] are published by the UK

17. https://www.nauticalcharts.noaa.gov/publications/coast-pilot/
18. https://msi.nga.mil/Publications/SDEnroute
19. https://charts.gc.ca/publications/sailingdirections-instructionsnautiques-eng.html
20. https://cartes.gc.ca/publications/sailingdirections-instructionsnautiques-fra.html
21. https : / / www . admiralty . co . uk / publications / publications-and-reference-guides /

Hydrographic Office (UKHO) and cover the main commercial shipping routes and ports worldwide. Written in English and updated weekly, they contain "essential information to support port entry and coastal navigation for all classes of ships at sea" and are available to buy as hardback paper publications or e-books.

Some information is shared and exchanged between hydrographic services around the world, making it possible for them to produce and continually update Sailing Directions that cover regions that are not under their cartographic responsibility. This also means that much information is regularly duplicated, translated and stored independently, unnecessarily adding to workloads, increasing the data storage space required and favouring the propagation of errors. However, despite containing the most detailed information available for certain maritime areas, in particular those under French sovereignty or jurisdiction, the *Instructions nautiques* are often dismissed by mariners worldwide in favour of publications written in English.

2.4 The *Instructions nautiques* as a Case Study

In 2020, the Shom and all other authorities involved in the maritime domain were instructed by the prime minister of the French Republic to digitise their nautical information as much as possible (Gouvernement de la République française 2020). The same instruction designates the Shom as the national coordinator of the processing, formatting and digitisation of nautical information, as well as the supervision of its distribution.

The knowledge contained within the *Instructions nautiques* currently only exists in this form: French natural language text. For the Shom, this requires drafting, performing all validity checks and carrying out updates of the text manually. For the users, limited to French-speakers, it requires manually identifying the appropriate volume, then selecting the right chapter, section or subsection by using the table of contents or using a search function in their PDF reader to locate the desired information, and then reading it in detail. As it stands, it is difficult to exploit and reuse the knowledge contained within the *Instructions nautiques* given its unstructured form. In order to conform to the ministerial instruction of 2020, the Shom is looking to structure and digitise the content of the *Instructions nautiques*, which are dense with geographic information, and facilitate its distribution.

It is difficult to integrate geographic information from text-based sources into geographic information system (GIS) models, which require highly-

admiralty-sailing-directions

structured complete data with direct spatial referencing (Keller et al. 2018; Elliott and Gillies 2011). The open-world assumption of semantic Web technologies makes knowledge graphs a better solution for modelling and storing geographic information extracted from natural language text, and thus making it accessible and reusable (Janowicz et al. 2022; H. Chen et al. 2018; Melo and Martins 2017; Kuhn et al. 2014; Stadler et al. 2012). We therefore chose a knowledge graph as the solution for digitally storing a structured version of the content of the *Instructions nautiques*. This will come with many benefits for the Shom and for users of the *Instructions nautiques*.

A geospatial knowledge graph of the content of the *Instructions nautiques* could transform the manual processes for producing and updating the *Instructions nautiques* as detailed in section 2.3.3 that are currently used by the Shom. Instead of manually analysing the entire series of *Instructions nautiques* to determine the impact that a new piece of information will have as in step 1a, and instead of manually searching for the lines to be updated as in steps 3b and 4b, the knowledge base could be queried to automatically identify the relevant lines in the text.

To improve the efficiency and accuracy of these processes, and thereby increase the reliability of the *Instructions nautiques*, the Shom could apply reasoning to the knowledge graph to automatically identify and correct errors that would otherwise put the users of the *Instructions nautiques* in danger. For example, the spatial relations between entities as described in the text could be verified by using the geographic positions of the entities and vice versa. To increase the exhaustiveness of the textual content of the *Instructions nautiques*, the Shom could use inference rules to infer new knowledge from the knowledge already present in the knowledge graph. For example, the description of a spatial entity whose geographic position is described only by geographic coordinates could be improved by adding a description of its position in relation to other nearby entities.

If the knowledge graph contained multilingual labels, a semi-automatic or automatic text generation system could be implemented to help produce high-quality automatic translations of the text of the *Instructions nautiques*, making them quickly and easily available in other languages and thereby increasing their potential user base. This in turn would increase the competitiveness of the *Instructions nautiques* with regards to other Sailing Directions with global coverage written in English such as the *Sailing Directions (Enroute)* and the *ADMIRALTY Sailing Directions*. If the information contained within the other nautical publications produced by the Shom was also structured in geospatial knowledge graphs,

"The beacons and buoys carrying a light and/or a fog signal are described in the *Livre des feux et signaux de brume - LD (Saint-Pierre-et-Miquelon - Petites Antilles - Guyane)*."

Extract 2.4.1 – Translated from the original French text: "Les balises et bouées porteuses d'un feu et/ou d'un signal de brume sont décrites dans le Livre des feux et signaux de brume - LD (Saint-Pierre-et-Miquelon - Petites Antilles - Guyane)." (Shom 2021e, p. 67)

the Shom could link related pieces of information from different sources. For example, instead of citing another publication like in extract 2.4.1, the *Instructions nautiques* text could be linked directly to the relevant text in the other publication.

To modernise user access to the content of the *Instructions nautiques*, the geospatial knowledge graph could be used as the basis of a digital platform, allowing users to interrogate the content of the *Instructions nautiques* via faceted search in different languages or even by selecting their area of interest on a RNC or an ENC. The Shom could integrate knowledge and information from internal and trusted external sources to this platform to reduce the number of different resources needing to be consulted by users of the *Instructions nautiques* during itinerary planning. For example, live access to the tide predictions and weather forecast for the time and place indicated by the user could be provided.

Previous PhD projects have already dealt with certain aspects of the digitisation of the *Instructions nautiques*. Sauvage-Vincent (2017) studies the use of a controlled language to express the knowledge contained within the *Instructions nautiques* using a textual and visual grammar. The aim of the language, called Inaut, is to serve as a pivot between the personnel writing the text, the production of the printed [22] or digital publications, and the interaction with knowledge graphs and navigation equipment. Laddada (2018) presents a coastal route recommendation system for maritime navigators. It is based on an ontological model that describes the spatial, temporal and semantic components of spatial entities, which are taken into account during the formalisation of a coastal navigation trajectory. The aim is for navigators to be able to access personalised coastal route recommendations according to precise departure and arrival locations, meteorological conditions and daylight levels [23]. Both Inaut and the coastal route recommendation system are designed to function with a knowledge graph that covers the spatial entities of the coastal maritime environment and the content of the *Instructions nautiques*, however neither author defines such a graph.

22. Until recently, the *Instructions nautiques* were produced in both paper and digital formats.
23. The same landmarks cannot always be used during the day and during the night.

The aim of this thesis is to provide an approach for constructing geospatial knowledge graphs from heterogeneous text-based sources. We apply our research to the text of the *Instructions nautiques* to test our approach and because of the interest in structuring and georeferencing their content. A complete geospatial knowledge graph of the content of the *Instructions nautiques* would offer new possibilities for their production, maintenance and use, and would also make it possible to implement the research work carried out by Sauvage-Vincent (2017) and Laddada (2018).

2.5 Conclusion

Our research into approaches for creating geospatial knowledge graphs is well-suited to being applied to the *Instructions nautiques*, which are a corpus of natural French-language texts on the maritime environment. Considerable potential benefits exist for the producers of the *Instructions nautiques*, the Shom, as well as for their users. A geospatial knowledge graph of the contents of the *Instructions nautiques* would allow the Shom to more efficiently produce and maintain them by automating parts of the process. It would also allow the Shom to exploit the capabilities of linked data to provide users of the *Instructions nautiques* with access to their content in novel ways. The user base could be expanded thanks to increased accessibility, new features and making text available in other languages. All these possibilities are based on the existence of a geospatial knowledge graph of the contents of the *Instructions nautiques*.

This work is part of a larger ongoing project at the Shom and other hydrographic services worldwide to rethink how maritime data is produced, structured and analysed internally, and how it can be harnessed to expand services and improve user experiences (Murat 2023).

Chapter 3

ATONTE: a Methodology for Knowledge Graph Creation from Text and Experts

3.1 Introduction

This chapter is dedicated to giving an overview of the ATlantis Ontology and kNowledge graph development from Texts and Experts (ATONTE) Methodology for the creation of knowledge graphs from text and experts, which we will present in detail in chapters 4 to 6. We developed and refined ATONTE empirically whilst working to create a geospatial knowledge graph of the content of the *Instructions nautiques*, which we introduced in the previous chapter.

The ATONTE Methodology is for the creation of knowledge graphs from text and experts, but if the aim is to create a *geospatial* knowledge graph, it also requires structured geographic data as a source. ATONTE has been designed for use in situations where the aim is to structure all or part of the knowledge contained in a text corpus, potentially complemented by the knowledge of domain experts, and disambiguate and link the entities. Although ATONTE can be used to create knowledge graphs from texts on any subject, it is worth noting that it is particularly suited to dealing with corpora that contain spatial knowledge.

Spatial knowledge is distinct from other types of knowledge for many reasons. It is often based on subjective human perception and social agreements whilst also being objectively grounded in the real world, albeit then measured by human-defined conventions (Kuhn 2005). This duality can be illustrated by comparing the two ways in which we reference spatial knowledge. Indirect spatial referencing (see definition in section 2.3.2.3) is the way in which humans tend to share spatial knowledge, whilst direct spatial referencing (see definition in section 2.3.2.2) is the way humans have developed to "objectively" represent spatial knowledge. The former leads

to vagueness and uncertainty in locations and in relations whilst the latter guarantees some precision (Hart and Dolbear 2007). In the case where direct spatial referencing is used, the spatial relations between any spatial entities can be calculated thanks to dedicated methods (Clementini and De Felice 1997). In other words, becomes possible to refer the positions of the spatial entities with respect to each other, thereby inferring indirect spatial referencing.

Working with corpora that contain spatial knowledge such as travel guides, itinerary descriptions and historical documents is particularly interesting because they may represent the only source of knowledge about some individual spatial entities in sparsely-surveyed regions. They may also give representations of known environments at a different scale, from a different viewpoint, with a different granularity, with a different perception, in a different language or at a different time [1] compared to what has previously been recorded and what currently exists in structured reference resources (Jiménez–Badillo et al. 2020; Elliott and Gillies 2011; Beall 2010; Kuhn 2005).

ATONTE can be divided into three main components. The first component involves manually developing a domain ontology whose scope allows modelling the main knowledge contained within the text corpus. Consulting domain experts is recommended to help orient the model towards the desired use case. The second component deals with the automatic extraction of entities and relations from the text corpus. The third and final component involves populating the knowledge graph with the extracted entities and relations, and performing entity canonicalisation.

In section 3.2 we present our review of the related work for this chapter, which we have divided into three parts. First, in section 3.2.1 we discuss the different definitions for knowledge graphs and related terms. Then, in section 3.2.2 we give an application- and domain-independent review of knowledge graph creation surveys before reviewing individual approaches in section 3.2.3. We summarise our findings in section 3.2.4 and then discuss how we developed the ATONTE Methodology with respect to the related work in section 3.2.5. In section 3.3 we present the ATONTE Methodology and give an overview of its components before concluding in section 3.4. The three main components of the ATONTE Methodology will be presented in detail in chapters 4 to 6.

1. Time of day, time of year, time in history, etc.

3.2 Related Work

3.2.1 Knowledge Graph Definitions

Hogan et al. (2021) define a knowledge graph as "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities". A similar definition is given by N. Noy et al. (2019), who state that knowledge graphs describe "objects of interest and connections between them", adding that many "practical implementations impose constraints on the links in knowledge graphs by defining a *schema* or *ontology*". Hogan et al. note that a schema "defines a high-level structure for the knowledge graph", the RDF Schema (RDFS) being a prominent standard, but that "the semantics of terms used in a graph can be defined in much more depth" than what is possible with a schema, for example by using the Web Ontology Language (OWL). Following a review of existing definitions, Ehrlinger and Wöß (2016) suggest that a "knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge". Weikum et al. (2021) distinguish between knowledge graphs and knowledge bases by stating that knowledge graphs are a type of knowledge base that contain only binary relations. According to Weikum et al. (2021), knowledge bases "comprise salient information about entities, semantic classes to which entities belong, attributes of entities, and relationships between entities". The ISO/IEC 21838-1:2021 standard defines a knowledge base as a "combination of an ontology with a collection of data which terms in the ontology have been used to describe, classify or connect" (International Organization for Standardization and International Electrotechnical Commission).

3.2.2 Knowledge Graph Creation Surveys

Weikum et al. (2021) carry out a comprehensive survey of the fundamental concepts and practical methods for creating and curating large knowledge graphs from online content and unstructured text sources. They divide the knowledge graph construction process into four major tasks: knowledge discovery, entity canonicalisation, knowledge graph augmentation, which are part of the knowledge graph creation step, and knowledge graph cleaning, which makes up the knowledge graph curation step. These four tasks are described as follows:

— Knowledge discovery involves defining the scope of the graph and selecting the most appropriate knowledge sources, and then performing

entity detection to find mentions of entities in the selected sources.

— Entity canonicalisation is the task of identifying synonymous mentions and disambiguating them according to a reference repository of entities.

— Knowledge graph augmentation involves adding properties to the entities present in the knowledge graph, either by adding their attributes or by linking them to other entities via relations.

— Knowledge graph cleaning requires corroborating statements, removing invalid ones and ensuring the quality of the graph for the duration of its lifespan.

Four main types of input sources for knowledge discovery are cited by Weikum et al., in descending order of priority:

— Premium sources such as Wikipedia [2] or GeoNames [3]

— Semi-structured data such as infoboxes

— Natural language text

— Mass-user data about online behaviour such as queries or clicks

Weikum et al. recommend that priority be given to premium sources because they contain the most reliable knowledge and can provide prominent entities with which to populate the core of a knowledge graph. However, premium sources are usually poor in *long-tail* entities, which are entities that are uncommon or obscure. Weikum et al. detail various methods for further populating a knowledge graph established from premium sources by discovering and typing entities from semi-structured and textual sources. They include dictionary-based (Keller et al. 2018; Moncla 2015; Gaio et al. 2012), pattern-based (Reyes-Ortiz 2019; Yangarber et al. 2000) and machine learning approaches (Sugathadasa et al. 2018; Wang et al. 2018; F. M. Suchanek et al. 2006), and are sometimes used in combination with one another (Lamotte et al. 2020; Qiu et al. 2020; Nguyen and Cao 2007). When premium sources cannot be exploited first, discovering and typing entities directly from semi-structured and textual sources is known as taxonomy induction, or ab-initio taxonomy construction.

According to Weikum et al., entity canonicalisation primarily refers to three tasks: entity linking, coreference resolution and entity matching. When dealing with a pre-populated knowledge graph, entity linking involves linking newly-discovered mentions of entities in semi-structured and textual sources with their corresponding entities already present in the

2. https://en.wikipedia.org/
3. https://www.geonames.org/

knowledge graph. Other authors indicate that new mentions can alternatively be linked with their corresponding entities in external knowledge graphs (Hogan et al. 2021). This task can be carried out with algorithmic (Ling et al. 2015; Ceccarelli et al. 2013) or learning-based methods (Shen et al. 2023). Both methods take into account mention-entity popularity, which considers the global popularity of the candidate entities, mention-entity context similarity, which compares the similarity of the text surrounding a mention with a description of the entity, and entity-entity coherence, which compares the semantic similarity of the candidate entities of co-occurring mentions. When dealing with mentions of entities that do not yet exist in the knowledge graph, which are usually long-tail entities, Weikum et al. note that they should not be linked to any entity and should be classed as *null* or `rdf:nill`. These entities could later become candidates for inclusion in an updated version of the knowledge graph. Coreference resolution deals with the grouping of mentions that refer to the same entity into an equivalence class of mentions. The mentions could use different names or even descriptions to refer to the same entity. The mentions within an equivalence class can later be linked via entity linking to their corresponding entity in the knowledge graph. The task of entity matching has similar considerations to entity linking but involves matching entities that exist in multiple resources, without a base knowledge graph, for example via name or context similarity comparisons (Mudgal et al. 2018).

In their survey, Weikum et al. explain the different methods for knowledge graph augmentation, which include pattern-based (F. Suchanek et al. 2008) and machine learning approaches (Yao et al. 2019). They discuss the case in which the attributes and relations are pre-defined as well as open information extraction for predicate discovery in which the schema is unknown.

Finally, they present methods and good practices for knowledge graph curation. The quality and completeness of a knowledge graph can be evaluated thanks to various metrics (Paulheim 2017), the knowledge it contains can be cleaned via consistency reasoning (Bonatti et al. 2011), and life cycle management can be aided by provenance tracking and versioning (Hoffart et al. 2013). Weikum et al. note that it is desirable for knowledge graph creation and curation to be as automatic as possible, but that this statement is dependent upon the requirements for the final knowledge graph. For example, if a knowledge graph is destined for advanced and fine-grain usage in which correctness is of utmost importance, more human involvement may be necessary in the creation process to achieve a high-quality

knowledge graph.

Hogan et al. (2021) provide a thorough introduction to knowledge graphs that includes a discussion about the place that a schema occupies in the creation of knowledge graphs. The schema for a knowledge graph can either be generated based on external sources before creating the knowledge graph or extracted from the graph itself once it has already been established, the latter of which is similar to the open information extraction technique suggested by Weikum et al. (2021) for predicate discovery. Hogan et al. describe two different ways in which a knowledge graph schema, or ontology, can be generated from external sources: ontology engineering and ontology learning. Ontology engineering involves applying a methodology to develop an ontology in a primarily manual manner (Shimizu et al. 2022; Peroni 2016a; Vrandečić et al. 2005; N. F. Noy and McGuiness 2001; Ferndández et al. 1997) whilst ontology learning involves automatically or semi-automatically extracting concepts and definitions from a source (Cimiano 2006; Buitelaar et al. 2005). Hogan et al. suggest that an ontology produced though ontology learning could be used as the starting point for an ontology engineering process, which could manually validate and improve the learnt ontology. Hogan et al. define three types of schemata that can be extracted from established knowledge graphs: semantic schemata, validating schemata and emergent schemata. Semantic schemata allow defining the meaning of high-level terms via the use of standards such as RDFS [4] and OWL [5]. Such schemata in turn allow the inference of new knowledge within the graph. If we wish to validate existing graph knowledge we must apply a validating schema, which allows defining constraints on the content of the graph using a language such as Shapes Constraint Language (SHACL) [6]. Domain experts are required to partake in the definition of semantic and validating schemata. Alternatively, emergent schemata can be extracted automatically from the latent structures present in a graph when an ontology is not present. This is known as graph summary or ontology discovery (Čebirić et al. 2019) and can be used to aid with the task of defining a semantic or validating schema.

Hogan et al. divide the main sources that can be used for knowledge graph creation and enrichment into three types: text sources, markup sources such as HTML, and structured sources in formats such as CSV and JSON. The former corresponds to one of the types suggested by Weikum et al. (2021), *natural language text*, whilst the latter two would both correspond to the same type: *semi-structured data*. Hogan et al. therefore

4. https://www.w3.org/TR/rdf12-schema/
5. https://www.w3.org/TR/owl-overview/
6. https://www.w3.org/TR/shacl/

give a narrower but stricter definition of the possible input sources. They also note that the flexible nature of knowledge graphs lends itself to first creating a core that can later be incrementally enriched as required.

Ma (2022) presents a review of knowledge graph construction approaches and knowledge graph applications in the geosciences. They group knowledge graph creation techniques into top-down approaches and bottom-up approaches. According to Ma, top-down approaches begin with delimiting the domain of the model and defining a conceptual model of entities, categories and relationships. This model is then formalised and later populated with domain knowledge. Ma describes this process as allowing the transformation of a human representation of a domain into a machine-readable format (Garvie 1995). This approach corresponds to what Hogan et al. (2021) called ontology engineering. Bottom-up approaches, according to Ma, involve using natural language processing (NLP) or text mining on largely unstructured sources such as social media posts and open-access publications and then using the results to guide knowledge graph construction (Fan et al. 2020). This approach, on the other hand, corresponds to what Hogan et al. (2021) called ontology learning. Ma argues that although bottom-up approaches are able to process large volumes of data to quickly produce substantial knowledge bases, they lack the precise representations achieved through top-down approaches and often require human intervention to refine their structure.

3.2.3 Knowledge Graph Creation Approaches and Examples

Iglesias et al. (2023) explain that creating knowledge graphs from unstructured data such as text requires two main steps: one that involves defining a semantic layer to describe the data, and one that involves manipulating the data to extract and link entities. This corresponds to the idea by Hogan et al. (2021) that one of the ways for a knowledge graph schema to be generated is to base it on external sources before creating the knowledge graph. Chessa et al. (2022) implement the steps suggested by Iglesias et al. to construct a knowledge graph from a data lake [7] containing textual data. They first create a semantic layer in the form of a domain ontology before identifying and extracting entities from the text and automatically generating Resource Description Framework (RDF) triples. Detailed information on the ontology development and entity extraction process is not given.

Janowicz et al. (2022) introduce the KnowWhereGraph, a cross-domain knowledge graph containing high-resolution spatial and temporal informa-

7. A data lake is a system that can store large amounts of raw data in various formats (Fang 2015).

tion about subjects such as administrative boundaries, climate, crops and transportation. It is oriented towards applications in environmental intelligence and makes it possible for decision makers to quickly and easily access current or historical information about any region on earth. It is continuously populated from heterogeneous data sources, which requires an integration process that can handle potentially noisy, missing and contradictory data. To satisfy real-time data integration whilst maintaining a high-quality graph model, Janowicz et al. combine top-down and bottom-up ontology engineering approaches.

Mansfield et al. (2021) present their experience of creating an enterprise knowledge graph from structured data. They use the Pay-as-you-go Methodology (Sequeda et al. 2019) for designing and building enterprise knowledge graphs from relational databases. Mansfield et al. argue that automatic approaches for capturing domain knowledge from databases are seldom able to capture the complexity of a domain and therefore advocate more manual approaches that facilitate knowledge capture from users, creators and maintainers of the database.

Schröder et al. (2021) introduce the Spread2RML approach, which predicts RDF Mapping Language (RML) [8] mappings for noisy spreadsheets in order to create knowledge graphs of the semi-structured data. They argue that it eliminates the need to perform a direct conversion from spreadsheets to RDF statements whilst still allowing users to adjust and correct mapping definitions.

Simsek et al. (2021) present an approach and tools for creating, hosting, curating and deploying knowledge graphs. Their approach involves defining mappings from hierarchical data sources about tourism in formats such as JSON and XML to Schema.org [9] vocabularies. They show in a case study that their approach scales well to large amounts of data, however they note that real-world industry data can pose problems even when structured, such as incompleteness.

Kertkeidkachorn and Ichise (2018) introduce T2KG, an approach for creating open-domain knowledge graphs from natural language text. The first step involves identifying entity mentions in text, which is called knowledge discovery by Weikum et al. (2021), and mapping them to existing identical entities in a knowledge graph, which is called entity linking by Weikum et al., or creating new knowledge graph entities if an identical match is not found. Then, an unspecified open information extraction technique is used to extract relation triples, composed of two entities and a relation, from the text. Duplicate entities are grouped via a process

8. https://rml.io/specs/rml/
9. https://schema.org/

called coreference resolution by Weikum et al., relations are transformed into individual predicates and literal objects are left as such. Finally, the predicates are mapped to existing identical predicates in a knowledge graph using a hybrid combined rule-based and similarity-based approach.

3.2.4 Summary

Our review reveals that there is neither a universal method for creating knowledge graphs, nor a universal way of defining, formalising and categorising knowledge graph creation approaches.

The sources that can be used as input to the process of creating a knowledge graph are generally divided into three or four types including at least structured data such as existing knowledge graphs, semi-structured data, and unstructured data such as natural language text (Hogan et al. 2021; Weikum et al. 2021). Knowledge can be extracted from input sources using either a predefined schema or through open information extraction (Ma 2022; Weikum et al. 2021; Kertkeidkachorn and Ichise 2018). The former implies generating the schema, usually an ontology, before creating the knowledge graph whilst the latter requires first carrying out the extraction task and creating the knowledge graph before being able to extract its schema (Hogan et al. 2021). When dealing with structured or semi-structured input sources, it is possible to define mappings between the source and the schema to automatically create a knowledge graph (Schröder et al. 2021; Simsek et al. 2021; Kertkeidkachorn and Ichise 2018). It is generally accepted that knowledge graph creation approaches should be automatised as much as possible and require as little human involvement as possible, but that this implies a trade-off between efficiency and correctness (Ma 2022; Hogan et al. 2021; Weikum et al. 2021). To maximise control over the knowledge graph structure, and to be able to integrate domain expert knowledge, manual ontology development approaches can be used (Ma 2022; Mansfield et al. 2021).

3.2.5 Discussion

We designed the ATONTE Methodology in light of our application context as presented in chapter 2, which requires creating a geospatial knowledge graph to represent the content of our corpus of the *Instructions nautiques*, and with respect to the recommendations and experiences published in the related work that we have reviewed in this section. The characteristics of our application context can be summarised with the following points:

— We wish to structure the content of a natural language text corpus as a knowledge graph
— Our corpus contains spatial knowledge meaning that we must create a geospatial knowledge graph
— We will need to disambiguate and associate direct spatial referencing to the spatial entities in the graph
— Our corpus covers a complex and technical domain
— We will require the input of domain experts to orient the structure of knowledge graph (the ontology) towards its future application

Although our application context required creating a geospatial knowledge graph from a text corpus dense with spatial knowledge, we decided to generalise and formalise the approach that we implemented as a methodology for the creation of geospatial or non-geospatial knowledge graphs from text. The remainder of this section is dedicated to positioning the ATONTE Methodology according to the related work reviewed in section 3.2.

ATONTE is a semi-automatic approach for knowledge graph creation from unstructured natural language text and expert knowledge, and optionally structured geographic data. The aim of the ATONTE Methodology is to structure all or part of the knowledge contained within a text corpus for a specific knowledge graph application according to an ontology defined with the help of domain experts. We consider that the knowledge content of the corpus is almost complete, lacking only the direct spatial referencing for each spatial entity mentioned in the corpus in the form of geographic coordinates in the case where a *geospatial* knowledge graph is required.

The direct integration of knowledge from premium sources into the knowledge graph as recommended by Weikum et al. (2021) occurs during the third and final stage of ATONTE, and only if a *geospatial* knowledge graph is being created, to add the geographic coordinates from a to spatial entities in the graph. In this case, the premium source that provides the geographic coordinates is a reference geographic resource.

As indicated by Mansfield et al. (2021), interacting with domain experts facilitates the correct modelling of complex domains. ATONTE therefore requires domain experts to participate in the definition of the schema to orient its structure towards the future application of the knowledge graph. If the text corpus covers a technical domain, it is likely to contain long-tail entities (Weikum et al. 2021) that can be better integrated into a schema by domain experts rather than an automatic approach. In ATONTE there is a significant human component in the definition of the schema, to ensure high correctness and coverage (Weikum et al. 2021).

ATONTE relies on a predefined entity and predicate extraction schema in the form of an ontology as opposed to using open information extraction for predicate discovery (Weikum et al. 2021). This is necessary to be able to integrate the knowledge of domain experts and to limit the extraction to entities and relations defined with the domain experts in the ontology.

Using the terms defined by Iglesias et al., ATONTE first deals with generating a semantic layer before manipulating the data to extract and link entities. According to the definitions provided by Hogan et al. (2021), ATONTE involves generating the schema for the knowledge graph based on external sources before creating the graph itself, where the schema is the ontology, developed via ontology engineering, and the external sources are the text corpus and the domain experts.

In ATONTE, the ontology is developed as a seed ontology that covers the core concepts of the domain and that can later be extended, as suggested by Hogan et al. (2021), by analysing the knowledge extracted from the corpus. ATONTE therefore employs a top-down approach for knowledge graph creation (Ma 2022).

ATONTE requires carrying out both entity discovery and predicate discovery before populating the knowledge graph rather than following the ordering suggested by Weikum et al. (2021). This allows the attribute and relation information to be leveraged during entity disambiguation. However, ATONTE is in line with the trend detailed by by Weikum et al. (2021) towards the use of deep neural networks for knowledge discovery, as we perform both entity and predicate extraction with a Transformer network (Vaswani et al. 2017). Such approaches have been shown to give better results than rule-based approaches for knowledge extraction from unstructured text (Nismi Mol and Santosh Kumar 2023).

3.3 Our Approach: ATONTE

3.3.1 Overview

The ATONTE Methodology for the creation of knowledge graphs from unstructured natural language text and expert knowledge, and optionally structured geographic data, is composed of three main stages. It also includes a preliminary stage to be carried out before starting to implement the methodology, which is designed to help with verifying that it is the right knowledge graph creation methodology to be used for a given project. A schema that illustrates the main inputs, outputs and processes involved in ATONTE was presented in figure 1.4.1 on page 6. The stages involved in ATONTE are:

0. Feasibility Study
1. Ontology Development from Text and Experts
2. Entity and Relation Extraction from Text
3. Information Structuring and Entity Disambiguation

In the following sections we present each of the stages involved in ATONTE.

3.3.2 ATONTE Stage 0: Feasibility Study

Before applying our methodology, we recommend carrying out a lightweight feasibility study to verify that it is suited to the application. Each of the following requirements should be satisfied for ATONTE to be a good choice of geospatial knowledge graph creation methodology:

— The aim of the geospatial knowledge graph is to represent, in a structured fashion, some or all of the (spatial) knowledge contained in a text corpus

— The structure of the geospatial knowledge graph is not yet known and needs to be defined for a specific application without necessarily mirroring the structure of the original corpus

— The text corpus contains spatial elements and uses direct and/or indirect spatial referencing

— The text corpus is accessible in a clean digital format

— Experts of the text corpus domain (producers, maintainers, users) and of the geospatial knowledge graph application (future producers, maintainers, users) are known and can be consulted on an ad-hoc basis during the knowledge graph creation process

3.3.3 ATONTE Stage 1: Ontology Development from Text and Experts

The first component of ATONTE is dedicated to defining the structure of the future knowledge graph via the manual development of a domain ontology from the text corpus with the help of domain experts. To ensure the geospatial aspect of the knowledge graph, the ontology must contain concepts that allow modelling spatial entities, their properties and the spatial relations between them.

The ontology development process begins with carrying out groundwork, which includes becoming familiar with the corpus, analysing related domain resources, defining the knowledge graph application, creating a

preliminary informal dataset, identifying domain experts, and and dividing the domain into subdomains.

The next step involves creating the documentation that will define the scope of the ontology. A motivating scenario, a list of informal competency questions and a glossary of terms used are drafted, using the knowledge present in the text corpus and that volunteered by the domain experts, for each subdomain.

Once the documentation has been validated by domain experts, the modelling phase begins. This is carried out empirically, for each subdomain in parallel, by manually semi-formalising the knowledge contained in extracts of the text corpus corresponding to the subdomain. Recurrent concepts are grouped and expressed in a formal language to build up a vocabulary of classes and properties for each subdomain.

Each subdomain model undergoes a series of three tests, to verify its agreement with the documentation, its global coherence and its ability to be populated with instances of triples. The subdomain models can then be merged together to create an ontology of the full domain of the text corpus. Finally, a manual refactoring process aligns the model with pertinent and useful external semantic resources.

The full domain ontology development stage, and the way in which it related to the state of the art, is presented in detail in chapter 4.

3.3.4 ATONTE Stage 2: Entity and Relation Extraction from Text

The second component of ATONTE deals with automatically extracting geographic information from the text corpus using a supervised deep learning approach. Again, to ensure the geospatial aspect of the knowledge graph, spatial entities and the spatial relations between them must be extracted.

Guided by the structure of the ontology, an annotation scheme is designed to define the elements of the text corpus that we aim to extract. The annotation scheme isolates the entity type when it is present within the entity name, giving nested annotations. A representative portion of the corpus is selected and formatted before being manually annotated according to the scheme.

The annotated dataset is then used to train a pretrained language model to identify and classify the annotated elements. Once trained, the model can be used to extract the desired geographic information from the entire text corpus.

The full geographic information extraction stage, and the way in which

it related to the state of the art, is presented in detail in chapter 5.

3.3.5 ATONTE Stage 3: Information Structuring and Entity Disambiguation

The third and final component of ATONTE addresses the semi-automatic population of the ontology with the entities and relations extracted from the text corpus to create a knowledge graph.

RDF triples are automatically produced to structure the information extracted from the text corpus according to the domain ontology. Concepts extracted from the text but not yet represented in the ontology can be added to extend its coverage of the domain.

The entities are deduplicated and disambiguated via automatic entity linking with a reference resource. The entity types identified thanks to the nested annotations are used to help this process.

The full ontology population stage, and the way in which it related to the state of the art, is presented in detail in chapter 6.

3.4 Conclusion

ATONTE is a methodology to create knowledge graphs from text and experts, and optionally structured geographic data. It is composed of three main stages: ontology development from text and experts, entity and relation extraction from text, and entity disambiguation and structuring to create a knowledge graph. In the following three chapters, detailed explanations of the components that make up ATONTE are given. Each chapter includes a review of the related work specific to the component. We also demonstrate how we implemented ATONTE to develop and populate the coAsTaL mAritime NavigaTion InstructionS (ATLANTIS) Ontology from a text corpus made up of *Instructions nautiques*.

Chapter 4

Ontology Development from Text and Experts

4.1 Introduction

In this chapter we present the first component of the ATlantis Ontology and kNowledge graph development from Texts and Experts (ATONTE) Methodology: a methodology for the development of a domain ontology from text and experts. We demonstrate how we implemented it to develop the coAsTaL mAritime NavigaTion InstructionS (ATLANTIS) Ontology, a geospatial seed domain ontology of the *Instructions nautiques*, from text and experts.

Ontology development can be carried out manually, with significant human involvement, semi-automatically, or automatically, with little human involvement. Despite being more time-consuming, manual ontology development can yield better results (Jiménez–Badillo et al. 2020) as it ensures high correctness and coverage (Weikum et al. 2021). A manual approach facilitates the participation of domain experts, who can ensure the correct modelling of complex domains (Mansfield et al. 2021) and also the correct handling of long-tail entities (Weikum et al. 2021).

So as to be able to easily integrate contributions from domain experts and to have total control over the knowledge graph structure, which is not necessarily reflected in the text corpus, we adopt a manual approach to ontology development from text. The ontology must have a solid core that can be automatically enriched (adding classes and properties) and populated (adding instances). In the case where the ontology will be automatically enriched, it suffices to create a seed ontology.

In section 4.2 we present the related work for this chapter, which we have divided into three parts. First, in section 4.2.1 we discuss the different definitions for ontologies. Then, in section 4.2.2 we review existing maritime ontologies before analysing existing ontology development methodologies

in section 4.2.3. Section 4.3 is dedicated to presenting the first component of ATONTE, which consists of a methodology for the manual development of domain ontologies from text and experts. For each step in the methodology we give a comprehensive domain-independent description of its purpose and detail the tasks to be carried out. Then, in section 4.4 we illustrate the methodology by showing how we implemented it, step-by-step, on the *Instructions nautiques*. In section 4.5 we present the results of this implementation: the ATLANTIS Ontology, which we then evaluate in section 4.6 before concluding in section 4.7.

4.2 Related Work

4.2.1 Ontology Definitions

The ISO/IEC 21838-1:2021 standard defines an ontology as a "collection of terms, relational expressions [1] and associated natural-language definitions together with one or more formal theories designed to capture the intended interpretations of these definitions" (International Organization for Standardization and International Electrotechnical Commission 2021). It defines a formal theory as a "collection of definitions and axioms expressed in a formal language", which is a "language that is machine readable and has well-defined semantics". It states that an ontology is an artefact for use by humans and by computers: the terms and relational expressions are expressed using natural language, and are also captured in a machine-readable formal language with well-defined semantics. There exist many formal languages that can be used for ontology implementation (Maniraj and Ramakrishnan 2010), including those that are based on description logic such as Web Ontology Language (OWL) and those that are based on first-order logic such as Common Logic (CL). First-order logic-based languages allow more expressivity than description logic-based languages and can therefore formally capture the implications of more complex axioms (International Organization for Standardization and International Electrotechnical Commission 2021). However, description logic-based languages are for the most part decidable [2], which means that they can be used for logical reasoning by computers (International Organization for Standardization and International Electrotechnical Commission 2021).

Within the field of computer science, ontologies are generally classi-

1. A relational expression is an "expression used to assert that a relation obtains", where a *relation* refers to the real-world link between entities (International Organization for Standardization and International Electrotechnical Commission 2021).

2. A decidable language is a "language for which membership can be decided by an algorithm that halts on all inputs in a finite number of steps" (Black 1999).

fied into two levels: top-level ontologies (also known as upper ontologies, high-level ontologies or foundation ontologies) and domain ontologies (Biemann 2005). Top-level ontologies are generic by definition, and serve as the basis for more specific ontologies: domain ontologies. The same classification of ontology levels is given in the ISO/IEC 21838-1:2021 standard: a top-level ontology "is created to represent the categories that are shared across a maximally broad range of domains" whilst a domain ontology is an "ontology whose terms represent classes or types and, optionally, certain particulars (called 'distinguished individuals') in some domain" (International Organization for Standardization and International Electrotechnical Commission 2021).

Guarino (1998) suggests another, finer division of levels. Alongside top-level ontologies and domain ontologies, they define task ontologies and application ontologies. Task ontologies are considered to be at the same level as domain ontologies, both of which are specialisations of top-level ontologies. Domain ontologies describe a given domain, such as vessels or coastal landmarks, whilst task ontologies describe a specific activity, such as navigating. Application ontologies are even more specific than domain and task ontologies. They are designed to describe both a given domain and a specific task, in other words, to satisfy a chosen application.

A seed ontology covers the fundamental aspects of its subject area in such a way that it can be easily expanded without changing its core structure (Weinstein and Alloway 1997).

4.2.2 Maritime Ontologies

We conducted a review of existing semantic resources on the maritime domain by searching in catalogues such as Linked Open Vocabularies (LOV) [3] and by searching for published articles that describe them. We identified one ontology that deals specifically with maritime navigation. The remainder of the maritime-related ontologies and thesauri identified during our review are dedicated either to navigation modelling for surveillance and security, or to scientific research in marine biology, chemistry or geology, with a focus on marine life and the environment.

4.2.2.1 Ontologies for Maritime Navigation

Malyankar (2001) proposes an ontology for maritime information and nautical chart symbology based on official sources such as the *United States Coast Pilot*. The aim is to offer users of these books a platform that

3. https://lov.linkeddata.es/

can retrieve elements from XML-based marked-up volumes of the *United States Coast Pilot* via queries. This could have constituted a useful basis for our work, however the ontology does not seem to have been published on the Web and we were unable to find any trace of it apart from in the publications that describe it.

4.2.2.2 Ontologies for Maritime Security

Vandecasteele and Napoli (2012) present a spatial ontology, associated with a geographical inference engine, to automatically identify suspicious vessels and their likely behaviour in a bid to improve maritime surveillance techniques. The European e-Compliance project worked to create an ontology on maritime regulations that apply to vessels and ports (Hagaseth et al. 2016). The aim of the project was to develop tools to help reduce the administrative load on actors in the maritime domain by creating and managing machine-readable regulations. Liang and Zhai (2018) introduce an ontology built to help construct linked data for shipping. As far as we are aware, none of the three above ontologies have been published on the Web.

4.2.2.3 Ontologies for Marine Life and the Environment

Tzitzikas et al. (2013) present a top-level ontology to improve semantic interoperability of marine data for biodiversity between scientific disciplines. The ontology introduced by Leadbetter et al. (2010) is dedicated to marine biology and the evolution of the maritime environment. Neither of these two ontologies have been published on the Web.

The Semantic Web for Earth and Environmental Terminology (SWEET) Ontologies are a set of around 200 ontologies that cover Earth system sciences, originally developed by the NASA Jet Propulsion Laboratory (Raskin and Pan 2005). Two SWEET Ontologies contain elements that are related to our work: Property Space Direction [4], which includes terms for expressing directions such as `angle` and `bearing`, and Realm Hydrosphere Body [5] which includes terms for meteorological and oceanographic phenomena.

The GEneral Multilingual Environmental Thesaurus (GEMET) [6], developed by the European Environment Agency, is dedicated to general terminology for the environment. It contains multilingual terms for different types of spatial entities, including those that can be found in the

4. http://sweetontology.net/propSpaceDirection/
5. http://sweetontology.net/realmHydroBody/
6. http://www.eionet.europa.eu/gemet

maritime environment, such as `bay`, `port` and `reef`.

EuroVoc [7] is a multilingual thesaurus dedicated to the activities of the European Union. It contains some terms related to the maritime environment such as `cargo vessel`, `fishing regulations` and `harbour installation`.

The *Thésaurus Eau* [8], a set of vocabularies on the subjects of water and biodiversity, contains terms that refer to different types of spatial entities related to the maritime domain such as `waterway` and `anchorage`.

The NERC Vocabulary Server (NVS) is managed by the British Oceanographic Data Centre and the National Oceanography Centre, and is financed by the UK Natural Environment Research Council (NERC). It is dedicated to marine science and hosts thesauri covering oceanography. The following thesauri are of particular interest to us: the Oregon Coastal Atlas Coastal Erosion Thesaurus [9], the MIDA Coastal Erosion Thesaurus [10] and the Marisaurus Thesaurus [11]. They contain terms for different types of maritime spatial entities such as `beacon` and `shipwreck`. The NVS also hosts two thesauri that contain terms for different types of vessels such as `fishing vessel` and `naval vessel`: the World Meteorological Organisation voluntary observing ship category [12] and the SeaVoX Platform Categories [13].

A first version of the Maritime Domaine Ontology [14] has been developed by the Open Simulation Platform. Their aim is to be able to use the ontology as a framework for attributing properties to digital models of elements from the maritime domain such as hulls, motors, waves, winds and currents.

Finally, the high-level domain-independent PROTON Ontology contains some useful terms for describing the maritime environment in its extent module [15] such as `beacon`, `lighthouse` and `water current`.

None of the existing maritime ontologies that we found cover more than a few of the concepts required to model the *Instructions nautiques*, which is why in the next section we review ontology development methodologies in view of creating an ontology from scratch. We will align our ontology with these existing resources as much as possible.

7. https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc
8. http://thesaurus.oieau.fr/thesaurus/
9. http://vocab.nerc.ac.uk/collection/A02/current/
10. http://vocab.nerc.ac.uk/collection/A04/current/
11. http://vocab.nerc.ac.uk/collection/P21/current/
12. http://vocab.nerc.ac.uk/collection/C31/current/
13. http://vocab.nerc.ac.uk/collection/L06/current/
14. https://opensimulationplatform.com/mdo/
15. http://www.ontotext.com/proton/protonext

4.2.3 Ontology Development Methodologies

The vast number of manual domain ontology development methodologies available means that we have created a shortlist of 10 to analyse in this section. We selected the most-well known methodologies, some of which have been updated by their creators (Peroni 2016a; Kotis and Papasalouros 2010; Jarrar and Meersman 2008), have been adapted or extended by members of the community (Moor et al. 2006; Vrandečić et al. 2005), or have been used to develop well-established ontologies covering a variety of domains. The methodologies that make up our shortlist are:

1. METHONTOLOGY
   — Created by Ferndández et al. (1997)
2. Ontology Development 101
   — Created by N. F. Noy and McGuiness (2001)
3. On-To-Knowledge (OTK)
   — Created by Sure and Studer (2001)
   — Updated by Sure (2003)
4. Developing Ontology-Grounded Methods and Applications (DOGMA)
   — Created by Jarrar and Meersman (2002)
   — Updated by the same authors (Jarrar and Meersman 2008)
5. DOGMA Meaning Evolution Support System (DOGMA-MESS)
   — Created by Moor et al. (2006)
6. DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies (DILIGENT)
   — Created by Vrandečić et al. (2005)
7. Human-Centered Ontology Engineering Methodology (HCOME)
   — Created by Kotis and Vouros (2006)
   — Updated by Kotis and Papasalouros (2010)
8. Networked Ontologies (NeOn)
   — Created by Suárez-Figueroa et al. (2012)
9. Simple Agile Methodology for Ontology Development (SAMOD)
   — Created by Peroni (2016b)
   — Updated by the same author (Peroni 2016a)
10. Modular Ontology Modeling (MOMo)
    — Created by Shimizu et al. (2022)

Table 4.2.2 on page 52 gives a comparative overview of some of the main features of the 10 methodologies and table 4.2.1 on page 51 shows the order in which the principal activities that they contain are carried out.

4.2.3.1 METHONTOLOGY

METHONTOLOGY (Ferndández et al. 1997) is a structured methodology for the manual development of ontologies from scratch. It was created empirically by the authors during the development of an ontology in the chemical domain. METHONTOLOGY details a series of six activities to be carried out in order, the necessary techniques to perform them, and the deliverables to be produced at the end of each activity. The first activity involves writing an ontology specification document in natural language. This document should specify the desired characteristics of the ontology such as its objective, its users, its formality, its scope and its granularity. The second activity consists of knowledge acquisition. It is recommended to use sources such as experts, written documents and existing ontologies to find and define concepts, their properties and their relations for the ontology. These elements are then structured during the third activity to create a conceptual model that corresponds to the specification document created during the first activity. Ferndández et al. note that, in reality, these first three activities are to be carried out simultaneously. The fourth activity involves integrating concepts and definitions from existing ontologies. For each concept integrated from another ontology, a record must be made of its origin and any changes made to its definition. During the fifth activity, the ontology is defined in a formal language. The sixth and final activity consists of evaluating the ontology. Ferndández et al. recommend using the evaluation guide written by Gómez-Pérez et al. (1995), as well as writing a document that describes the way in which the ontology was evaluated and the errors found.

4.2.3.2 Ontology Development 101

Ontology Development 101 (N. F. Noy and McGuiness 2001) is an initial ontology development guide, aimed at beginners, that is based on the ontology development experiences of the authors. It is an iterative methodology that advises starting by developing a draft of the ontology before iteratively revising and refining it. With each iteration, finer details can be added to the ontology. This methodology divides the ontology development process into seven steps. Like METHONTOLOGY, the first step involves describing the domain, scope and use of the ontology. To help with this process, it is recommended to write a list of competency questions. Competency questions are written in natural language and give an indication of the questions to which a knowledge graph based on the final ontology should be able to give an answer. The second step requires examining existing ontologies in case all or part of one or some could be reused or extended.

During the third step, a full list of the terms and properties that the ontology should contain is written. In Ontology Development 101, steps four and five are closely linked and therefore N. F. Noy and McGuiness recommend carrying them out in parallel. Step four consists of defining the ontology classes and their hierarchy, according to the list of terms drafted during step three, using an ontology editing software. This can be done either by adopting a top-down approach and starting with the most general terms, a bottom-up approach and starting with the most specific terms, or a combination of the two. Step five requires creating the properties of the classes according to those written in the list of terms during step three. Step six involves describing the class properties, for example by assigning their domain and range. The seventh and final step consists of instantiating the classes with individuals. At the end of the guide, N. F. Noy and McGuiness list some more specific pieces of advice such as how to define and name classes in an ontology.

4.2.3.3 OTK

OTK (Sure 2003; Sure and Studer 2001) is a set of tools for knowledge management projects. It includes an ontology development methodology for domain and application ontologies. Before beginning the development of an ontology, OTK requires carrying out a feasibility study in the same way as in CommonKADS (Schreiber et al. 1999), a methodology to support structured knowledge engineering. The feasibility study must take into account the scientific feasibility of the project, as well as economic and technical aspects. As in METHONTOLOGY and Ontology Development 101, the first step in OTK involves writing an ontology requirements specification document (ORSD). This document should contain six elements:

1. The domain and goal of the ontology
2. Design guidelines such as granularity estimations and naming conventions
3. Knowledge sources including existing semantic resources, documentation and domain experts
4. Potential users of and uses for the ontology
5. Competency questions
6. Details of the software environment in which the ontology will be implemented

At the end of this first step, a hierarchy should be made out of the key concepts and relations to create what Sure and Studer call a *baseline ontology*.

Three different approaches are suggested for this task: top-down, bottom-up or middle-out, which involves starting with the most relevant concepts before defining more and less specific ones, although Sure and Studer note that the best approach may be to use a combination. The second step in OTK is the refinement phase and involves producing an ontology that fulfils all the elements included in the ORSD written in step one. The intervention of domain experts can help to expand and refine the baseline ontology, which is then transformed into a formal language to give the final ontology. The third step is dedicated to evaluating the ontology, first by verifying that it adheres to the ORSD. Then, it should be implemented in a prototype software environment and tested by users. The fourth and final step in OTK is the maintenance phase. It requires defining rules that dictate how, at what frequency and by whom the ontology will be updated, tested and newly released.

4.2.3.4 DOGMA

Inspired by databases, DOGMA (Jarrar and Meersman 2008; Jarrar and Meersman 2002) is an approach for developing formal domain ontologies that focuses on usability, shareability and reusability. The specificity of this approach lies in its division of domain axiomatisations and application axiomatisations, which results in application-independent ontologies. DOGMA does not give a series of steps to follow or a list of tasks to carry out but rather describes a specific ontology architecture. To be able to implement the DOGMA approach, one must therefore already be knowledgeable in the development of ontologies from scratch. The DOGMA ontology architecture divides the ontology in two parts: the *ontology base* and the *commitment layer*. The ontology base is a stable core for the domain ontology, in which the concepts and axioms are suited to all possible use for the ontology. A commitment layer, of which there may be many, represents the axioms developed for one specific application of the ontology. Each application of the ontology must reuse at least part of the ontology base. Some parts of the commitment layers may also be shared if the applications are similar.

4.2.3.5 DOGMA-MESS

DOGMA-MESS (Moor et al. 2006) is an adaptation of the DOGMA approach that is dedicated to the development of interorganisational ontologies. DOGMA-MESS prescribes the use of a permanent domain- and application-independent core, which Moor et al. call the *meta-ontology*. The rest of the model is divided into layers that have varying access

and modifiability, which favours interorganisational use. Like DOGMA, DOGMA-MESS is an ontology development approach that describes a specific ontology architecture rather than a methodology that can be followed step-by-step.

4.2.3.6 DILIGENT

DILIGENT (Vrandečić et al. 2005) is a methodology for the development of domain ontologies that positions itself as an extension of METHONTOLOGY and OTK. It distinguishes itself from these two methodologies by focusing on the use and the evolution of the ontology, and caters for the possibility of automating part of the evolution process. DILIGENT divides the ontology development process into five activities: *build*, *local adaptation*, *analysis*, *revision* and *local update*. The first activity involves the building of a core ontology as a collaborative process between domain experts, users, knowledge engineers and ontology engineers. The core ontology should not be a complete representation of the domain but rather a representation of the fundamental unequivocal aspects. The second activity requires users to work with the core ontology and adapt a local version of it according to their needs. During the third activity a *control board*, made up of selected domain experts, users, knowledge engineers and ontology engineers, analyse the changes made locally by all users during the previous activity and choose which to integrate into the core ontology. The fourth activity involves the control board regularly revising the core ontology so that local ontologies do not deviate too far from it. The fifth and final activity requires users to update their local ontologies to the new version of the core ontology. Like DOGMA-MESS, DILIGENT is oriented towards the development of ontologies that are used simultaneously by multiple people or groups with different needs. DILIGENT is composed of five activities that detail an approach to the life cycle management of a shared ontology. In order to be able to implement DILIGENT, one must therefore already be familiar with the fundamental aspects of ontology development like for DOGMA and DOGMA-MESS. Both methodologies are therefore based on the concept of a stable core ontology that can be personalised by different users in parallel.

4.2.3.7 HCOME

HCOME (Kotis and Vouros 2006) is for developing and evaluating ontologies. It divides the life cycle of an ontology into three phases: specification, conceptualisation and exploitation. One or more objectives are assigned to each phase of the ontology life cycle. HCOME requires that the

ontology developers work independently of one another or in collaboration with each other, depending on the phase. The first phase involves working together to produce specification documents for the ontology and agreeing upon its aim and scope. During the second phase, members of the project work individually using the approach of their choice to develop an ontology according to the specification. The third and final phase requires the members of the project to share and test each others' personal ontologies before evaluating them conversationally. Globally agreed-upon aspects of the individual ontologies are merged to create a final shared ontology. Like DOGMA, DOGMA-MESS and DILIGENT, HCOME does not provide specific details on how to approach the development of a suitable ontological model and therefore can only be used by those who already understand the key steps in this process. Instead, HCOME gives an overview of how the collaborative development of an ontology can be organised and managed.

4.2.3.8 NeOn

NeOn (Suárez-Figueroa et al. 2012) is a methodology for ontology engineering that offers a number of different routes to follow to develop an ontology, rather than a one-size-fits-all approach. The authors affirm the importance of reusing ontologies, parts of ontologies, ontology design patterns and non-ontological resources, hence the concept of *networked* ontologies. NeOn has been developed specifically to aid in the collaborative construction of ontological networks and is composed of four elements:

1. A glossary of the processes and activities involved in the construction of ontological networks.
2. Nine scenarios that describe how to build ontologies and ontology networks using the processes and activities included in the glossary.
3. Two ontological network life cycle models that describe how the processes and activities can be organised into phases.
4. Methodological guidelines for the processes and activities.

The first of the nine scenarios describes the key processes and activities to build an ontology from scratch, without reusing an existing knowledge resource as a base. It begins with the definition of the requirements that the final ontology should fulfil, which involves writing an ORSD and a list of competency questions. Then, Suárez-Figueroa et al. recommend using terms included in the ORSD to search for relevant knowledge resources such as ontologies, non-ontological resources and ontology design patterns that could aid the development process. Using the results of the search and the ORSD, the ontology network life cycle should be defined including

an evaluation of the necessary human resources for the project. The final three activities consist of the organisation of knowledge as a conceptual model, the formalisation of this model for example using description logic or rules, and the implementation of the model in a formal representation language. The other eight scenarios described by NeOn are all based on the first and each one includes additional processes and activities. Examples of such processes and activities are the reuse or re-engineering of ontological or non-ontological resources, the reuse of ontology design patterns, and ontology alignment. Scenarios can be combined to create a more fitting scenario that includes all the necessary processes and activities for an ontology network development project, but any combination must include scenario one. Suárez-Figueroa et al. indicate that every ontology network development scenario should also include the following activities: knowledge acquisition, documentation, configuration management, evaluation and assessment.

4.2.3.9 SAMOD

SAMOD (Peroni 2016a; Peroni 2016b) is a domain ontology development methodology that is suitable for use by those with little knowledge of semantic Web technologies. It is an iterative methodology that includes the step-by-step production of documentation. At the end of each iteration, a preliminary version of the final ontology is published. SAMOD requires the participation of two types of people: domain experts and ontology engineers. The domain experts should be capable of describing in detail the domain of the ontology in natural language. The ontology engineers should be capable of constructing an ontology in a formal language based on the informal descriptions provided by the domain experts. SAMOD has three main steps, which are to be carried out and repeated in an iterative fashion until a complete and satisfactory ontology has been produced. In the first step, a subpart of the domain is identified and isolated. The first step is initially carried out only on this subpart of the domain to represent it as an ontological model, called a *modelet* by Peroni. It requires the domain experts and ontology engineers to work together to produce three documents:

1. A motivating scenario: composed of a name to describe the subpart of the domain, a natural language description of it and representative examples.
2. A list of informal competency questions: composed of natural language questions that address the requirements of the subpart of the domain, and their corresponding answers.

3. A glossary of terms: composed of terms used in the former two documents along with their definitions.

The ontology engineers then work alone to produce a modelet in OWL according to the three documents produced with the domain experts. Peroni suggests using a middle-out approach, reusing ontology design patterns and choosing interpretable class and property names that adhere to naming conventions. Once the modelet has been created, the ontology engineers run a *model test* on it by verifying its consistency, for example by using a reasoner, and by verifying its content according to the motivating scenario. If the modelet passes the model test, the ontology engineers proceed to the *data test*. This requires creating an *exemplar dataset* by implementing the examples in the motivating scenario. If it is possible to do so, the data test is passed. Finally, a *query test* is carried out by mapping the informal competency questions to SPARQL queries and running them on the exemplar dataset. The query test is passed if the queries return the expected answered, as indicated in the list of informal competency questions. If any of the tests fail, the ontology engineers must return to the modelling process and then repeat the three tests. If all of the tests are passed, the same process is carried out on the next subpart of the domain to create a second modelet. The second step is then to merge the second modelet with the first to create the *current model*. Step three involves refactoring the current model by integrating elements from existing ontologies, adding annotations to the model and creating property and class restrictions and axioms. Once step three has been completed for the first time, the process is continued iteratively. Other subparts of the domain are selected and undergo all three steps until it has all been modelled.

4.2.3.10 MOMo

MOMo (Shimizu et al. 2022) is another domain ontology development methodology that has been created to promote and facilitate the reuse or adaptation of existing ontologies to other projects. It focuses on modular development and the reuse of ontology design patterns. According to Shimizu et al., it is often simplest to create a brand new ontology rather than trying to reuse or adapt an existing one. They note that there are four main reasons that explain why reusing or adapting existing ontologies is a difficult task.

1. Differences between the desired granularity and that of existing ontologies.
2. Lack of conceptual clarity in existing ontologies.

3. Lack of adherence to good modelling practices in existing ontologies.
4. Lack of assistance in the reuse or adaptation of existing ontologies.

MOMo is based on the eXtreme Design Methodology (Blomqvist et al. 2016) and adds in particular the use of graphical schema diagrams. To support this way of working, Shimizu et al. present Comprehensive Modular Ontology IDE (CoModIDE), an integrated development environment that has been produced to facilitate the implementation of MOMo.

MOMo is a workflow composed of 10 steps, although Shimizu et al. note that some steps may be carried out simultaneously and that the results of some steps may require making modifications to previous steps. The first step is to write a short description of the use cases for the ontology as well as a list of possible data sources. Step two involves writing competency questions, which can help to refine the use cases description and the list of sources. The third step involves determining the key notions present in the domain to be modelled, a task which can be aided by the outputs of the two first steps. Each notion will eventually become a module in the ontology. During the fourth step, pattern libraries should be used to identify templates for each module. Shimizu et al. recommend using a well-curated library of ontology design patterns such as Modular Ontology Design Library (MODL) rather than using crowd-sourced collections such as those available at `ontologydesignpatterns.org`. Step five involves creating schema diagrams, which Shimizu et al. (2022) define as "labeled graphs that indicate OWL entities and their (possible) relationships", for all the models that will be part of the ontology, using the patterns identified during the previous step where possible. The sixth step requires writing documentation and defining axioms for each module. The documentation should include a schema diagram, the formal OWL axioms, alternative formal representations of the axioms using description logic syntax, for example, and finally natural language representations of the axioms. Step seven requires combining the schema diagrams from each module into one ontology schema diagram, and step eight involves adding axioms that cover more than one module. For step nine, Shimizu et al. give advice on naming classes and properties in a way that makes the ontology more user-friendly and therefore easier to reuse. The tenth and final step requires producing the formal model in the form of an OWL file, for which it is recommended to use CoModIDE.

4.2.3.11 ACIMOV

The Agile and Continuous Integration for Modular Ontologies and Vocabularies (ACIMOV) methodology is an ontology and vocabulary develop-

	METHONTOLOGY	Ontology Development 101	OTK	DOGMA	DOGMA-MESS	DILIGENT	HCOME	NeOn (Scenario 1)	SAMOD	MOMo
Writing ORSD	1st	1st	1st				1st	1st	1st	1st
Writing competency questions	1st	1st	1st					1st	1st	2nd
Knowledge acquisiton	2nd		1st				2nd	2nd	1st	1st
Studying existing ontologies	2nd	2nd	1st				2nd	2nd	3rd	1st
Conceptualisation, defining concepts and relations	3rd	3rd	1st				2nd	4th		3rd, 5th
Formalisation, defining a semi-computable model								5th		6th
Identifying ontology design patterns								6th	1st	4th
Creating diagrams									1st	5th, 7th
Integrating knowledge from external sources	4th	2nd	1st				2nd		3rd	
Implementing concepts as classes in formal language	5th	4th	2nd			1st	2nd			10th
Implementing relations as properties in formal language	5th	5th	2nd			1st	2nd			10th
Creating restrictions and axioms		6th							3rd	6th, 8th
Merging modules or different versions						3rd, 5th	3rd		2nd	7th
Naming classes and properties appropriately			1st						1st	9th
Testing or putting ontology into practice			3rd			2nd	3rd		1st, 2nd	
Evaluation	6th		3rd				3rd			
Populating ontology		7th								
Maintenance			4th			4th				

Table 4.2.1 – Summary of the steps involved, and the order in which they are carried out, in the 10 ontology development methodologies under review.

	METHONTOLOGY	Ontology Development 101	OTK	DOGMA	DOGMA-MESS	DILIGENT	HCOME	NeOn	SAMOD	MOMo
Number of steps	6	7	4			5	3	6	3	10
Modular						✓	✓	✓	✓	✓
Iterative		✓				✓	✓		✓	
Multi-actor					✓	✓	✓			
Domain experts	2nd		1st, 2nd			1st, 2nd, 3rd, 4th	2nd		1st	1st, 2nd, 3rd, 5th
Feasibility study			0th					3rd		

Table 4.2.2 – Summary of the characteristics, and the step number or numbers in which they are carried out if applicable, of the 10 ontology development methodologies under review.

ment methodology (Hannou et al. 2023). It possesses many of the characteristics for which we were searching in an ontology development methodology: it is a modular methodology that encourages collaboration with domain experts, but unfortunately it was published after we had completed this part of the project.

4.2.4 Summary

According to the general classification of ontology levels as described by Biemann (2005) we aim to produce a domain ontology, whilst according to the classification defined by Guarino (1998) we are seeking to produce an application ontology. We will use OWL, a description logic-based language, because we do not need to formally capture the implications of very complex axioms but we need for it to be able to be used by computers for logical reasoning.

As can be seen in table 4.2.1, most of the methodologies that we studied in section 4.2.3 begin with a similar phase of knowledge acquisition, documentation writing or the writing of competency questions. The moment at which concepts from existing ontologies or other resources are integrated differs between the methodologies. It is done at the beginning of the development process in Ontology Development 101 but is the last step of SAMOD. Specific instructions for developing modular ontologies are given in NeOn, SAMOD and MOMo. OTK, NeOn and SAMOD give

instructions for evaluating the ontologies developed with their methodology. METHONTOLOGY recommends an evaluation guide published elsewhere (Gómez-Pérez et al. 1995) and HCOME underlines the need to carry out an evaluation of the ontology without offering a strategy for doing so. Evaluation is not mentioned in Ontology Development 101, DOGMA, DOGMA-MESS nor in MOMo. As shown in table 4.2.2, DOGMA-MESS, DILIGENT and HCOME are oriented towards interorganisational or multi-actor applications because they allow or require the ontology to be developed in parallel by the different parties involved. METHONTOLOGY and OTK are very similar: both begin with knowledge acquisition and specification writing before moving on to informal domain modelling, which is then transformed into a formal language. Both give instructions to evaluate the final ontology. The steps in SAMOD and MOMo are also very similar. They are based on modular development during which the ontology is constructed little by little, either by adding a new part of the domain to the model during each iteration (SAMOD) or by modelling each part of the domain individually first and then merging them (MOMo). NeOn can be distinguished from the other methodologies by the fact that it offers many approaches to developing an ontology or a network of ontologies. It requires of the ontology engineers to carry out a detailed analysis of the project before starting in order to be able to choose the best combination of processes and activities.

Out of the 10 ontology development methodologies, the one most suited to our needs within the context of ATONTE is SAMOD (Peroni 2016a) for three main reasons. Firstly, SAMOD requires establishing a solid core for the ontology that is extended iteratively as much as is required. This fits well with our need to have a seed ontology that can be enriched automatically. Secondly, SAMOD has a high level of involvement by domain experts, which we have already established as being necessary. Thirdly, SAMOD includes tests to evaluate the final ontology. Other methodologies that contain useful elements are MOMo (Shimizu et al. 2022), in particular the modelling phase, and NeOn (Suárez-Figueroa et al. 2012), in particular the refactoring phase.

However, no single ontology development methodology fully fits the needs of ATONTE. In section 4.3 we present the methodology that we created, taking inspiration from SAMOD, MOMo and NeOn. We apply our methodology in section 4.4 to our corpus of *Instructions nautiques* with the help of domain experts.

4.3 Our Methodology

4.3.1 Overview

Our methodology for the development of domain ontologies from text and experts is composed of four main steps, the latter two of which are to be carried out in an iterative fashion. It also includes a preliminary step to be carried out before starting to implement the methodology, which is designed to help with verifying that it is the right ontology development methodology to be used for a given project. Figure 4.3.1 shows a flowchart that represents the main steps involved in our methodology. An overview the main tasks involved in each step is given below, and then each step is presented in detail in sections 4.3.2 to 4.3.6:

0. Feasibility Study
    - Verifying methodology is suited
1. Groundwork
    - Familiarisation with the corpus
    - Identifying and analysing sources of domain knowledge
    - Identifying domain experts
    - Defining ontology application
    - Creating preliminary dataset of semantic triples
    - Dividing domain into subdomains
2. Producing Documentation
    - Motivating scenario
    - Informal competency questions
    - Glossary
3. Structuring, Implementing and Testing Subdomain Models
    - Conceptualising subdomain models
    - Implementing subdomain models in OWL
    - Creating subdomain datasets using Resource Description Framework (RDF)
    - Iterations of subdomain models and datasets
    - Testing subdomain models and datasets
4. Merging, Refactoring and Aligning
    - Merging subdomain models to create full model
    - Merging subdomain datasets to create exemplar dataset
    - Refactoring and aligning full model
    - Testing full model and exemplar dataset

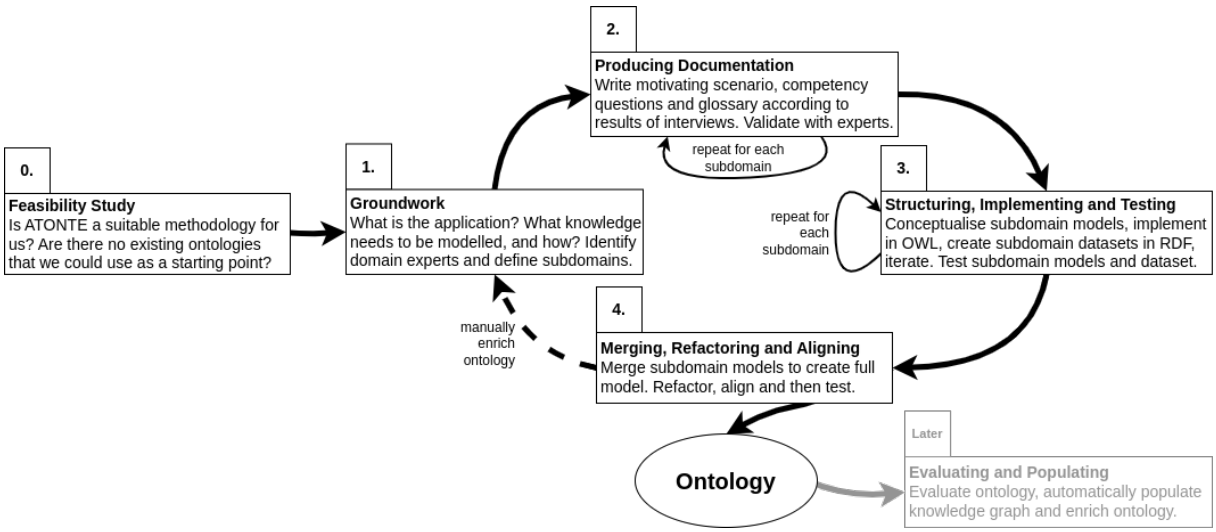


Figure 4.3.1 – Flowchart showing the steps involved in our methodology and how they are to be carried out (Rawsthorne et al. 2022b).

4.3.2 Step 0: Feasibility Study

Before applying our methodology, we recommend carrying out a lightweight feasibility study to verify that it is suited to the application. Each of the following requirements should be satisfied for this to be a good choice of ontology development methodology:

— The aim of the ontology is to model all or part of the knowledge contained within a text corpus, combined with the knowledge of domain experts, with a final application in mind

— The end users of the ontology application are known and can be consulted, along with domain experts (who can be the same people), on an ad-hoc basis during the development process

— No other ontologies that could be used for the final application already exist and are published on the Web

4.3.3 Step 1: Groundwork

The first task is to study and become familiar with the content of the text corpus that contains the knowledge to be modelled.

Whilst carrying out the first task, other sources of domain knowledge such as existing semantic resources, reference documents or specifications should be identified and analysed. These will show how authoritative bodies interpret the domain and organise the knowledge within it, and could serve as templates or inspiration. They may also provide definitions of concepts or relationships. This exercise helps us to gain a better understanding of the domain and how it is organised, leading to a more reliable model.

Another important source of domain knowledge can be interactions with different types of domain experts. They should be identified and invited to participate as early as possible, to be able to plan the development process according to their availability. As demonstrated by Mansfield et al. (2021), the domain experts can be users or producers of the corpus. Consultations with domain experts can take the form of meetings, interviews or presentations with feedback sessions, or a combination of the three. Regardless of the form of the interaction, it is important to present the ontology development project simply and pedagogically, by using easy-to-understand domain-based examples. The more the domain experts understand about the reasons the ontology is being built and what its construction entails, the more likely it is that they will be able to give useful advice and information. Different types of domain experts could be consulted in different ways and at different stages in the ontology development process. They can be called to aid during this first step to help understand the corpus or other sources of domain knowledge, but supposing that they would be participating on a voluntary basis, it can be judicious to hold back on provoking these interactions until Step 2 when concrete needs have been identified.

In parallel, define the ontology application or its potential use cases: how will the ontology be ultimately used and for what purpose or purposes? This definition will influence the form that the final model will take.

From this early stage, attempts should be made at informally structuring the knowledge contained in various relevant extracts of the corpus to be modelled. This can be done by manually decomposing the knowledge in an extract into finer-grained chunks, favouring a subject-predicate-object structure where possible, to create semantic triples. The application of the ontology should be taken into account here: how should the knowledge be modelled in order for it to satisfy the needs of the application? This process of trial and error of creating a *preliminary dataset* should be carried out with different extracts until a rough structure of the main concepts of the domain converges. It can also help us become more familiar with and increase our understanding of the corpus domain and content.

Once we have an overview of the extent of the domain, it should be divided into coherent subdomains according to the main themes that have emerged, unless the knowledge contained within the corpus spans only a very small domain. This will facilitate the documentation (Step 2) and modelling (Step 3) tasks, during which the subdomains are dealt with individually. Each subdomain will have its own documentation and will be modelled as its own small ontology, all of which will eventually be merged

together to create the final full ontology. It is likely that there will be some overlap between subdomains meaning that some classes and properties will appear in more than one subdomain model.

4.3.4 Step 2: Producing Documentation

Before the modelling can continue, the production of documentation should be initiated. The three pieces of documentation to be produced for each subdomain are identical to those produced in SAMOD (Peroni 2016a):

— Motivating scenario

— List of informal competency questions

— Glossary

However, we suggest that, thanks to the initial modelling phase during the preliminary dataset construction, we are able to work independently of domain experts to produce initial drafts of the documentation.

A *motivating scenario* should be drafted for each subdomain identified; it will be subject to modification as the model evolves. The motivating scenario should consist of:

— A name for the subdomain

— A natural language description of the theme of the subdomain

— A list of extracts that demonstrate all the ways in which the subdomain theme is represented in the text and their corresponding semantic triples

— A list of the main concepts and characteristics related to the subdomain theme

The description of the subdomain theme should highlight how and why it needs to be modelled in the context of the ontology application. Reuse the extracts identified during the groundwork step to start creating the lists of extracts. Extracts may match the theme of more than one subdomain, but should only be added to the motivating scenario of the most relevant one. Add as many extracts as are required to give a complete and balanced picture of the way in which the subdomain theme is expressed in the text. Include any semantic triples that were created from the extracts during the groundwork step. The extracts and the semantic triples can then be used as a source from which to identify the main concepts and characteristics related to the subdomain theme. If possible, categorise the characteristics as being optional or obligatory for each concept identified.

The next piece of documentation to be produced for each subdomain is a *list of informal competency questions*. These questions, also written in natural language, must demonstrate the requirements that should be satisfied by the final ontology that are expressed in the corresponding motivating scenario. A model answer should be given for each question. The information that allows the question to be answered should figure in the extracts in the motivating scenario. At this point, it can be useful to consult domain experts to validate the lists of competency questions, or to help write them as is recommended in MOMo (Shimizu et al. 2022). This can be done, for example, by conducting targeted interviews or hosting meetings with a range of domain experts. The outcomes of these interactions can help to refine not only the competency questions but also the motivating scenarios.

Finally, a *glossary* should be compiled for each subdomain, defining its specific vocabulary within the context of the project.

During this step, consult domain experts by conducting meetings, interviews or presentations with feedback sessions to find out what knowledge from the corpus needs to be modelled in the ontology, where their priorities lie, what knowledge needs to be added (if any) and how it all needs to be structured in order to be useful to them. They could also be called upon towards the end of this step to enrich and validate the documentation.

4.3.5 Step 3: Structuring, Implementing and Testing Subdomain Models

Once a first draft of every piece of documentation has been produced, the modelling process can resume. For each subdomain, analyse the semantic triples produced during the groundwork phase (section 4.3.3) and inserted into their corresponding motivating scenario (section 4.3.4). Group together the subjects/objects and the predicates that serve the same purpose. There may be classes and properties that better fit the theme of another subdomain, in which case they should be copied to their corresponding subdomain model. If the concepts had been modelled differently in the other subdomain, study both versions and adapt the model to be able to represent all occurrences of the concepts.

Implement the subjects, objects and properties belonging to each subdomain as OWL Classes, NamedIndividuals, ObjectProperties and DatatypeProperties, using the vocabulary defined in the subdomain glossary to name them when possible: this is the first iteration of the subdomain model. Rewrite the semantic triples associated with one extract formally as RDF triples using the first iteration of the subdomain model. This is the first

iteration of the set of subdomain triples, which we will call the *subdomain dataset*. Move on to the next extract in the subdomain motivating scenario and try to structure its content by creating RDF triples using the newly-created classes and properties. If this task cannot be performed satisfactorily, modify the subdomain model accordingly. Continue in an iterative fashion with all other extracts in the subdomain motivating scenario until the subdomain model has stabilised and you have a solid set of triples specific to the subdomain. It may be necessary to structure additional unseen extracts from the corpus before arriving at a stable subdomain model.

To the greatest possible extent, work on all subdomains in parallel. This makes it possible to have a constant and complete view of the domain and therefore to be able to work on the development of each subdomain model in a complementary, rather than independent way. It also makes it easier to keep track of, and ensure the consistency of, the inevitable overlapping elements that appear in more than one model, thereby minimising the refactoring that will be required (Conesa et al. 2011). During the modelling phase, there should be a constant back-and-forth within and between subdomain models, as well as subdomain datasets, to ensure compatibility. By the end of the modelling phase, all the semantic triples contained within the preliminary dataset should have been implemented in OWL according to at least one of the subdomain models.

During this iterative phase of manual RDF triple creation and model enrichment, a version control system (VCS) can be used to keep track of the changes made to the subdomain model. This makes it possible to easily retrieve earlier versions of a model during the development process, facilitates the subdomain model merging process and also makes it easier to integrate ulterior changes.

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 subdomain model and then read through the motivating scenario written for the subdomain model and manually check that it corresponds to its description. For the data test, verify the validity of the subdomain model by modelling unseen extracts from the text. If it is possible to model the extract according to its corresponding subdomain model, the test is passed. For the query test, translate the natural language competency questions into SPARQL queries and run them on the subdomain dataset 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 carrying out all the tests

again.

4.3.6 Step 4: Merging, Refactoring and Aligning

Now the subdomain models can be merged to create the *full model*. This can be done by exporting the .owl files of the models and manually combining them, removing all duplicate classes, properties and individuals in the process. 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 full model has been created. Merge the subdomain datasets to create the *exemplar dataset* and carry out the three tests on the full model and the exemplar dataset.

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 restrictions, axioms and inferences. The elements of the refactoring process are the same as those given in SAMOD by Peroni (2016b). A detailed description of how to reuse existing semantic knowledge resources is given in NeOn (d'Aquin 2012). After the refactoring process has been carried out, the model should undergo a final testing cycle.

4.4 Application to the *Instructions nautiques*

In this section we apply our methodology, which we described theoretically in section 4.4, to our corpus of *Instructions nautiques* with the help of domain experts.

In this section and subsequent ones, we provide examples of the documentation created during the implementation of our approach on the *Instructions nautiques* in shaded grey-coloured boxes. Code listings are displayed with syntax highlighting and line numbers on a grey background. Listing 4.4.1 shows all the prefixes used.

4.4.1 Step 0: Feasibility Study

— Is the aim of the ontology to model all or part of the knowledge contained within a text corpus, combined with the knowledge of domain experts, with a final application in mind?

✓ Yes: the aim of our ontology is to model the knowledge contained within our corpus of the *Instructions nautiques*, which are de-

```
# Established
@prefix geof: <http://www.opengis.net/def/function/geosparql/> .
@prefix geom: <http://data.ign.fr/def/geometrie#> .
@prefix gsp: <http://www.opengis.net/ont/geosparql#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# Ours
@prefix atln: <http://data.shom.fr/def/atlantis#> . # base IRI for the ontology
@prefix ent: <http://data.shom.fr/id/spatialentity/> . # for named individuals
    that are spatial entities
@prefix inp: <http://data.shom.fr/id/inparagraph/> . # for named individuals
    that are paragraphs from the Instructions nautiques
@prefix tco: <http://data.shom.fr/id/codes/atln/typeofcolour/> . # for named
    individuals that are a type of colour
@prefix tdi: <http://data.shom.fr/id/codes/atln/typeofdirection/> . # for
    named individuals that are a type of direction
@prefix tse: <http://data.shom.fr/id/codes/atln/typeofspatialentity/> . # for
    named individuals that are a type of spatial entity
```

Listing 4.4.1 – All prefixes used in the code listings in this chapter, written in Turtle syntax.

scribed in detail in section 2.3.1, combined with the knowledge of domain experts to orient the model towards the final application. Our ontology will ultimately be automatically enriched, meaning that we only need to create a seed ontology that represents the core structure of the domain. The final application of our ontology is to provide a model to structure a knowledge graph to represent the content of the *Instructions nautiques*. The purpose of the knowledge graph is to provide the *Service hydrographique et océanographique de la Marine* (Shom) with more efficient and connected ways of producing the *Instructions nautiques*, and to offer users of the *Instructions nautiques* more efficient and connected ways of consulting their content.

— Are the end users of the ontology application known and can they be consulted, along with domain experts (who can be the same people), on an ad-hoc basis during the development process?

  ✓ Yes: the end users of our ontology application are Shom employees and current users of the *Instructions nautiques*: we consider both to be *Instructions nautiques* domain experts. They are available to be consulted during our project.

— Do no other ontologies that could be used for the final application

already exist and are they published on the Web?

- ✓ No: we conducted a thorough search for existing maritime ontologies, the results of which are detailed in section 4.2.2. We found publications referring to unpublished ontologies but found no published ontologies that cover more than a few of the concepts required to model the content of the *Instructions nautiques*. Although we identified some useful classes, we found no suitable properties.

4.4.2 Step 1: Groundwork

4.4.2.1 Familiarisation with the Corpus

Our text corpus is composed of 15 volumes of *Instructions nautiques*. We carried out a full analysis of the content of our corpus and the way in which it is presented at the level of individual volumes and at the level of the entire series, the results of which are described in detail in section 2.3.1 from page 2.3.1 onwards.

4.4.2.2 Identifying and Analysing Sources of Domain Knowledge

Whilst analysing the *Instructions nautiques*, we also consulted documents produced by three key organisations in the world of maritime navigation: the Shom, the International Hydrographic Organization (IHO) and the International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA), to find out how the knowledge contained within the *Instructions nautiques* is referred to in other official documentation. We also wanted to find out what kind of complementary knowledge or information is provided in other documentation and therefore how the ontology needs to be constructed in order for it to be easily connected to them.

As well as the *Instructions nautiques* series, we analysed other publications by the Shom such as *Signalisation maritime* (Shom 2016), which gives the main maritime signalling rules and communication signals, *Symboles, abréviations et termes utilisés sur les cartes marines papier* (Shom 2019b), which explains the symbols, abbreviations and terms used on paper nautical charts, and a document that describes the *Balisage maritime* database (Shom 2019a), which is a database of navigation marks [16]. The IHO publications that we used are the *S-32 IHO Hydrographic Dictionary*

16. A navigation mark is defined as an "artificial or natural object of easily recognisable shape or colour, or both, situated in such a POSITION that it may be identified on a CHART or related to a known navigational instruction. Alternative term for visual AID TO NAVIGATION. Includes both BUOYs and BEACONs (fixed artificial navigation mark)." (Hydrographic Dictionary Working Group 2019).

(Hydrographic Dictionary Working Group 2019), the authoritative reference for multilingual hydrographic related terms and definitions used in IHO publications, and the *S-57 IHO Object Catalogue* (International Hydrographic Organization 2000), object catalogue of the IHO transfer standard for digital hydrographic data. Finally, we studied the *NAVGUIDE* (International Association of Marine Aids to Navigation and Lighthouse Authorities 2018), the IALA marine aids to navigation manual.

4.4.2.3 Identifying Domain Experts

We identified two types of domain experts of the *Instructions nautiques*: those who produce them and those who use them. The two groups interact very differently with the *Instructions nautiques*. On the one hand, the producers understand the entire editorial process of the *Instructions nautiques* and have a global vision of their content and organisation. They write and edit the *Instructions nautiques* to be in accordance with cartographic representations and other nautical publications. On the other hand, the users are navigators accustomed to consulting the *Instructions nautiques*, alongside nautical charts and other resources, to efficiently find the information they need during itinerary planning.

4.4.2.4 Defining Ontology Application

The purpose of our ontology is to provide a model to structure a geospatial knowledge graph to represent the content of the *Instructions nautiques*. We primarily aim to represent spatial entities, their locations via direct and indirect spatial referencing, their characteristics and the relations between them, as well as maritime navigation guidelines and instructions for entering ports, which rely heavily on the descriptions of spatial entities. A more detailed description of the ontology and knowledge graph application is given in section 2.4 from page 19 onwards.

4.4.2.5 Creating Preliminary Dataset of Semantic Triples

We selected a range of extracts from the text of the *Instructions nautiques*, aiming to cover the entire scope of the corpus. For example, extract 4.4.1 is dense with references to spatial entities and their characteristics. It also includes references to typical oceanographic conditions and gives navigation guidelines for a certain type of vessel. The configuration of the spatial entities mentioned in this extract are shown in the raster navigational chart (RNC) in figure 4.4.1. Document 4.4.1 shows how we decomposed the knowledge contained within extract 4.4.1 into semantic

"30 M east of Makemo, Nihiru (16° 42' S — 142° 50' W), an atoll of 5 M in diameter, is well wooded, except in the south. On the motu of the SW point, which shelters the village of Tatake, the old lighthouse (15 m) is partially hidden by vegetation (visible at 3 M). To the south of the atoll, the current flows NE and can create violent eddies near the SE point.
CHANNEL. — A channel for small boats, oriented at 116°, 100 m long, 4 m wide and 1.50 m deep, has been dug to the NW of the motu. Posts mark it. A landing point has been created there."

Extract 4.4.1 – Translated from the original French text: "À 30 M à l'Est de Makemo, Nihiru (16° 42' S — 142° 50' W), atoll de 5 M de diamètre, est bien boisé, sauf au Sud. Sur le motu de la pointe SW, abritant le village de Tatake, l'ancien phare (15 m) est partiellement masqué par la végétation (visible à 3 M). Au Sud de l'atoll, le courant porte au NE et peut produire de violents remous près de la pointe SE. PASSE. — Une passe pour embarcations, orientée à 116°, de 100 m de long, 4 m de large et 1,50 m de profondeur, a été creusée au NW du motu. Des piquets la balisent. Un point de débarquement y est aménagé." (Shom 2021f, p. 143)

triples.When it was not clear how to model a given piece of knowledge, we paused and searched for other extracts that covered the same subject to get more insights into how the concept is dealt with in our corpus. We frequently consulted the related domain resources cited in section 4.4.2.2 to know, in particular, how they categorise the spatial entities that are specific to the maritime environment and heavily relied-upon for navigation purposes.

4.4.2.6 Dividing Domain into Subdomains

Having closely studied the domain of the *Instructions nautiques*, we divided it into four subdomains:

— Spatial entities, their types and properties, and the spatial relations between them

— Navigation instructions, guidelines, rules and regulations that indicate where and when is navigable or not and by what type of vessel or craft

— Vessels and crafts that can be used for coastal navigation, their types and their properties

— Temporalities, and meteorological and oceanographic phenomena, their types and their properties

4.4.3 Step 2: Producing Documentation

4.4.3.1 Interviews with Users of the *Instructions nautiques*

To better understand how the *Instructions nautiques* are used, we conducted a series of semi-structured interviews with some of their users. We

Extract, split into sentences

1. 30 M east of Makemo, Nihiru (16° 42' S — 142° 50' W), an atoll of 5 M in diameter, is well wooded, except in the south.
2. On the motu of the SW point, sheltering the village of Tatake, the old lighthouse (15 m) is partially hidden by vegetation (visible at 3 M).
3. To the south of the atoll, the current flows NE and can create violent eddies near the SE point.
4. CHANNEL. — A channel for small boats, oriented at 116°, 100 m long, 4 m wide and 1.50 m deep, has been dug to the NW of the motu.
5. Posts mark it.
6. A landing point has been created there.

Decomposition into triples, per sentence

1. spatial entity 1 - is a - spatial entity / spatial entity 1 - is called - Makemo / atoll 1 - is a - atoll / atoll 1 - is called - Nihiru / atoll 1 - is 30 M from - spatial entity 1 / atoll 1 - is east of - spatial entity 1 / atoll 1 - has geographic coordinates - 16° 42' S, 142° 50' W / atoll 1 - has diameter - 5 M / atoll 1 - is - wooded / atoll 1 - has region - spatial entity 2 / spatial entity 2 - is a - spatial entity / spatial entity 2 - is called - south Nihiru / spatial entity 2 - is not - wooded
2. motu 1 - is a - motu / atoll 1 - has region - point 1 / point 1 - is a - point / point 1 - is called - SW point / motu 1 - is part of - point 1 / village 1 - is a - village / village 1 - is called - Tatake / motu 1 - shelters - village 1 / lighthouse 1 - is a - lighthouse / lighthouse 1 - has height - 15m / lighthouse 1 - has visibility - partial / lighthouse 1 - is visible at - 3 M
3. current 1 - is a - current / current 1 - is south of - atoll 1 / current 1 - flows - northeast / spatial entity 3 - is a - spatial entity / atoll 1 - has region - point 2 / point 2 - is a - point / point 2 - is called - SE point / spatial entity 3 - is near - point 2 / current 1 - is dangerous at - spatial entity 3
4. channel 1 - is a - channel / channel 1 - is for - small boats / channel 1 - has orientation - 116° / channel 1 - has length - 100 m / channel 1 - has width - 4 m / channel 1 - has depth - 1.5 m / channel 1 - is northwest of / motu 1
5. posts 1 - is a - set of posts / channel 1 - is marked by - posts 1
6. landing point 1 - is a - landing point / landing point 1 - is located in - channel 1

Document 4.4.1 – The decomposition into semantic triples of the knowledge contained within extract 4.4.1.

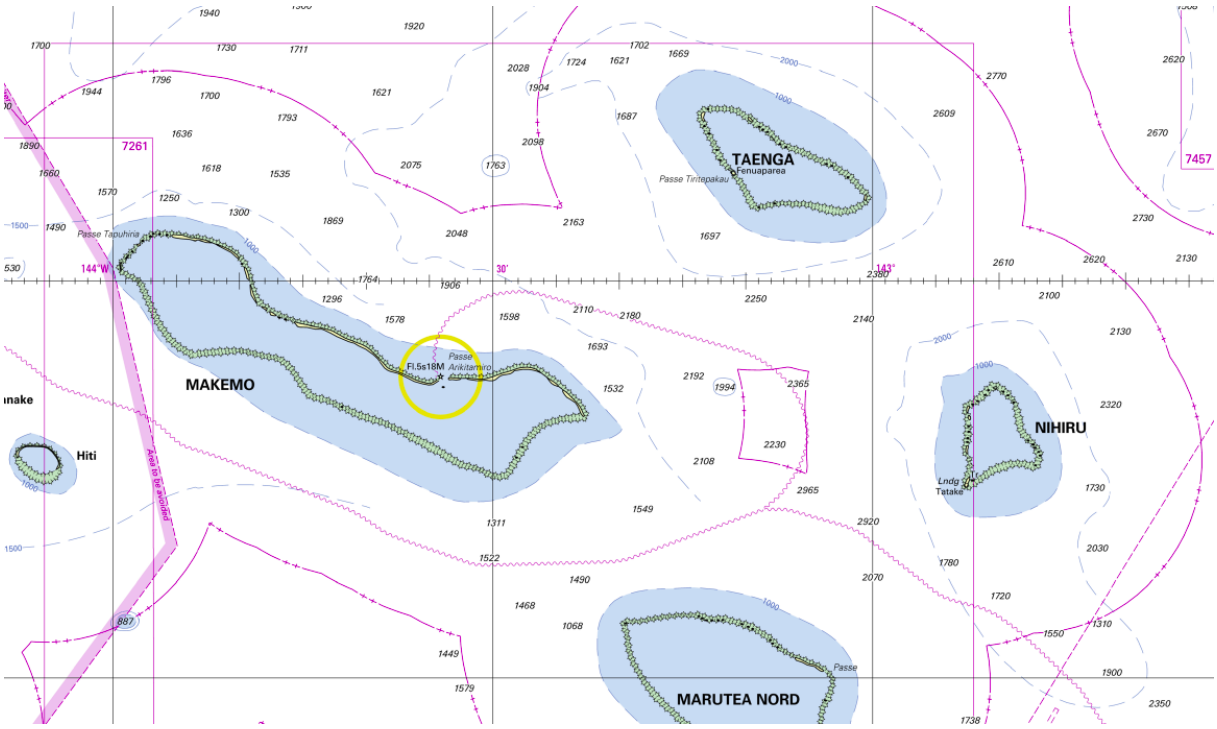


Figure 4.4.1 – Extract of the *Assemblage des cartes marines (RasterMarine)* RNC product published by the Shom showing the relative positions of the Makemo and Nihiru atolls in the Tuamotu archipelago, French Polynesia (Shom 2023).

held the interviews over a period of weeks, during which time we worked in parallel on the documentation. The results of the first interviews helped us to define appropriate informal competency questions and refine the content of the motivating scenarios. During the later interviews, we were able to have our documentation validated and approved by domain experts.

In total we interviewed 15 people with different levels of experience of using the *Instructions nautiques* in their studies or for their work, in the military or civilian domain. We held 10 interviews of 30 to 60 minutes long with individuals or small groups, in person where possible and otherwise virtually. We spoke with five students and four military instructors at the *École navale*, the French Naval Academy, and three civilian instructors from the *École nationale supérieure maritime* (ENSM), the French National Maritime Academy where merchant navy officers are trained. The questionnaire that we wrote to help direct the interviews can be found in appendix A on page 169. At the beginning of each interview we presented the objective of our project as extracting, organising and storing the information contained within the *Instructions nautiques* in a way that makes it possible to use the information differently, in a format unlike their current one. We gave the example of being able to offer quicker and more efficient ways of accessing the information contained within the *Instructions nautiques* that would not require taking the time to read the full text. We explained that the aim of the interviews was to analyse the needs of current

users of the *Instructions nautiques* to be able to direct our work towards appropriate solutions.

Because of the open-ended questions in the questionnaire, the interviews often tended towards unstructured discussions. The order of the questions in the questionnaire was not always respected because some questions were answered by our interviewees before we had asked them. We received a large quantity of comments, criticisms and suggestions in a disorderly fashion. We therefore decided to present the results of the interviews by theme rather than question-by-question.

Also due to the open-ended format of the questionnaire, we cannot assume completeness in the responses of our interviewees. For example, only five people out of 12 cited the use of nautical charts along side the *Instructions nautiques*, however we can suppose that in reality they are heavily used by all of our interviewees. We must therefore bear in mind that the absence of a particular statement does not mean that it is not associated with certain uses of the *Instructions nautiques* or with the reality of the *Instructions nautiques*. This applies to all of the questions in the questionnaire.

4.4.3.1.1 General Needs Some of our interviewees talked about their general needs concerning the *Instructions nautiques*. Three out of the 12 people interviewed explained that they need access to precise and well-structured information whilst at sea and two people insisted on the fact that they appreciate being able to read full text whilst on land. One person noted their need to be able to work with images, as having only the text is difficult. Another person indicated that the *Instructions nautiques* serve to answer the following question: "what can I expect to see?". Finally, one person said that the *Instructions nautiques* should be "something practical, convenient and reliable".

4.4.3.1.2 How the *Instructions nautiques* are Used During the interviews, we asked each person to describe how they use the *Instructions nautiques*: in which situations and in what manner. The most common answer was the following: when we already know the region where we're going, the *Instructions nautiques* serve no purpose; we learn off by heart the general information about a region very quickly. This comment was made by eight out of our 12 interviewees. Seven people said that the *Instructions nautiques* are mainly used to prepare a mission or a watch [17] in advance. Half of our interviewees explained that the way that they use the *Instructions*

17. A watch is a division of time during which a vessel operator is on shift.

nautiques changes depending on whether they are on land or at sea. At sea, the *Instructions nautiques* are less used: they are only used if there is a doubt about something, or if the route or arrival port has to be changed at the last minute. Five of our interviewees noted that the *Instructions nautiques* are not used to search for a specific piece of information but rather to find out what is written about a given place, to discover a new zone or to better understand an environment. As well these four frequent remarks, we identified five other comments each made by more than one person. First of all, the *Instructions nautiques* are important to be able to make a link between the nautical chart and what can be seen in reality. Second, it is necessary to read all of the *Instructions nautiques* at least once. Third, the *Instructions nautiques* are only used by leisure navigators in difficult areas. Fourth, the *Instructions nautiques* are mainly used to help fill out a *fiche de traversée*, a 10-15 page document that must be prepared ahead of each mission that contains the departure and arrival times, routes, the weather forecast, the stations to be contacted, the traffic separation scheme (TSS) or straights to be navigated, pilotage instructions, danger zones and telephone numbers.

Nine other remarks were made, each by one interviewee, which shows the diversity in the use of the *Instructions nautiques*. One person said that they consider the *Instructions nautiques* to be "like a travel guide, to have a description of the country and to have information such as the currency used". One person mentioned that they use the paper version of the *Instructions nautiques* on land and a PDF version on board a vessel. One person specified that the *Instructions nautiques* are never used without a nautical chart. One person said that they use the *Instructions nautiques* to search for information to complement the information that they find elsewhere, meaning that the *Instructions nautiques* come in second place to another source. One person indicated that they do not open a volume of the *Instructions nautiques* to search for a piece of information such as a telephone number if they could search for it on the Web using their smartphone. One person mentioned that they use the *Instructions nautiques* to have information in French. One person explained that the *Instructions nautiques* are used during the preparation of a mission before having made the request for nautical charts. This is because making a request for nautical charts is expensive in time and in money, so it is important to first check, with the help of the *Instructions nautiques*, that the vessel will really be able to navigate in the targeted zone. One person told us that a mission is always planed from quay to quay, even if it is obligatory to use the services of a pilot to enter the destination port. They

explained that it is useful to already have an idea of the manoeuvres that should be carried out by the pilot in case there are any disagreements or misunderstandings, for example because of a language barrier, during the operation. One person mentioned that their students draw diagrams using the information given in the *Instructions nautiques*. Finally, one person indicated that they consult the *Instructions nautiques* to fill in a *passage planning*, a regulatory document that must be filled out for each mission. This document covers the whole itinerary from quay to quay and must include information such as passage points and landmarks.

4.4.3.1.3 Information Searched For in the *Instructions nautiques*

We asked our interviewees what they search for in the *Instructions nautiques* in priority. In their responses, we identified three clear categories:

— general information

— ports

— coastal navigation

General Information All of our interviewees mentioned that they search for at least one piece of general information in the *Instructions nautiques*. Within this category, three quarters of people replied that they search for information about the communications that must be made with semaphores or stations: who should be called and when, the telephone numbers or the radio channel frequencies. Seven out of 12 people cited photos, in particular photos of landmarks, as something that they search for. One person added that photos are more useful than the textual descriptions full of abbreviations that can be found in the *Feux et signaux de brume* publications. Two people mentioned that they search for general information or generalisations, and the same number of people noted that they search for particular information about the weather. The following elements were mentioned each by one person amongst the 12 interviewed: special cases (such as exceptions to rules), speed limits and general reminders (such as where a boundary will be crossed).

Ports The second most frequently cited category that we identified concerns ports. Out of the 12 people interviewed, 11 said that they search for information about ports, of which nine search for information about landings, docking, port approaches and port entries. Five out of the 12 interviewees search for information about pilotages: either instructions or information about the vessel used for the pilotage. Three people mentioned that they search for the administrative procedures to follow before arriving

in a port. The other port-related information that is searched for includes: the entry times of a port (two people), the positioning of quays or berths in a port (two people), the height of the quays in a port (one person) and the dimensions of locks (one person).

Coastal Navigation After general information and ports, the third most mentioned category is coastal navigation. Out of the 12 people interviewed, 10 indicated that they search for information about navigation outside of ports. Within this category, leading lines are the most searched-for: half of the interviewees cited them as something that they look for. Information about landmarks are searched for by five people out of 12. Four people search for the characteristics of lights or night lights, and the same number search for navigation guidelines or recommended routes in regions with which they are not yet familiar. Three people cited bearings as something that they search for and two people search for qualitative information about currents. A few other elements were each cited once: straights, specific information about the oceanography of a region, navigation in narrow channels, soundings and draughts.

4.4.3.1.4 Comments on the Current *Instructions nautiques*

During the interviews, we asked our interviewees for their opinion on the current *Instructions nautiques*, whether they pose any problems or whether there are any cases in which they do not suffice. We have divided the comments that we received on this subject into five categories:

— the text

— the photos

— the up-to-dateness of the content

— the organisation or format of the content

— consultation practices

Textual Content Half of the people that we interviewed indicated that there is too much information in the *Instructions nautiques*, that the information is difficult to digest and that the large sections of text are impractical for finding information. However, four out of 12 people said that they find the *Instructions nautiques* to be complete. One person said that they find the *Instructions nautiques* to be written in a very precise manner, and that having precise descriptions and photos is a big advantage. This person also said that the fact that the *Instructions nautiques* are written in full sentences makes them interesting to read, and if something is not

interesting to read then the reader risks missing out on important information. One person said that the literary style is not a problem for them personally but that it can pose difficulties for others. We also received more specific comments on the content of the text in the *Instructions nautiques*. One person indicated that leading lines, which are written "in a novel-like way", are well explained but incomplete. Another person said that the maritime vocabulary is consistent, and that it should be kept so. One person said that indications such as "a square tower" are not very useful and that a photo or a more precise description of its characteristics or its location would be more useful. Finally, one person mentioned that sometimes the *Instructions nautiques* indicate the differences between nautical charts from different countries, such as a change in the geodesic system, and that this type of information is useful to have.

Apart from the comments about the lack of photos in the *Instructions nautiques*, which we discuss in the next section, we received many comments about what is lacking in the text of the *Instructions nautiques*, or elements that the people that we interviewed would like to see in the *Instructions nautiques*. Out of the 12 people interviewed, 10 made at least one remark about something that was missing or something that should be added. Four people said that the *Instructions nautiques* lack feedback from other navigators. One person added that feedback from other navigators can be useful to learn things like that there are lots of fishing nets floating in the water somewhere, or that somewhere else there is a strong current. Three people would like to see nautical charts integrated within the *Instructions nautiques*. One person explained that this need comes from the fact that the computers that are used to consult the *Instructions nautiques* are not located in the same place as the nautical chart consultation tables on a vessel. Another person attributed this need to a difficulty in placing an object shown in a photo in the *Instructions nautiques* on a nautical chart. Two people said that all the lights are not cited in the *Instructions nautiques*. Two people also indicated that there is a lack of information for noncommercial zones and uncommon destinations. One person noted that sometimes the *Instructions nautiques* do not include all the information about an area, for example buoys may be indicated on the nautical chart but not in the text, making the task of getting one's bearings in a new location more difficult.

Five of our interviewees made suggestions on what they would like to see in the *Instructions nautiques*. Concerning the addition of visualisations, one person said that they would like to see general views of pilot stations and waiting areas and another person would like to have precise diagrams

such as those found in the *Pilotes Côtiers*[18]. One person said that they would like to have a section dedicated to photos at the end of each volume of the *Instructions nautiques*. Another person suggested adding a symbol, such as a flag, next to old passages of text to distinguish them from more up-to-date sections. The addition of reminders about changes of time zones was suggested by one person. Finally, one person suggested adding an indication of how well an object or part of an object appears on a radar, to help navigators in bad weather conditions and when visibility is reduced, which renders the usual landmarks useless.

Photographs Many comments were made about the photos in the *Instructions nautiques*. On the one hand, two people out of 12 said that they think that there are already enough photos in the *Instructions nautiques*. On the other hand, three quarters of our interviewees made at least one negative comment about the photos in the *Instructions nautiques*. Seven people out of 12 said that there are not enough, of which four said that there a not enough photos of landmarks in particular. Photos of leading lines, access channels and ports were each cited as lacking by one person. One person also noted that there is a lack of close-up photos, such as of locks. Three people said that they have found that the photos are not always up-to-date, whilst two people said that the photos are not all very clear. Two people also said that the photos seem to be distributed randomly in the *Instructions nautiques* volumes.

Up-to-Dateness of Content On the subject of the up-to-dateness of the content of the current *Instructions nautiques*, half of the people that we interviewed were unsatisfied. Three people said that, in general, the volumes do not seem to be kept up-to-date. As we mentioned in the previous section, three people noted that the photos are not always up-to-date. One person indicated that the phone numbers, very high frequency (VHF) channels and port access hours are sometimes outdated.

Organisation and Formatting of Content Regarding the organisation and the formatting of the content of the *Instructions nautiques*, five out of 12 people indicated that they find it difficult to identify the correct volume to consult to access the information about a specific geographic region given the counterintuitive codes used to name the volumes (C22, D21, K11, etc.). However, three people said that it is easy to find the correct section within a volume thanks to the summary at the beginning of each volume. Three

18. The *Pilotes Côtiers* are coastal navigation guides dedicated to leisure navigators.

people also mentioned that it is easier to find the information that they are looking for in the paper editions of the *Instructions nautiques* compared to the PDF versions because they are easier to leaf through. As we have already noted, two people said that they find the photos in the *Instructions nautiques* to be distributed somewhat randomly. One person out of the 12 interviewed told us that the current format, which includes one or multiple general chapters followed by chapters that are organised geographically from port to port, is practical and easy to use. One person also suggested that the *Instructions nautiques* could be divided into smaller versions to facilitate finding the right information and to facilitate reading.

Consultation Practices Some of the people that we interviewed made comments about the ways in which they consult the *Instructions nautiques*, or the consultation practices of the students that they teach. The general opinion surrounding paper and PDF editions of the *Instructions nautiques* are divided. One person told us that the PDF editions are practical whilst another person told us that consulting the paper editions are more practical and also can be consulted in a more confidential way. Regarding the updates to the *Instructions nautiques*, three people out of 12 said that it is laborious to have to download the new version of a PDF edition each week. However, one person remarked that updating the paper editions of the *Instructions nautiques* is difficult and therefore that updating the PDF editions is simpler. Finally, regarding students learning how to use the *Instructions nautiques*, three people told us that their students are too dependent on the Web when searching for information for which they should in fact search in the *Instructions nautiques*. Two people said that nowadays their students have difficulties reading, that they are no longer accustomed to reading books and that they are not comfortable with consulting them, which means that they prefer using the Web to search for information. Another person told us that the lack of knowledge surrounding the *Instructions nautiques* amongst students, making them too dependent on the Web even though they should depend more on the *Instructions nautiques*, which are more reliable, and use the Web only when necessary. Apart from their over-dependence on the Web, one person mentioned that their students tend to directly consult nautical charts even though they are taught to first consult the *Instructions nautiques* and then the nautical charts.

4.4.3.1.5 Other Resources We divided the different resources consulted along side the *Instructions nautiques* by our 12 interviewees into four cat-

egories (some resources may belong to more than one category). In descending order of popularity, they are:

— Other resources produced by the Shom

— Online resources

— Resources produced internally by the *Marine nationale*

— Resources produced by other national hydrographic services

Other Shom Publications All the people that we interviewed said they they use at least one other resource produced by the Shom along side the *Instructions nautiques*. The *Feux et signaux de brume* publications were cited by seven people, the *Courants de marée* by six people and the *Annuaire des marées* also by six people. As we previously indicated, five people cited nautical charts as a resource that they consult along side the *Instructions nautiques*. Three people said that they use the *Radiosignaux* publications, and two people said that they consult the online *Groupe d'Avis aux Navigateurs* (GAN) corrections [19]. Two other Shom websites were cited once each: the data portal *data.shom.fr* [20] for charts such as Litto3D, and the tides portal *maree.shom.fr* [21]. Finally, the *Guide du Navigateur* [22] publications were cited by one person out of the 12 interviewed.

Online Resources Out of the 12 people that we interviewed, eight said that they use at least one online resource along side the *Instructions nautiques*. As we noted in the previous paragraph, four people mentioned using online resources produced by the shom: the online GAN corrections, the data portal *data.shom.fr* and the tides portal *maree.shom.fr*. Three people said that they carry out generic searches on the Web to find either images (two people), especially of landmarks, or information such as telephone numbers (one person). Two people out of 12 mentioned that they use Google Earth [23] to see where they are going to go and to consult images of landmarks, and one person mentioned using Google Maps [24] satellite images. The *Atlas of Pilot Charts* [25], published by the National Geospatial-Intelligence Agency (NGA), were cited by two people as being a source of meteorological and environmental information such as currents.

19. https://diffusion.shom.fr/gan
20. https://data.shom.fr/
21. https://maree.shom.fr/
22. https://diffusion.shom.fr/ouvrages/ouvrages-generaux/guide-du-navigateur-volume-1.html, https://diffusion.shom.fr/ouvrages/ouvrages-generaux/guide-du-navigateur-volume-2.html, https://diffusion.shom.fr/ouvrages/ouvrages-generaux/guide-du-navigateur-volume-3.html
23. https://earth.google.com/
24. https://maps.google.com/
25. https://msi.nga.mil/Publications/APC

The *Sailing Directions (Enroute)*[26], also published by the NGA, were cited by one person. One person said that they consult the *ADMIRALTY Digital List of Lights*[27], published by the UK Hydrographic Office (UKHO), which are similar to the *Feux et signaux de brume* publications produced by the Shom. The following pieces of software were cited each by one person out of 12: *MaxSea*[28] for leisure navigation, *MétéoConsult*[29], *OpenCPN*[30] and an application that provides information about currents that the interviewee did not name.

Internal Resources Regarding the resources produced internally by the *Marine nationale*, we will take into account only the five students and the four instructors from the *École navale*. Six of these nine people mentioned the use of resources produced internally within the *Marine nationale* along side the *Instructions nautiques*: half cited the use of the *Guides des ports*, which are regularly-updated documents about entering ports (see fuller explanation in section 4.4.3.1.9), and half cited ad-hoc feedback documents known as *RETEX*, which are written by colleagues upon their return from a mission. One person mentioned using internal military documents without giving further details.

Resources from Outside France Four out of the 12 people that we interviewed indicated that they use resources that are equivalent to the publications produced by the Shom along side the *Instructions nautiques*. Two out of 12 cited the *ADMIRALTY Sailing Directions*[31]. As we have previously mentioned, two people cited use of the *Pilot Charts*, one person the *ADMIRALTY Digital List of Lights* and one person the *Sailing Directions (Enroute)*.

4.4.3.1.6 The Ideal *Instructions nautiques* Tool

We asked the 12 people that we interviewed to describe what the "ideal *Instructions nautiques* tool", based on the *Instructions nautiques*, would look like. We specified that this tool could replace the current *Instructions nautiques*, that it could integrate elements from other sources and that it could take the form of

26. https://msi.nga.mil/Publications/SDEnroute
27. https://www.admiralty.co.uk/digital-services/admiralty-digital-publications/admiralty-digital-list-of-lights
28. http://comen.maxsea.fr/maxsea/
29. https://www.meteoconsult.fr/
30. https://opencpn.org/
31. https://www.admiralty.co.uk/publications/publications-and-reference-guides/admiralty-sailing-directions

something other than a physical book or a static PDF. We identified three categories of responses in their answers:

— Comments on the potential format of the tool

— Suggestions of elements from other sources that could be integrated into the tool

— Criticisms regarding the creation of a digital tool based on the *Instructions nautiques*

Format Regarding the possible format of a tool based on the *Instructions nautiques*, eight out of the 12 interviewees suggested a digital tool based on an interactive nautical chart. The chart would allow the user to select a geographic zone and then see the elements and important information associated with the zone. It was suggested that the following elements feature on the chart: landmarks, dangers, main routes, up-to-date information about buoys, for example, the rules in vigour in each port and lighthouse signals. By clicking on these elements on the chart, supplementary information and photos could be displayed. Four people suggested either a function that allows categorising the information displayed according to the type of navigation selected by the user, or a function that allows the user to filter the categories of information to be displayed on the chart. In parallel to the idea of an interactive nautical chart, two other ideas for the tool stand out in terms of popularity. The first idea, cited by seven people out of 12, involves having access to the full text of the *Instructions nautiques*, organised in the same way as in the current volumes, somewhere in the tool. The second idea, also cited by seven people, involves having less formally-written text visible and more summaries and recap charts whilst conserving a link to the full text. Moving away from the idea of a tool based on an interactive chart, two people out of 12 said that the ideal tool for them would be interactive versions of the current *Instructions nautiques* PDF editions. Three other comments were made on the possible format of a tool based on the *Instructions nautiques*, each one made by one person. First of all, one person would like a tool that has regular updates and that updates by itself. The second person would like to be able to search using keywords, such as “channel X access”, “isle of Y lighthouse” or “port Z leading line”. The third person suggested integrating a participative aspect in the tool, which would allow users to suggest additions or modifications to the textual or iconographic content from their own experiences and observations.

Integration of Elements from Other Sources Out of the 12 people that we interviewed, 10 replied favourably to the integration of information or elements from other sources in a tool based on the *Instructions nautiques*. The element cited the most often was the nautical chart. As we have already mentioned, eight people said that they would like to see nautical charts as part of the tool. Those that we interviewed often mentioned giving access to full PDF publications in the tool, either by dedicating a page or a tab to them or by providing access to their content via an interactive chart. The publications cited are the following: *Radiosignaux* (three people), *Feux et signaux de brume* (three people), *Courants de marée* (two people), the GAN corrections for the *Instructions nautiques* (two people). As well as these Shom publications, two people out of 12 suggested the integration of the Météo-France [32] weather forecast in the tool. Finally, regarding the integration of more visual elements: seven people would like to see more photos and two people would like to see videos of vessels entering ports. Recording videos of vessels entering ports is a widespread practice in the French Navy, on a personal basis. One person indicated that a main video of 20 seconds would suffice for a vessel entering a port, with shorter clips of 3-4 seconds of the vessel at 500, 200 and 50 metres from the port, for example.

Criticisms Contrary to these positive suggestions, we also received criticisms surrounding the idea of creating a digital tool based on the *Instructions nautiques*. Three people out of 12 made negative comments about such a tool. One person said that getting used to a new tool requires too much time and that it would be better to keep the current format of the *Instructions nautiques*. Another person said that the more a tool is linked to the Web, the more complicated it becomes. The third person said that having a digital tool is not important to them. This person is of the opinion that the advantages of the *Instructions nautiques* in their current form are that there is a continuity from port to port, that there is a continuity between volumes and therefore between countries, and that having one volume dedicated to one region allows keeping a continuity from the general to the specific. For this person, these advantages would be lost by creating an information storage system and dissociating the information from its original volume. Another comment made by this person regards the potential search engine integrated in the tool: if the search engine does not find any results for a given search query, how do we know whether the answer really does not exist or whether we asked the wrong question?

32. https://meteofrance.com/

This problem is less important when working with the current *Instructions nautiques*.

4.4.3.1.7 Thoughts on Spatial Relations We asked 10 out of our 12 interviewees to give their opinions on two different ways of describing the position of a spatial entity: the first is by giving its geographic coordinates and the second is by describing the position of the entity via its spatial relations with other entities. Half of these 10 people said that both are important and that they serve different purposes. Spatial relations help to mentally locate oneself and to locate an entity on a chart, whilst geographic coordinates bring precision. Two people told us that spatial relations are important when reading a radar screen given that geographic coordinates are not displayed on radar screens. Two other people said that spatial relations are more important than geographic coordinates. However, one person out of 10 was of the opinion that geographic coordinates are more important than spatial relations and another person said that the utility of spatial relations depends on the quality of the writing.

4.4.3.1.8 Utility of Images in 3D All of the four people with whom we discussed 3D images were against the idea. Their reasoning was either linked to their uselessness or to the danger of being too absorbed in a model instead of reality.

4.4.3.1.9 Other Comments During the interviews, a certain number of remarks were made about subjects related to the *Instructions nautiques* such as:

— Other Shom publications
— Publications not produced by the Shom
— Teaching at the *École navale* and the ENSM
— The future of the project

Feux et signaux de brume Two people made comments about the *Feux et signaux de brume* publications produced by the Shom during the interviews. One person said that they do not use the *Feux et signaux de brume* because they can obtain all the information that they contain elsewhere. For example, the characteristics of lights can be found on digital nautical charts and on the digital navigation system found on vessels. The other person said that the *Feux et signaux de brume* publications are not very easy to use, and that it results in them consulting the chart directly even though they know that they should first consult the book.

Radiosignaux We received two comments about the *Radiosignaux* publications produced by the Shom during the interviews. One person mentioned that these publications overlap somewhat the *Instructions nautiques*, but that they contain more information. Another person explained that having information presented in tables as in the *Radiosignaux* publications is more practical than the way that information is presented in the current *Instructions nautiques*.

Pilotes Côtiers The *Pilotes Côtiers* are coastal navigation guides dedicated to leisure navigators. During the interviews, one person explained that the *Pilotes Côtiers* are well made and contain precise diagrams. According to this person, the *Pilotes Côtiers* are like small versions of the *Instructions nautiques*, which is a format that is suited to leisure navigators because they have more time to read.

Guides des ports The *Guides des ports* are documents produced internally by the *Marine nationale*. Each time that a vessel belonging to the *Marine nationale* enters a port, the crew is obliged to add or update the information, photos and videos in the guide. A *Guide du port* contains general information about the port, about the way in which the pilotage service works (and therefore whether the manoeuvre must be well prepared in advance or not), about the type of tugboat, about the port entrance and general feedback on the experience, all written in day-to-day language. The content of the *Guides des ports* is therefore more targeted towards *Marine nationale* vessels than the *Instructions nautiques*. However, updates to these documents can be sporadic: those for frequently-visited ports are updated regularly whilst for rarely-visited ports there is a risk that they are outdated. During the interviews, one person said that the *Guides des ports* are indispensable because they contain the key elements from the *Instructions nautiques* and recent feedback. Another person told us that a good *Guide du port* is better than an edition of the *Instructions nautiques*. This was explained by another person who said that the *Instructions nautiques* are left aside in favour of the *Guides des ports*, which are much more used. They mention that the two publications are very similar but that the *Guides des ports* are more targeted to the way in which the *Marine nationale* navigates, and that they contain recent feedback. Finally, one person suggested combining the *Guides des ports* with the *Instructions nautiques*: to complete the *Instructions nautiques* with the *Guides des ports*.

ADMIRALTY Sailing Directions During the interviews, some people mentioned the *ADMIRALTY Sailing Directions*, either to talk about their ease of use or to make comparisons between them and the *Instructions nautiques*. One person said that the *ADMIRALTY Sailing Directions* are very expensive but that they are well-made and that they have a good geographic coverage. Another person indicated that they have a better geographic coverage than the *Instructions nautiques* and that they are therefore used by francophone navigators even though the *Instructions nautiques* are of a higher quality. On the other hand, one person said that the information in the *ADMIRALTY Sailing Directions* is more directly available than the information in the *Instructions nautiques*. To help them find the right volume of the *ADMIRALTY Sailing Directions* for their needs, one person told us that they use online forums. Finally, one person said that the Shom publications are not used at all in their working environment, even amongst francophone crew members, because the *ADMIRALTY Sailing Directions* are much more practical.

Sailing Directions (Enroute) The *Sailing Directions (Enroute)* were mentioned by one person. This person said that the advantages of these publications are that they have a global coverage, they contain ready-made *fiches de traversée*[33], and they are freely available online. They also added, however, that the ergonomic design of these publications is of average quality.

ADMIRALTY Digital List of Lights One person made comments about the *ADMIRALTY Digital List of Lights*, a British equivalent of the *Feux et signaux de brume* produced by the Shom. According to this person, the *ADMIRALTY Digital List of Lights*, which take the form of a piece of software, is very well made. This person said that they would like to something comparable for the *Instructions nautiques*.

Classes at the École navale Regarding the teaching surrounding the *Instructions nautiques* at the École navale, one person told us that the instructors present the *Instructions nautiques* to the students, explain how to use them to prepare an itinerary, and explain how to use them along side the electronic navigation system found onboard vessels.

33. A *fiche de traversée* is a 10-15 page document that must be prepared ahead of each mission that contains the departure and arrival times, routes, the weather forecast, the stations to be contacted, the TSS or straights to be navigated, pilotage instructions, danger zones and telephone numbers.

Classes at the ENSM Regarding navigation classes at the ENSM, one person said that the instructors spend a lot of time getting to know the *Instructions nautiques* before teaching students about them. The *Instructions nautiques* feature early on in the teaching programme for the students at the ENSM because they are a priority tool and they are not necessarily very easy to use. In during the first classes, the instructors give students simple questions, such as "What is the height of Tower X?", to which they have to find the answer by searching in the *Instructions nautiques*. According to this interviewee, the students find it frustrating to have to work on these French publications whilst in reality everyone uses the *ADMIRALTY Sailing Directions*. Regarding the writing style of the *Instructions nautiques*, some students get lost in the verbose paragraphs and can therefore miss important information. This interviewee said that around half of the students have this problem, whilst the other half appreciate having something interesting to read. Finally, they added that the students believe that Wikipedia is almost as reliable as the official documentation. Another interviewee said that a great deal of attention is given to navigation publications, especially the *Instructions nautiques* and especially for the departure and arrival of a vessel.

Continuation of the Project The people that we interviewed were on the whole very interested in our project and we received multiple requests to be kept up to date about our work. We also received two suggestions about the remainder of the project. One person suggested that, after completing our series of interviews, we create an online questionnaire with the main questions that we asked during the interviews to gather more responses. Having noticed a considerable difference in the text between different volumes of the *Instructions nautiques*, another person wondered whether we would need to take into account the author of the text when working on the extraction of geographic information.

4.4.3.1.10 Discussion Overall, we saw a great diversity in the responses we received during the interviews. However, on closer examination, it is possible to identify a number of common threads running through a large proportion of the responses.

The *Instructions nautiques* are mainly used during the preparation of a mission and rarely whilst the mission is ongoing. They are generally used to provide an overview of the route to be taken and to help draft the official preparatory documents. The completion of these preparatory documents requires identifying specific information such as routes, stations

to be contacted, piloting instructions and the telephone numbers of various services. These pieces of information are amongst the most sought-after elements in the *Instructions nautiques*.

If we look more closely at the specific information that is searched for in the *Instructions nautiques*, priority is given to four main elements:

— Information on the communications to be carried out in different circumstances is required in order to know who to contact, when and how (telephone number or VHF channel)

— Photos, especially of landmarks, are highly sought after to get a better idea of what to expect to see on the horizon and what the coastal landscape actually looks like

— Port information, especially about approaches and entrances, is also highly sought after, as well as administrative information, pilotage information and information more specific to each port

— Landmarks, leading lines and lights

With regard to the comments about the *Instructions nautiques* in their current form, certain remarks were made repeatedly. There was a consensus on the idea that today, the *Instructions nautiques* contain too much information or do not present information in the right way. More generally, consultation of current *Instructions nautiques* is not considered to be efficient. As well as proposing other ways of accessing important information, there were calls for more photos, especially of landmarks, and for photos and information to be reliable and up to date.

The other sources of information used along side the *Instructions nautiques* are mainly other publications by the Shom: *Feux et signaux de brume*, *Courants de marée*, *Annuaire des marées* and nautical charts. We also noted a willingness to consult other sources that take the form of digital tools.

Most of those we interviewed are in favour of the development of a digital tool based on the content of the *Instructions nautiques*. The most popular idea is the creation of a platform that consists of an interactive nautical chart that gives visual and personalisable access to the information and photos contained within the current *Instructions nautiques* series. The addition of more photos, and even videos, would be appreciated. Whilst the way in which key information is presented could be optimised by introducing more summary paragraphs and standardised tables, it is impossible to ignore the want to conserve access to the full text version of the *Instructions nautiques*. There is also a minority that are reluctant to seeing the *Instructions nautiques* evolve.

Finally, we noted that some of those that we interviewed are driven to use equivalents of the *Instructions nautiques* produced by other countries' hydrographic services, in particular the UK *ADMIRALTY Sailing Directions* and the US *Sailing Directions (Enroute)*, because of their greater geographic coverage and thanks to the fact that they are available in English.

This series of interviews provided us with a lot of information about the way in which the *Instructions nautiques* are used today in different working environments. It also taught us what a sample of users feel is lacking in the current *Instructions nautiques* as well as the way in which they would like to see the *Instructions nautiques* evolve.

The knowledge graph that we aim to construct will contain geospatial knowledge. We can therefore envisage the development of an *Instructions nautiques*-based tool in the form of an online platform featuring an interactive nautical chart. This chart could provide visual access to the content of the knowledge graph and therefore the content of the *Instructions nautiques* as we know them today. We also plan to offer a search engine that would enable the knowledge base to be queried with specific questions. Thanks to the interviews, we have understood that the architecture of the knowledge graph should be oriented primarily around queries on communications to be made, photos, ports, landmarks, alignments and lights. We have also learned that users would like to be able to read summaries of key information or pre-prepared summary sheets. It would be possible to integrate such alternative versions of the text into the platform, created from the structured content of the knowledge graph. Finally, we could consider adding geographical coordinates and descriptions of spatial relations to the text. At a later stage, elements from other sources would need to be integrated through links that could be made between this knowledge graph and other sources. According to the results of the interviews, the integration of more photos should be a priority. The GAN *Instructions nautiques* corrections[34], which are currently distributed in PDF format, could be integrated directly into the knowledge graph, which would update any product linked to the graph in real time. Adding an indication of the latest update to any textual or iconographic element of the product could reassure users that the product is up to date and reliable.

4.4.3.2 Motivating Scenario

We wrote the motivating scenarios with the help of the reference documents described in section 4.4.2.2 and refined them after having analysed

34. https://diffusion.shom.fr/gan

and synthesised the use and the importance of the relevant concepts and properties in the *Instructions nautiques* according to our interviews with their users. We wrote four motivating scenarios, one for each subdomain identified in section 4.4.2.6 on page 64. Their names are as follows:

— Maritime Navigation Guidelines

— Maritime Spatial Entities and Spatial Relations

— Temporalities, Meteorological and Oceanographic Phenomena

— Maritime Vessels

Extract 4.4.1 on page 64, the knowledge contained within which is decomposed into semantic triples in document 4.4.1 on page 65, matches the themes of the Maritime Spatial Entities and Spatial Relations subdomain, the Temporalities, Meteorological and Oceanographic Phenomena subdomain and the Maritime Vessels subdomain. We decided that it is most relevant to the Maritime Spatial Entities and Spatial Relations subdomain and therefore included it in that motivating scenario.

A simplified version of the motivating scenario that we wrote for the Maritime Navigation Guidelines subdomain is shown in document 4.4.2. The full motivating scenario for the Maritime Navigation Guidelines subdomain can be found in appendix section B.1 on page 173. See appendix sections C.1, D.1 and E.1 on pages 177, 181 and 185 respectively for the full motivating scenarios for the three other subdomains.

4.4.3.3 Informal Competency Questions

We wrote the informal competency questions after having started the interview process, once we had identified the most common pieces of information searched for by users of the *Instructions nautiques*. To validate our informal competency questions, we shared them with our interviewees during the later interviews. We asked our interviewees to validate the relevance of the questions and to optimise them when necessary.

Some of the informal competency questions that we wrote for all four of our subdomains are shown in document 4.4.3 on page 86. The full lists of informal competency questions for each subdomain can be found in appendix sections B.2 on page 175, C.2 on page 179, D.2 on page 182 and E.2 on page 186.

4.4.3.4 Glossary

We compiled the glossaries to explain the subdomain-specific vocabulary that we used in the rest of the documentation. We used definitions from

MOTIVATING SCENARIO

Name
Maritime Navigation Guidelines
Description
The *Instructions nautiques* contain many maritime navigation guidelines. Maritime navigation guidelines are pieces of information, instructions or prohibitions that concern all possible actions in the maritime domain. Such actions are most commonly navigating or remaining stationary on the water. Maritime navigation guidelines can also come in the form of contact information. An instruction can be advisory or obligatory, in which case a decree is cited. A prohibition is necessarily obligatory and cites a decree. A piece of information can be linked to a decree. Navigation guidelines can be dependent on local conditions such as temporality (time of day, season, etc.) or meteorological and oceanographic conditions. They can also be targeted at maritime vessels with specific characteristics such as size or origin.
Extracts
See extracts 2.3.3 (p. 10), 2.3.4 (p. 10) and 2.3.6 (p. 18).
Main Concepts and Characteristics
A maritime navigation guideline must be of one of the following types:

- information
- instruction
- prohibition
- contact information
- decree

Maritime navigation guidelines typically have the following optional or obligatory characteristics:

- type of guideline [obligatory]
- region to which guideline applies [obligatory: information | prohibition, optional: instruction]
- spatial entity to be followed according to guideline [optional: instruction]
- action to be carried out according to guideline [obligatory: instruction]
- action prohibited by guideline [obligatory: prohibition]
- decree at origin of guideline [optional: information | instruction | prohibition]
- name of decree [obligatory: decree]
- local condition under which guideline is valid [optional: information | instruction | prohibition | contact information]
- target of guideline [optional: information | instruction | prohibition | contact information]
- exception to guideline [optional: information | instruction | prohibition]
- complementary information [optional]

Document 4.4.2 – The motivating scenario for the Maritime Navigation Guidelines subdomain.

INFORMAL COMPETENCY QUESTIONS

Maritime Navigation Guidelines

1. What are the navigation instructions for The Great Western Pass? (Extract 2.3.3)
   — It is not recommended to take The Great Western Pass.
2. How can the North Channel be accessed? (Extract 2.3.4)
   — During the day, the channel entry access route is oriented at approximately 114° towards the southern extremity of the summit of Mont Mahinia, or towards the northern slope of Mont de la Selle.

Maritime Spatial Entities and Spatial Relations

1. What does Barn Hill Point look like from a boat on the water? (Extract 2.3.1)
   — Barn Hill Point is the extremity of a narrow craggy peninsula that reaches 1 M SSW of Taliokoaka, a headland 60 m tall. This peninsula is lined with white limestone cliffs that stand out when illuminated by the sun.
2. Is The Great Western Pass marked? (Extract 2.3.3)
   — No, The Great Western Pass is unmarked.
3. What is the orientation of the North Channel entry access route? (Extract 2.3.4)
   — During the day, the channel entry access route is oriented at approximately 114° towards the southern extremity of the summit of Mont Mahinia, or towards the northern slope of Mont de la Selle.
4. What colour is the pyramid of Île Pigued? (Extract 2.3.6)
   — The colour of the pyramid of Île Pigued is white.
5. What landmarks are there on the Île de Batz? (Extracts 2.3.6 and 4.4.2)
   — On the Île de Batz, Île de Batz bell tower, Notre-Dame de Bon Secours chapel, a semaphore, a semaphore tower and a lighthouse serve as landmarks.

Temporalities, Meteorological and Oceanographic Phenomena

1. What is the climate like on the Kerguelen Islands? (Extract 2.3.2)
   — The climate is cold, humid and very windy.
2. Does it snow on the Kergeulen Islands? (Extract 2.3.2)
   — On the coastal plains, snow can fall at any time of year but rarely lasts more than a few days.

Maritime Vessels

1. Is the alignment of Île de Batz clock tower and the pyramid of Île Pigued visible to all vessels? (Extract 2.3.6)
   — The alignment of Île de Batz clock tower and the pyramid of Île Pigued is visible to small vessels up to around 0.6 M to the east of 'Le Menk' turret (at half tide) and, for vessels with higher bridges, up to the north of the turret.

Document 4.4.3 – Some informal competency questions for all four of our subdomains.

the reference documents described in section 4.4.2.2 on page 62 onwards where possible, and give definitions for concepts that we created for the purpose of the project.

The glossary that we wrote for the Maritime Navigation Guidelines subdomain is shown in document 4.4.4 on page 88. See appendix sections C.3, D.3 and E.3 on pages 180, 182 and 186 respectively for the full glossaries for the three other subdomains.

4.4.4 Step 3: Structuring, Implementing and Testing Subdomain Models

4.4.4.1 Conceptualising Subdomain Models

For each subdomain, we analysed the semantic triples produced during the groundwork phase and inserted them into their corresponding motivating scenario. We then grouped together the subjects/objects and the predicates that serve the same purpose. Document 4.4.5 on page 89 summarises the subjects and objects that we extracted from the semantic triples in document 4.4.1 on page 65, which we had assigned to the Maritime Spatial Entities and Spatial Relations subdomain, and document 4.4.6 on page 90 summarises the predicates. This extract contained some concepts and predicates that were better suited to other subdomain themes and which we therefore added to their corresponding models. They are identified by asterisks (*) in documents 4.4.5 and 4.4.6.

During this process we made some changes to the way we had decomposed the knowledge into semantic triples. For example, in document 4.4.1 we defined atoll 1 - is - wooded and spatial entity 2 - is not - wooded. When taking into account the application of the ontology, we realised that for users who were searching for information about the visual aspect of a spatial entity it would be more useful to group together all types of visual characteristics, whether affirmative or negative, and have only one corresponding affirmative predicate rather than an affirmative and a negative one. Otherwise, looking up the visual aspect of a spatial entity would imply searching for affirmative and negative statements about it, which is not intuitive. Document 4.4.6 shows the single affirmative predicate has visual aspect and the affirmative and negative named individuals wooded and not wooded that we created to implement this change.

4.4.4.2 Implementing Subdomain Models in OWL

Listing 4.4.2 shows declarations of OWL Classes, NamedIndividuals, ObjectProperties and DatatypeProperties that correspond to some of the

GLOSSARY

Maritime Navigation Guidelines

Term	Definition
Action	Any possible action that can be executed in the maritime domain. The most common are 'navigating' and 'remaining stationary' (on the water). Other examples include 'dragging', 'fishing' and 'swimming'.
Contact information	A type of maritime navigation guideline that gives one or more ways of contacting a service that may need to be reached before or during navigation.
Information (piece of)	A type of maritime navigation guideline that is purely informative and can be linked to a decree.
Instruction	A type of maritime navigation guideline that indicates how an activity is to be performed. It can be advisory or obligatory, in which case a decree is cited.
Local condition	*See glossary for Temporalities, Meteorological and Oceanographic Phenomena subdomain*
Maritime domain	*See glossary for Maritime Spatial Entities and Spatial Relations subdomain*
Maritime navigation guideline	A piece of information, an instruction or a prohibition that concerns any possible action that can be executed in the maritime domain.
Maritime navigation guideline type	A category of maritime navigation guideline. There are five categories of maritime navigation guideline: information, instruction, prohibition, contact information, decree.
Maritime vessel	*See glossary for Maritime Vessels subdomain*
Material characteristic	*See glossary for Maritime Vessels subdomain*
Navigating	An action that involves the deliberate movement of a vessel on the water.
Prohibition	A type of maritime navigation guideline that indicates an action that may not performed. It is necessarily obligatory and cites a decree.
Remaining stationary	An action that involves avoiding the movement of a vessel at a given position on the water by mooring or by dropping the anchor. Places where the action of remaining stationary can be executed are called 'stopping places'.
Target	The type of maritime vessel for which a maritime navigation guideline applies.

Document 4.4.4 – The glossary for the Maritime Navigation Guidelines subdomain.

Maritime Spatial Entities and Spatial Relations
Subjects/Objects

- spatial entity
  - atoll, motu, point, village, lighthouse, channel, set of posts, landing point
- visual aspect
  - wooded, not wooded
- visibility
  - partial
- oceanographic phenomenon*
  - current*
- cardinal direction
  - northeast
- vessel**
  - small boat**

*To be added to the Temporalities, Meteorological and Oceanographic Phenomena subdomain model.
**To be added to the Maritime Vessels subdomain model.

Document 4.4.5 – The grouping together of the subjects and objects in document 4.4.1 on page 65 that serve the same purpose.

concepts and predicates featuring in documents 4.4.5 and 4.4.6. Again, during the implementation process we were driven to make some changes to the informal modelling presented in these documents thanks to the possibilities presented by OWL. For example, in document 4.4.6 we defined the following three predicates: has region, is part of and is located in. During their implementation we realised that is part of and is located in represented the same meaning: that the spatial footprint of one spatial entity was contained within that of another, and that has region had the exact inverse meaning. We therefore merged is part of and is located in into one property, `atln:isPartOf`, and standardised the name of its inverse property, `atln:hasPart`, to match.

4.4.4.3 Creating Subdomain Datasets in RDF

We were then able to rewrite the semantic triples from document 4.4.1 as RDF triples, as shown in listing 4.4.3, using the first iteration of the subdomain model presented in listing 4.4.2.

Once the first draft of a formal model and a set of RDF triples had been produced for each subdomain, we continued structuring the extracts from the motivating scenarios according to the subdomain models and refining

Maritime Spatial Entities and Spatial Relations
Predicates

- [INDIVIDUAL] is a [spatial entity | oceanographic phenomenon]*
- [spatial entity] is called [STRING]
- [spatial entity] is east of [spatial entity]
- [spatial entity] is 30 M from [spatial entity]
- [spatial entity] has geographic coordinates [STRING]
- [spatial entity] has diameter [NUMBER + UNIT]
- [spatial entity] has region [spatial entity]
- [spatial entity] has visual aspect [wooded | not wooded]
- [spatial entity] is part of [spatial entity]
- [spatial entity] shelters [spatial entity]
- [spatial entity] has visibility [visibility]
- [spatial entity] is visible at [NUMBER + UNIT]
- [oceanographic phenomenon] is south of [spatial entity]*
- [oceanographic phenomenon] flows [cardinal direction]*
- [spatial entity] is near [spatial entity]
- [oceanographic phenomenon] is dangerous at [spatial entity]*
- [spatial entity] has target [vessel]**
- [spatial entity] has orientation [NUMBER + UNIT]
- [spatial entity] has length [NUMBER + UNIT]
- [spatial entity] has width [NUMBER + UNIT]
- [spatial entity] has depth [NUMBER + UNIT]
- [spatial entity] is northwest of [spatial entity]
- [spatial entity] is marked by [spatial entity]
- [spatial entity] is located in [spatial entity]

*To be added to the Temporalities, Meteorological and Oceanographic Phenomena subdomain model.
**To be added to the Maritime Vessels subdomain model.

Document 4.4.6 – The grouping together of the predicates in document 4.4.1 on page 65 that serve the same purpose.

```
atln:SpatialEntity rdf:type owl:Class ;
    rdfs:label "spatial entity"@en .

atln:Atoll rdf:type owl:Class ; # idem. for the other spatial entity types
    rdfs:subClassOf nav:SpatialEntity ;
    rdfs:label "atoll"@en .

atln:VisualAspect rdf:type owl:Class ;
    rdfs:label "visual aspect"@en .

atln:Wooded rdf:type owl:NamedIndividual ;
    rdfs:label "wooded"@en .

# for "is a" we will use rdf:type

# for "is called" we will use rdfs:label

atln:isEastOf rdf:type owl:ObjectProperty ; # can be declined for all
    cardinal directions
    rdfs:domain atln:SpatialEntity ;
    rdfs:range atln:SpatialEntity ;
    rdfs:label "is east of"@en .

atln:hasPart rdf:type owl:ObjectProperty ; # for "has region"
    rdfs:domain atln:SpatialEntity ;
    rdfs:range atln:SpatialEntity ;
    rdfs:label "has part"@en .

atln:isPartOf rdf:type owl:ObjectProperty ; # for "is part of" AND "is
    located in"
    owl:inverseOf atln:hasPart ;
    rdfs:domain atln:SpatialEntity ;
    rdfs:range atln:SpatialEntity ;
    rdfs:label "is part of"@en .
```

Listing 4.4.2 – The first draft of the OWL implementation in Turtle syntax of some of the concepts and predicates featuring in documents 4.4.5 and 4.4.6.

```
ent:0001 rdf:type atln:SpatialEntity ; # spatial entity 1 - is a - spatial
    entity
    rdfs:label "Makemo"@en . # spatial entity 1 - is called - Makemo

ent:0002 rdf:type atln:Atoll ; # atoll 1 - is a - atoll
    rdfs:label "Nihiru"@en ; # atoll 1 - is called - Nihiru
    atln:isEastOf ent:0001 ; # atoll 1 - is east of - spatial entity 1
    atln:hasVisualAspect atln:wooded ; # atoll 1 - is - wooded
    atln:hasPart ent:0003 ; # atoll 1 - has region - spatial entity 2
    atln:hasPart ent:0005 . # atoll 1 - has region - point 1

ent:0003 rdf:type atln:SpatialEntity ; # spatial entity 2 - is a - spatial
    entity
    rdfs:label "south Nihiru"@en ; # spatial entity 2 - is called - south
        Nihiru
    atln:hasVisualAspect atln:notWooded . # spatial entity 2 - is not - wooded

ent:0004 rdf:type atln:Motu ; # motu 1 - is a - motu
    atln:isPartOf ent:0005 ; # motu 1 - is part of - point 1
    atln:shelters ent:0006 . # motu 1 - shelters - village 1

ent:0005 rdf:type atln:Point ; # point 1 - is a - point
    rdfs:label "SW point"@en . # point 1 - is called - SW point

ent:0006 rdf:type atln:Village ; # village 1 - is a - village
    rdfs:label "Tatake"@en . # village 1 - is called - Tatake

ent:0007 rdf:type atln:Lighthouse ; # lighthouse 1 - is a - lighthouse
    atln:hasVisibility atln:partial . # lighthouse 1 - has visibility - partial
```

Listing 4.4.3 – Some of the semantic triples from document 4.4.1 rewritten as RDF triples in Turtle syntax, using the first iteration of the subdomain model presented in listing 4.4.2.

the models accordingly.

4.4.4.4 Iterations of Subdomain Models and Datasets

During the iteration process we made some changes to the way in which we had initially modelled the subdomains. In particular, we noticed that in the extract decomposed into semantic triples in document 4.4.1, eight different types of spatial entities are mentioned within six sentences. We understood that it would therefore be almost impossible for our ontology to exhaustively cover all types of spatial entities that could be mentioned in the corpus. To solve this problem, we decided to create a Simple Knowledge Organization System (SKOS) [35] thesaurus `atln:TypeOfSpatialEntity` for the different types of spatial entities, as shown on line 9 in listing 4.4.4

35. https://www.w3.org/2004/02/skos/

and also presented as a graph in figure 4.4.2, rather having an OWL class for each of them as in listing 4.4.2, lines 1 to 6. The formalism of a SKOS thesaurus is simpler than the rigorous descriptive formalism of ontologies defined with OWL, and does not require a formal description of the semantics (Pastor-Sánchez et al. 2009). That means that it will be easier to automatically enrich a thesaurus with new SKOS concepts than to automatically create new OWL classes during knowledge graph population. In addition, spatial entity types do not require the creation of any axioms on top of their hierarchical relationships, making a SKOS thesaurus the most efficient solution.

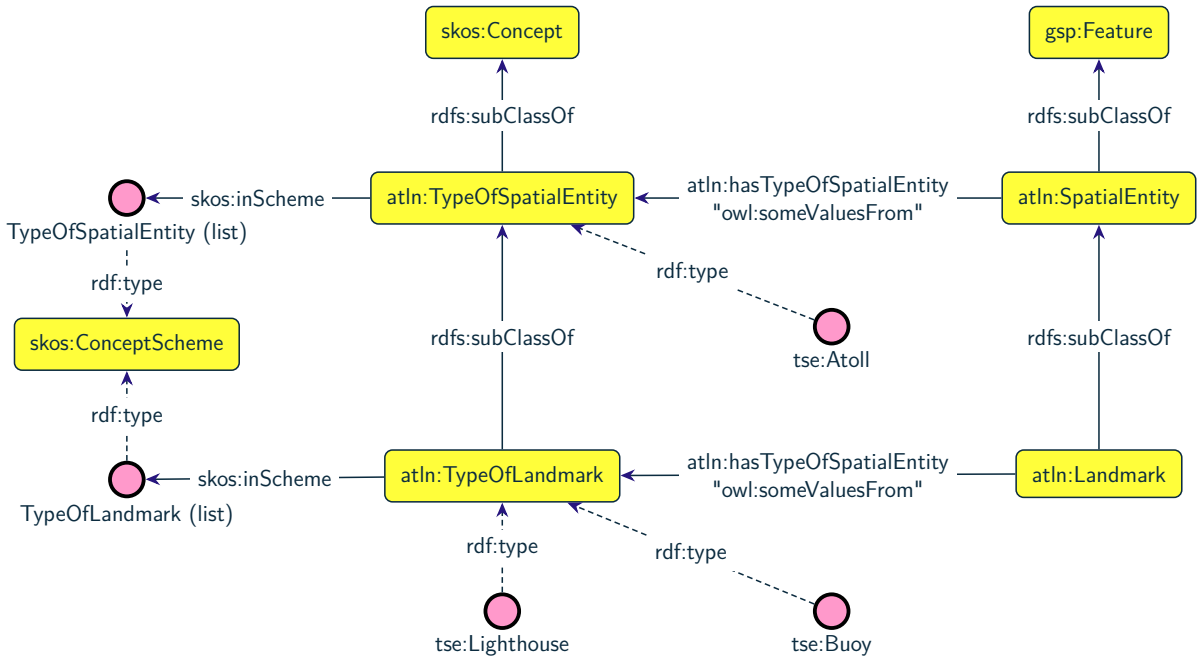


Figure 4.4.2 – Extract of the ATLANTIS ontology (Rawsthorne et al. 2022a) presented as a Graffoo diagram (Peroni 2013).

Within the `atln:TypeOfSpatialEntity` thesaurus we created three sub-thesauri: `atln:TypeOfLandmark`, `atln:TypeOfDanger` and `atln:TypeOfStoppingPlace` (see definition of 'Remaining stationary' in document 4.4.4). The declaration of the `atln:TypeOfLandmark` thesaurus can be found in listing 4.4.4, lines 30 to 36, and it is also presented as a graph in figure 4.4.2. The declaration of the `atln:TypeOfStoppingPlace` thesaurus can be found in listing 4.6.2, lines 9 to 15, on page 105. These thesauri serve to gather together the groups of spatial entities that we may want to isolate and query together. The `atln:TypeOfLandmark` thesaurus contains individuals such as `tse:Buoy` (see listing 4.4.4, lines 42 to 44, and figure 4.4.2) and `tse:Lighthouse` (see figure 4.4.2), the `atln:TypeOfDanger` thesaurus contains individuals such as `tse:Rock` and `tse:Shipwreck`, and the `atln:TypeOfStoppingPlace` thesaurus contains individuals such as `tse:Anchorage` and `tse:Mooring`. To show why this grouping together

```
atln:SpatialEntity rdf:type owl:Class ;
    owl:equivalentClass [ rdf:type owl:Restriction ;
                          owl:onProperty atln:hasTypeOfSpatialEntity ;
                          owl:someValuesFrom atln:TypeOfSpatialEntity
                        ] ;
    rdfs:subClassOf gsp:Feature ;
    rdfs:label "spatial entity"@en .

atln:TypeOfSpatialEntity rdf:type owl:Class ;
    rdfs:subClassOf skos:Concept ;
    rdfs:label "Spatial Entity Type"@en .

atln:hasTypeOfSpatialEntity rdf:type owl:ObjectProperty ;
    rdfs:subPropertyOf atln:hasCharacteristic ;
    rdfs:domain atln:SpatialEntity ;
    rdfs:range [ rdf:type owl:Restriction ;
                          owl:onProperty skos:inScheme ;
                          owl:hasValue <http://data.shom.fr/id/codes/atlantis/
                              typeofspatialentity/list>
               ] ;
    rdfs:label "has type of spatial entity"@en .

atln:Landmark rdf:type owl:Class ;
    owl:equivalentClass [ rdf:type owl:Restriction ;
                          owl:onProperty atln:hasTypeOfSpatialEntity ;
                          owl:someValuesFrom atln:TypeOfLandmark
                        ] ;
    rdfs:subClassOf gsp:SpatialEntity ;
    rdfs:label "landmark"@en .

atln:TypeOfLandmark rdf:type owl:Class ;
    owl:equivalentClass [ rdf:type owl:Restriction ;
                          owl:onProperty skos:inScheme ;
                          owl:hasValue <http://data.shom.fr/id/codes/atlantis/
                              typeoflandmark/list>
                        ] ;
    rdfs:subClassOf atln:TypeOfSpatialEntity ;
    rdfs:label "Type of Landmark"@en .

tse:Atoll rdf:type owl:NamedIndividual ,
                   atln:TypeOfSpatialEntity ;
    skos:prefLabel "atoll"@en .

tse:Buoy rdf:type owl:NamedIndividual ,
                  atln:TypeOfLandmark ;
    skos:prefLabel "buoy"@en .
```

Listing 4.4.4 – The declaration in Turtle syntax of a SKOS thesaurus to store the different types of spatial entities.

of certain spatial entities is useful, let us take the example of informal competency question number 5 for the Maritime Spatial Entities and Spatial Relations subdomain as shown in document 4.4.3. This question is looking for a list of all the landmarks in a given region, regardless of their type. By employing the `atln:TypeOfLandmark` thesaurus, we can obtain an answer to this question without specifying each possible type of landmark in the query. This example is developed further in section 4.4.4.5.

We added the `gsp:Feature` class derived from the GeoSPARQL ontology [36] to our Maritime Spatial Entities and Spatial Relations subdomain model, as can be seen in figure 4.4.2. GeoSPARQL is a standard for representing and querying geographic entities. This standard allows us to define the geometry of features and access their descriptions in well-known text (WKT) or Geography Markup Language (GML) via properties.

Our interviews with users of *Instructions nautiques* taught us that they use them in particular to gather information before embarking about the landmarks that they should be able to see and use for navigating in coastal waters. When looking for information about a landmark, users are searching for all possible characteristics that describe it. Before consulting the *Instructions nautiques*, they don't know what type of information will be available for any given landmark. We meet this specific need by creating a hierarchy in the object properties so that all the characteristics of an instance of the `gsp:Feature` class can be retrieved by inference. The object property `atln:hasCharacteristic` therefore has several sub-properties such as `atln:hasShape` and `atln:hasColour`.

To complement our series of interviews with users of the *Instructions nautiques*, we held two meetings with the writers of the *Instructions nautiques*. During these meetings, we realised that it is vital, for safety reasons, to retain the indications of relative importance mentioned in the text. Extract 4.4.2 on page 96 illustrates this problem. The hierarchy indicated by the words "above all" helps a navigator in poor visibility conditions by advising them to use the lighthouse rather than the semaphore tower as a landmark. In order to solve this problem, and other similar problems that may arise due to subtle but important nuances in the text, we decided to associate each instance with the original sentence that mentions it in the text. The data property `atln:hasAssociatedText` therefore has the `gsp:Feature` class as its `rdfs:domain` and `xsd:string` as its `rdfs:range`. The `xsd:string` is destined to capture the entire original sentence that mentions the entity in question.

During this iterative phase of manual RDF triple creation and model

36. https://opengeospatial.github.io/ogc-geosparql/geosparql11/index.html

"Viewed from the north, Île de Batz shows the semaphore tower (48° 44.78' N - 4° 00.69' W) and above all the lighthouse (48° 44.72' N - 4° 01.61' W), a 43 m high grey tower surrounded by houses."

Extract 4.4.2 – Translated from the original French text: "Vue du Nord, l'Île de Batz montre la tour du sémaphore (48° 44,78' N — 4° 00,69' W) et surtout le phare (48° 44,72' N — 4° 01,61' W), tour grise haute de 43 m, entourée de maisons." (Shom 2021a, p. 398)

enrichment we used the open-source Git [37] as a VCS. We used the Git branching feature to separately follow each subdomain model. This allowed us to be able to easily retrieve earlier versions of a model during the development process, and also made it easier to integrate ulterior changes.

4.4.4.5 Testing Subdomain Models and Datasets

We carried out the model tests to verify the consistency of the subdomain models using the HermiT reasoner [38] (version 1.4.3.456) integrated in the ontology editor Protégé [39]. We read through the four motivating scenarios to check that each subdomain model corresponds to its initial description.

We carried out the data tests by modelling unseen extracts from the *Instructions nautiques* according to the subdomain models, thereby verifying their validity.

The query tests were performed using the SPARQL query function of the RDF triplestore GraphDB [40]. Listing 4.4.5 shows informal competency question number 5 for the Maritime Spatial Entities and Spatial Relations subdomain (see document 4.4.3 on page 86) transformed into a SPARQL query. The query first identifies the entity that corresponds to the Île de Batz and then selects all the landmarks whose geometry is contained within the geometry of the Île de Batz entity. For each landmark, the query retrieves its type and, if available, its name. Table 4.4.1 shows the results of this query when executed on our subdomain dataset of RDF triples. It identified five landmarks, two of which are cited in the navigation instructions shown in extract 2.3.6 on page 18 whilst the remaining three are cited on another page, in the coastal landscape description shown in extract 4.4.2. Thanks to the ontology and this query, we were able to group together these landmarks that were dispersed across different pages in the text of the *Instructions nautiques*.

37. https://git-scm.com/
38. http://www.hermit-reasoner.com/
39. https://protege.stanford.edu/
40. https://graphdb.ontotext.com/

```
SELECT DISTINCT ?typeOfLandmark ?landmarkLabel WHERE {
  ?entity atln:hasTypeOfSpatialEntity tse:Island;
          rdfs:label ?label.
  FILTER(REGEX(STR(?label), "Batz"))
  ?entity geom:hasGeometry ?batzGeom.
  ?batzGeom gsp:asWKT ?batzWKT.
  ?landmark atln:hasTypeOfSpatialEntity ?typeOfLandmark.
  ?typeOfLandmark rdf:type atln:typeOfLandmark.
  ?landmark geom:hasGeometry ?landmarkGeom.
  ?landmarkGeom gsp:asWKT ?landmarkWKT.
  OPTIONAL {
    ?landmark rdfs:label ?landmarkLabel.
  }
  FILTER(geof:sfContains(?batzWKT, ?landmarkWKT))
}
```

Listing 4.4.5 – The SPARQL query that corresponds to the informal competency question "What landmarks are there on the Île de Batz?" (see question number 5 for the Maritime Spatial Entities and Spatial Relations subdomain in document 4.4.3).

typeOfLandmark	landmarkLabel
tse:Lighthouse	"Île de Batz lighthouse"@en
tse:Chapel	"Notre-Dame de Bon Secours chapel"@en
tse:BellTower	"Île de Batz bell tower"@en
tse:Semaphore	
tse:Tower	"semaphore tower"@en

Table 4.4.1 – Results of the SPARQL query displayed in listing 4.4.5.

4.4.5 Step 4: Merging, Refactoring and Aligning

4.4.5.1 Merging Subdomain Models to Create Full Model

Once all four subdomain models had each passed the model, data and query tests, we merged their respective .owl files to create the full model. During this process we had to merge certain duplicate declarations. Let us take as an example the `atln:hasSpatialRelationWith` object property. In document 4.4.6 we identified the [spatial entity] is east of [spatial entity] predicate, destined for the Maritime Spatial Entities and Spatial Relations subdomain model, and the [oceanographic phenomenon] is south of [spatial entity] predicate, destined for the Temporalities, Meteorological and Oceanographic Phenomena subdomain model. We subsequently created a `atln:hasSpatialRelationWith` object property in both subdomain models, and defined `atln:isEastOf` and `atln:isSouthOf` as sub-properties of it in their respective subdomain models. In the Maritime Spatial Entities and Spatial Relations subdomain model, the `atln:hasSpatialRelationWith` object property had been defined as

having the gsp:Feature class as its rdfs:domain and as its rdfs:range. However, in the Temporalities, Meteorological and Oceanographic Phenomena subdomain model, the atln:hasSpatialRelationWith object property had been defined as having the atln:MeteorologicalOrOceanographicPhenomenon class as its rdfs:domain and a union of the atln:MeteorologicalOrOceanographicPhenomenon and gsp:Feature classes and the as its rdfs:range. When we merged the .owl files of these two subdomain models, we merged the two definitions of the atln:hasSpatialRelationWith object property so as to have a union of the gsp:Feature and atln:MeteorologicalOrOceanographicPhenomenon classes as its rdfs:domain and as its rdfs:range, as shown in listing 4.4.6, lines 1 to 13.

```
1  atln:hasSpatialRelationWith rdf:type owl:ObjectProperty ;
2      rdfs:subPropertyOf owl:topObjectProperty ;
3      rdfs:domain [ rdf:type owl:Class ;
4                    owl:unionOf ( atln:MeteorologicalOrOceanographicPhenomenon
5                                  gsp:Feature
6                                )
7                  ] ;
8      rdfs:range [ rdf:type owl:Class ;
9                   owl:unionOf ( atln:MeteorologicalOrOceanographicPhenomenon
10                                 gsp:Feature
11                               )
12                 ] ;
13     rdfs:label "has a spatial relation with"@en .
14
15 atln:isEastOf rdf:type owl:ObjectProperty ;
16     rdfs:subPropertyOf atln:hasSpatialRelationWith ;
17     owl:inverseOf atln:isWestOf ;
18     rdfs:label "is east of"@en .
```

Listing 4.4.6 – The declaration of a reverse OWL ObjectProperty in Turtle syntax.

4.4.5.2 Merging Subdomain Datasets to Create Exemplar Dataset

We merged the .ttls files of all four subdomain datasets to create the exemplar dataset [41].

4.4.5.3 Refactoring and Aligning Full Model

We created axioms that automatically classify some entities according to their properties and that infer new knowledge.

41. https://github.com/umrlastig/atlantis-ontology/blob/main/triplets.ttls

"This route passes east of the starboard lateral beacon"

Extract 4.4.3 – Translated from the original French text: "Cette route laisse dans l'Ouest la balise latérale tribord" (Shom 2021a, p. 320)

To demonstrate why this can be useful, let us first take the example of navigation marks. In the *Instructions nautiques*, marks can be referred to in one of two ways. Either, they can be referred to only by their physical type (beacon, buoy, turret) and mark type (cardinal, safe water, isolated danger, lateral or special) without using the word "mark", as in extract 4.4.3. Or, they can be referred to using the word "mark", their physical type and their mark type, as in extract 4.4.4. A lateral mark can either signify port (left-hand), meaning that the vessel should keep the mark to its left to remain in safe waters, or starboard (right-hand) meaning that the vessel should keep the mark to its right to remain in safe waters. The IALA has divided the world into two regions: Region A, which includes Europe, Australia, New Zealand, Africa, the Gulf and some Asian countries, and Region B, which is comprised of North, South and Central America, Japan, Korea and the Philippines (International Association of Marine Aids to Navigation and Lighthouse Authorities 2018). In Region A, port lateral marks are always coloured red and starboard lateral marks are always coloured green. The inverse is true in Region B. In our ontology we have declared that the `atln:StartboardLateralMarkRegionA` class is equivalent to a `atln:Beacon`, a `atln:Buoy`, a `atln:Mark` or a `atln:Turret` that has a `atln:hasLateralType` property pointing towards the `tdi:Starboard` instance of the `atln:LateralDirection` class, and a `atln:hasColour` property pointing towards the `tco:Green` instance of the `atln:hasColourType` class. This means that every entity that is declared as an instance of the `atln:StartboardLateralMarkRegionA` class will automatically be inferred as being of the colour green with a `atln:hasColour` property pointing towards the `tco:Green` instance. Inversely, every entity declared as being a `atln:Buoy`, for example, with a `atln:hasLateralType` property pointing towards the `tdi:Starboard` instance and a `atln:hasColour` property pointing towards the `tco:Green` instance, will automatically be inferred as being part of the `atln:StartboardLateralMarkRegionA` class. The OWL implementation of the class definition is shown in listing 4.4.7.

Another example that demonstrates the benefit of creating axioms is based on spatial relations that employ the cardinal directions. Such spatial relations are heavily relied upon in navigation because they are constructed using an absolute frame of reference, which means that no viewpoint is in-

"Warning: the 'Grand Pot de Beurre' turret (48° 37.22' N — 4° 36.47' W), a port lateral mark in the Grand Chenal, near to the alignment at 135.7°, can lead to confusion."

Extract 4.4.4 – Translated from the original French text: "Attention : la tourelle « Grand Pot de Beurre » (48° 37,22' N — 4° 36,47' W), marque latérale bâbord du Grand Chenal, proche de l'alignement à 135,7°, peut porter à confusion." (Shom 2021a, p. 413)

volved (Levinson 1996). It also means that their inverse relations can be calculated and expressed via a spatial relation employing the opposite cardinal direction. For each spatial relation based on a cardinal direction defined as an OWL ObjectProperty in our ontology, we declared its inverse property. An example of such a declaration for the OWL ObjectProperty `atln:isEastOf`, which was originally declared in lines 18 to 21 of listing 4.4.2 on page 91, can be seen on line 17 of listing 4.4.6: `atln:isWestOf`.

After having finished the refactoring process, we manually aligned our ontology with all of the relevant semantic resources cited in section 4.2.2 that have been published on the Web. Listing 4.4.8 shows our refactoring and alignment process for the `atln:Wind` class and listing 4.4.9 shows our refactoring and alignment process for the `atln:Island` named individual.

4.4.5.4 Testing Full Model and Exemplar Dataset

We carried out the model, data and query tests one final time in the same manner as described in section 4.4.4.5 and published the full final model [42] and the final exemplar dataset [43] on the Web.

4.5 Results

The ATLANTIS Ontology currently contains 110 classes, 90 object properties, 90 data properties and 2190 axioms in total. It is a seed ontology that has the structure and fundamental elements necessary to model the knowledge contained within the *Instructions nautiques* in a way that has been validated by domain experts.

4.6 Evaluation

4.6.1 Evaluation through Testing

We evaluated the ATLANTIS Ontology in two different ways. The first evaluation occurred during the testing phase of the development method-

42. https://github.com/umrlastig/atlantis-ontology/blob/main/atlantis_ontology.owl
43. https://github.com/umrlastig/atlantis-ontology/blob/main/triplets.ttls

```
atln:StarboardLateralMarkRegionA rdf:type owl:Class ;
    owl:equivalentClass [ owl:intersectionOf ( [ rdf:type owl:Class ;
                                                 owl:unionOf ( atln:Beacon
                                                               atln:Buoy
                                                               atln:Mark
                                                               atln:Turret
                                                             )
                                               ]
                                               [ rdf:type owl:Restriction ;
                                                 owl:onProperty
                                                     atln:hasLateralType ;
                                                 owl:hasValue tdi:Starboard
                                               ]
                                               [ rdf:type owl:Restriction ;
                                                 owl:onProperty atln:hasColour
                                                     ;
                                                 owl:hasValue tco:Green
                                               ]
                                             ) ;
                          rdf:type owl:Class
                        ] ;
    rdfs:subClassOf atln:LateralMark ,
                    [ rdf:type owl:Restriction ;
                      owl:onProperty atln:hasLateralType ;
                      owl:hasValue tdi:Starboard
                    ] ,
                    [ rdf:type owl:Restriction ;
                      owl:onProperty atln:hasColour ;
                      owl:hasValue tco:Green
                    ] ;
    rdfs:label "Starboard Lateral Mark Region A"@en .
```

Listing 4.4.7 – The declaration of the `atln:StartboardLateralMarkRegionA` OWL Class in Turtle syntax.

ology (see section 4.4.4.5). Thanks to the query test we validated the pertinence of our ontological model according to the possibility to transform the informal competency questions defined in section 4.4.3.3 into SPARQL queries that return the correct answers. This way of evaluating an ontology is suggested by Hogan et al. (2021) and is implemented in many other studies (Barbe et al. 2023; Mansfield et al. 2021; Sequeda et al. 2019).

4.6.2 Evaluation through Reuse

In 2022 we supervised the internship of Benoit and Kergus (2022), who carried out a study to determine whether the ATLANTIS Ontology could be used to model Sailing Directions other than the *Instructions nautiques*.

```
atln:Wind rdf:type owl:Class ;
    owl:equivalentClass <http://sweetontology.net/phenAtmoWind/Wind> ,
                        <https://bimerr.iot.linkeddata.es/def/weather#Wind> ;
    rdfs:subClassOf atln:MeteorologicalPhenomenon ;
    rdfs:label "Wind"@en ,
               "Vent"@fr .
```

Listing 4.4.8 – The alignment of an OWL Class with existing semantic resources in Turtle syntax.

```
tse:Island rdf:type owl:NamedIndividual ,
                    atln:TypeOfSpatialEntity ;
    skos:exactMatch <http://data.europa.eu/bkc/017.03.04.0500> ,
                    <http://data.ign.fr/id/codes/topo/typederelief/Ile> ,
                    <http://thesaurus.oieau.fr/thesaurus/page/ark:/99160/368b1
                       4ee-4cc3-4184-babd-e1dca13e802b> ,
                    <http://www.eionet.europa.eu/gemet/concept/4514> ;
    rdfs:comment "See http://www.ontotext.com/proton/protonext#Island and
       http://sweetontology.net/realmLandCoastal/Island"@en .
    skos:altLabel "isle"@en ;
    skos:prefLabel "island"@en ,
                   "ile"@fr .
```

Listing 4.4.9 – The alignment of an OWL NamedIndividual with existing semantic resources in Turtle syntax.

They covered three series of Sailing Directions, all written in English: the *United States Coast Pilot*, the *Sailing Directions (Enroute)* and the *Canadian Sailing Directions*. To carry out the study, Benoit and Kergus applied our ontology development methodology as described in section 4.3 to the three publications, putting into practice the dashed arrow to return to Step 1 from Step 4 as shown in figure 4.3.1 on page 55.

The first part of their study deals with comparing the content of the *Sailing Directions (Enroute)*, the *United States Coast Pilot* and the *Canadian Sailing Directions* with that of the *Instructions nautiques*. This constitutes Step 1: Groundwork of our methodology and their findings are as follows.

The US *Sailing Directions (Enroute)*, which only cover non-domestic coastal waters, are found by Benoit and Kergus to have less detailed descriptions of coastlines and maritime navigation guidelines for French coastal waters than the *Instructions nautiques*, probably owing to the vastness of the geographic extent that they must cover. For example, the description of the Île de Batz and its associated channels, ports and dangers covers two pages in the *Instructions nautiques* (Shom 2021a) and only a quarter of a page in the *Sailing Directions (Enroute)* (National Geospatial-Intelligence Agency 2022). Whilst the type of information and the way in

which it is presented is similar in both publications, there is one piece of information that features in the *Sailing Directions (Enroute)* but that cannot be found in the *Instructions nautiques*: the Île de Batz is radar conspicuous.

Like the *Sailing Directions (Enroute)*, the *United States Coast Pilot* and the *Canadian Sailing Directions*, which only cover their respective domestic coastal waters, have a similar form and content to the *Instructions nautiques*. The *United States Coast Pilot* and the *Canadian Sailing Directions* have richer descriptions than the *Sailing Directions (Enroute)*, although they remain less detailed than those given in the *Instructions nautiques*. For example, they give few or no guidelines on typical routes or the best routes to take, and how to navigate them. Benoit and Kergus indicate that the *United States Coast Pilot* contain details about the history of certain spatial entities, for example after whom they have been named, which is not done in the *Instructions nautiques*. Like the *Sailing Directions (Enroute)*, the *United States Coast Pilot* contain indications about the radar visibility of spatial entities.

The larger and more diverse geographic coverage of these other Sailing Directions means that they cover parts of the ocean exposed to climate conditions unseen in the *Instructions nautiques*. Benoit and Kergus identified in particular that, unlike the *Instructions nautiques*, other Sailing Directions deal with icy conditions and parts of the ocean that can be iced over for some or all of the year.

After having analysed the content of the different series of Sailing Directions, Benoit and Kergus proceed to Step 2: Producing Documentation. They revise the motivating scenarios and glossaries, noting in particular the required information about radar visibility and ice conditions, add relevant extracts to the motivating scenarios and corresponding questions to the lists of informal competency questions.

To carry out Step 3: Structuring, Implementing and Testing Subdomain Models, Benoit and Kergus conceptualise and implement classes and properties in OWL that allow modelling the additional concepts and relations identified during Step 1. For some concepts and relations it was necessary to create new classes or properties whilst for others it was possible to adapt the definitions of existing ones to encompass the new needs. Listing 4.6.1 shows the declaration of a new property, `atln:hasRadarVisibility`, that allows storing a string that describes the radar visibility of a spatial entity. Listing 4.6.2 shows the declaration of a property that has been modified to encompass new needs. The `atln:hasTarget` property on line 17 was originally created to indicate the type of vessel for which a maritime navigation

guideline applies (see definition of 'Target' in document 4.4.4). Benoit and Kergus required a similar property to indicate the type of vessel for which a stopping place is suited and therefore decided to expand the declaration of the domain of the `atln:hasTarget` property to include stopping places contained within the `atln:StoppingPlace` class as well as maritime navigation guidelines contained within the `atln:NavigationGuideline` class as can be seen in lines 19 to 23 in listing 4.6.2. Benoit and Kergus expanded the datasets accordingly by adding RDF triples that represent the knowledge contained within the new extracts in the motivating scenarios. Finally, they carry out the three types of test to verify and validate the coherence of the models and the datasets, and their ability to provide correct answers to the competency questions.

```
1 atln:hasRadarVisibility rdf:type owl:DatatypeProperty ;
2     rdfs:subPropertyOf atln:hasCharacteristic ;
3     rdfs:domain atln:SpatialEntity ;
4     rdfs:range xsd:string ;
5     rdfs:label "has radar visibility"@en .
```

Listing 4.6.1 – The OWL implementation in Turtle syntax of a data property that allows indicating the visibility of a spatial entity on a radar.

Benoit and Kergus then accomplish Step 4: Merging, Refactoring and Aligning by merging the new versions of the subdomain models and the subdomain datasets to create the new full model and the new exemplar dataset.

The results of their study show that the ATLANTIS Ontology is capable of modelling most of the information contained within the other Sailing Directions studied, but that some differences in content between the *Instructions nautiques* and other Sailing Directions mean that ATLANTIS would require some minor additions to be able to fully represent them. Their work also demonstrates that our ontology development methodology is repeatable and that it can be applied to manually enrich existing ontologies as well as creating ontologies from scratch.

In total, Benoit and Kergus suggest the addition of 30 classes and 13 properties to generalise the ATLANTIS Ontology to the content of Sailing Directions other than the *Instructions nautiques*. This international version of the ATLANTIS Ontology can be freely consulted at an online repository [44].

44. https://github.com/umrlastig/atlantis-ontology/tree/main/atlantis_in_english

```
atln:StoppingPlace rdf:type owl:Class ;
    owl:equivalentClass [ rdf:type owl:Restriction ;
                          owl:onProperty atln:hasTypeOfStoppingPlace ;
                          owl:someValuesFrom atln:TypeOfStoppingPlace
                        ] ;
    rdfs:subClassOf atln:SpatialEntity ;
    rdfs:label "stopping place"@en .

atln:TypeOfStoppingPlace rdf:type owl:Class ;
    owl:equivalentClass [ rdf:type owl:Restriction ;
                          owl:onProperty skos:inScheme ;
                          owl:hasValue <http://data.shom.fr/id/codes/atlantis/
                              typeofstoppingplace/list>
                        ] ;
    rdfs:subClassOf atln:TypeOfSpatialEntity ;
    rdfs:label "Type of Stopping Place"@en .

atln:hasTarget rdf:type owl:ObjectProperty ;
    rdfs:subPropertyOf atln:hasSpecification ;
    rdfs:domain [ rdf:type owl:Class ;
                  owl:unionOf ( atln:NavigationGuideline
                                atln:StoppingPlace
                              )
                ] ;
    rdfs:range [ rdf:type owl:Restriction ;
                 owl:onProperty skos:inScheme ;
                 owl:hasValue <http://data.shom.fr/id/codes/atlantis/typeofves
                     sel/list>
               ] ;
    rdfs:label "has target"@en .
```

Listing 4.6.2 – The declaration in Turtle syntax of a SKOS thesaurus to store the different types of spatial entities that are stopping places, and the new object property that indicates the target vessel type for navigation guidelines or for stopping places.

4.7 Conclusion

In this chapter we first presented a thorough analysis of existing maritime ontologies and reviewed 10 well-established domain ontology development methodologies. Combined with the characteristics of our corpus and application, this led us to identify a need to create a new ontology development methodology that is suited to being used in situations in which the aim is to manually develop a domain ontology that represents the content of a textual corpus combined with the knowledge of domain experts. In section 4.3 we gave a step-by-step presentation of our new ontology development methodology, which is based on elements from SAMOD (Peroni 2016a), MOMo (Shimizu et al. 2022) and NeOn (Suárez-Figueroa et al.

2012). One of the main characteristics of our methodology is the creation of the preliminary datasets of semantic triples at the very beginning of the development process. This allows us to become more familiar with the corpus content as well as its organisation before beginning the formal implementation. In addition, having a better understanding of the corpus makes time spent with domain experts more efficient and beneficial because we are able to orient our questions more precisely. Our domain ontology development methodology makes up the first component of the ATONTE Methodology.

We demonstrated how we implemented our ontology development methodology to create the ATLANTIS Ontology, a geospatial seed ontology that covers the domain of the *Instructions nautiques*. During the ontology development process we carried out a series of interviews with users and producers of the *Instructions nautiques*. In this chapter we gave a comprehensive report of the findings of these interviews, which provide insights on user practices and desires concerning Sailing Directions. All the documentation, datasets and formal models linked to the ATLANTIS Ontology have been published on the Web [45].

The ATLANTIS Ontology could be improved by adding formal descriptions of more concepts to be able to generate more new knowledge via reasoning. This could be implemented once the ontology application has been better defined to be able to target the specific needs.

The next chapter presents the second component that makes up the ATONTE Methodology, which is dedicated to the extraction of geographic information from text. The extracted information will eventually be used to populate a knowledge graph according to the ontological model.

45. https://github.com/umrlastig/atlantis-ontology

Chapter 5

Entity and Relation Extraction from Text

5.1 Introduction

In this chapter we present the second component of the ATlantis Ontology and kNowledge graph development from Texts and Experts (ATONTE) Methodology, which deals with the automatic extraction of geographic information from text. We demonstrate how we implemented it to automatically extract spatial entities and relations from the text of the *Instructions nautiques*. The extracted information will be structured in order to populate the ontology and create a knowledge graph, as demonstrated in chapter 6.

The full domain of the coAsTaL mAritime NavigaTion InstructionS (ATLANTIS) Ontology is vast, which is why we chose to work on only one subdomain: Maritime Spatial Entities and Spatial Relations, for our application of this approach. We therefore focus on the automatic extraction of spatial entities and spatial relations from text. This involves automatically identifying mentions of spatial entities in the text, and capturing any information that describes the spatial relations between them.

During knowledge graph population, spatial entities become instances of ontological classes that represent their physical type and spatial relations become assertions of their corresponding object properties. To be able to correctly assign spatial entities to their corresponding ontological class, it is necessary to know their *type*. We make the assumption that the geographic name of a spatial entity often contains a common noun that indicates its type, such as *port* in "Port of Liverpool". Although this holds for many of the Romance languages, it is not applicable to all languages. Sometimes, the geographic name of a spatial entity contains more than one type noun, such as in "Robben Island Lighthouse". This increases the complexity of identifying a spatial entity's true type from its geographic name.

Whilst *flat* spatial entity extraction would simply aim to capture "Port of Liverpool" or "Robben Island Lighthouse" as the name of a spatial entity without seeking any further definitions, *nested* spatial entity extraction allows defining multiple layers of labels for the same text. We use the labels introduced by Moncla (2015), the definitions of which are as follows: geographic feature refers to common nouns that represent *types* of spatial entities, name refers to pure proper nouns and geographic name refers to the full name associated with a geographic feature. For our first example, *nested* spatial entity extraction would therefore aim to capture "Port of Liverpool" as the geographic name, "Port" as the geographic feature and "Liverpool" as the name. For our second example, nested spatial entity extraction would aim to capture initially "Robben Island" as one geographic name, with "Island" as the geographic feature and "Robben" as the name, and then also capture "Robben Island Lighthouse" as another geographic name, with "Lighthouse" as the geographic feature. This layered approach facilitates the identification of the correct type in cases where the geographic name contains multiple instances of a geographic feature.

By extracting *nested* as opposed to *flat* spatial entities, the geographic feature type of the entity is captured and an instance of the right ontology class can be created automatically. In some cases, its name gives an indication of its geographical location via an indirect spatial reference. These two extra pieces of information, the entity *type* and its *name*, facilitate the disambiguation task of linking the instance to the correct entry in a reference geographic resource (Southall et al. 2011).

By extracting the spatial relations between entities, assertions of object properties can automatically be created between instances. This information can also be used to aid disambiguation of named and unnamed entities, and increase confidence in the results thanks to spatial reasoning (Paris et al. 2017). In the case where a corresponding reference entry does not yet exist, a new entry can be created in the geographic resource, supported by the class of the instance and certain property information (depending on the fields present in the resource). The same reasoning applies to the creation and enrichment of gazetteers: by specifically identifying entity types during the extraction process, gazetteer entries can automatically be classed or assigned attributes and can more easily be disambiguated. The identification and extraction of the spatial relations in which spatial entities take part can increase the level of detail available in descriptions of gazetteer entries and their locations.

There are three different types of spatial relations: topological, directional and distance (Brageul and Guesgen 2007). Topological spatial rela-

"A fishing port lies 5.7 M to the ENE of Ras Magroua."

Extract 5.1.1 – Translated from the original French text: "Un port de pêche est établi à 5,7 M à l'ENE de Ras Magroua." (Shom 2021d, p. 181)

tions describe the inherent properties between objects, which can be disjoint (the garden is disjoint from (does not touch) the road), touching (the edge of the garden is the edge of the bike path) or enclosing (the bike is in the garden). Directional spatial relations describe the relative position of objects according to a specified frame of reference. The frame of reference can be intrinsic, relative (also known as deictic) or absolute (also known as extrinsic) (Shusterman and P. Li 2016; Brageul and Guesgen 2007; Dokic and Pacherie 2006; Levinson 1996). An intrinsic frame of reference is defined by inherent properties of the reference object, such as its front or back. A relative frame of reference is defined by the point of view used to describe the spatial relation. In the *Instructions nautiques*, descriptions of the coastline are always given from the point of view of a vessel on the water looking towards the coast. An absolute frame of reference is defined by an external system, such as the cardinal directions. Distance spatial relations describe how far objects are from one another according to a frame of reference that is composed of a distance system, a scale system and a type (Clementini et al. 1997). Our approach focuses on topological and directional spatial relations.

Texts that cover an international environment are likely to contain geographic names in languages other than the main language of the text, such as in extract 5.1.1. Although the original main text is written in French, the geographic name includes a geographic feature type written in romanised Arabic: "Ras", which means *cape*. It is important that this does not hinder the extraction process: geographic names and geographic feature types written in other languages should still be identified as such. The state of the art in information extraction from text relies on deep neural network language models (Nasar et al. 2021). Such models can be trained to deal with one or multiple languages and are referred to as pretrained *monolingual* or *multilingual* language models respectively. A multilingual ontology can then be used to aid the disambiguation of entities whose type is written in other languages (Stadler et al. 2012).

In section 5.2 we first give an overview of artificial neural networks (ANNs) (section 5.2.1) before reviewing related work in terms of flat and nested entity extraction and relation extraction from text (section 5.2.2) and in terms of comparisons of monolingual and multilingual language

models (section 5.2.3). Section 5.3 is dedicated to presenting the second component of ATONTE, which consists of an approach for the automatic extraction of nested entities and binary relations from text. We also demonstrate how we applied our approach to automatically extract nested spatial entities and binary spatial relations from the *Instructions nautiques*. In section 5.4 we present our results for the extraction task, which we then evaluate before concluding in section 5.5.

5.2 Related Work

In this section we give an overview of ANN, we review the current state of the art in flat and nested entity extraction and relation extraction from text, and finally examine studies that compare the performance of monolingual language models with that of multilingual ones.

In their infancy, entity and relation extraction from text were primarily performed using rule-based approaches that required manually developing rules built on grammar, syntax and punctuation to identify them. Such approaches have been combined with the use of dictionaries and successfully applied to the extraction of spatial entities from texts that describe the natural environment (Lamotte 2019; Moncla 2015). Classical machine learning approaches were then developed and achieved consistently higher performances in these tasks, and a trend away from rule-based approaches was documented by Nadeau and Sekine (2007). Research in both tasks is now dominated by approaches that apply deep learning techniques, which is why we only consider such techniques for review, although slower progress is being made in relation extraction (Nasar et al. 2021; J. Li et al. 2020; Yadav and Bethard 2018). Given that the tasks of entity extraction and relation extraction are not always studied together, we review work that deals with either task or both. We consider approaches designed for generic entities and relations, the suitability of which to spatial entities and relations would need to be verified, as well as those dedicated to spatial entities and relations, which are much less common.

5.2.1 Artificial Neural Networks

Before reviewing individual approaches to entity and relation extraction from text that use deep learning techniques, we will give a brief overview of ANN, on which deep learning is based. ANN are a type of machine learning model that are constructed in a similar way to animal brains (Krogh 2008). They are composed of a set of connected artificial neurons, mirroring animal brain neurons that are connected by synapses. The connections

in an ANN can transmit signals between artificial neurons, and these connections can be weighted to modify the strength of the signals that pass through it. An artificial neuron is activated if the sum of the signals that it receives is above a threshold value. When it is activated, an artificial neuron processes the sum of signals and transfers the resulting signal to its neighbouring neurons (Zou et al. 2009).

ANN can be *trained* either through a supervised learning process or an unsupervised one (Zou et al. 2009). Supervised learning consists of providing the network with a training dataset: a set of inputs and their corresponding outputs. The network processes the training inputs and adjusts the connection weights (from their initial random values) to minimise the difference between the output obtained and the training outputs. The weights that minimised this difference and therefore produced the best output are fixed and the trained network can be used to predict outputs on new inputs. Unsupervised learning requires providing the network only with a set of inputs, without their corresponding outputs. In this case, the network attempts to find patterns in the set of input data by adjusting the connection weights to minimise a function defined according to the application. The weights that minimised this function and therefore produced the best output are fixed and the training process is complete. This initial ANN training process is known as *pretraining*, and a pretrained ANN is called a *model*. A pretrained model can either be used directly (as an off-the-shelf solution) to predict outputs on new inputs, or it can be used as a starting point to develop a model better suited to a specific task or dataset. The further training of a pretrained model in this way, using a domain- and/or task-specific training dataset, is known as *fine-tuning*. The performance of a network can be evaluated by providing it with a test dataset, a set of inputs and their corresponding outputs, just like the training dataset but containing unseen data. The network predicts outputs from the inputs in the test dataset and the predicted outputs are compared with the outputs in the test dataset.

The two main types of ANN are feedforward neural network (FNN) and recurrent neural network (RNN). FNN are uni-directional forward-flowing networks, meaning that the information flows through the network from the input, through the neurons and to the output without ever going backwards (Fine 1999). RNN, on the other hand, are bi-directional networks in which information can flow forwards and backwards, meaning that the information can flow through loops and pass more than once through the same nodes (Jain and Medsker 1999). This characteristic allows RNN, unlike FNN, to retain information that has already passed through them to

influence the processing later on.

Long short-term memory (LSTM) networks are a type of RNN that have an extra feature that allow them to decide which information to retain and which information to discard (Greff et al. 2017). A bidirectional long short-term memory (BiLSTM) network consists of a pair of LSTM networks working together: one processes the input forwards through the network and the other processes the input backwards (Siami-Namini et al. 2019). This allows the network to have information from previous steps and from future steps whilst processing the current step.

A gated recurrent unit (GRU) is also a type of RNN that is very similar to a LSTM network but has fewer parameters (Dey and Salem 2017). GRU are therefore quicker to train than LSTM networks but also generally give inferior results, although this is application-dependent. A bidirectional gated recurrent unit (BiGRU) is like a BiLSTM but composed of two GRU instead of two LSTM networks.

Transformer networks (Vaswani et al. 2017) are similar to RNN in that they can retain information, but instead of processing the input sequentially like an RNN does, a Transformer network is able to process the input in parallel. A bidirectional Transformer architecture was used to create the Bidirectional Encoder Representations from Transformers (BERT) pretrained language model (Devlin et al. 2019), which takes text as input. BERT was initially trained on an English-language corpus of 3,300M words for natural language processing (NLP) tasks such as word or token classification, text prediction, question answering and sentiment analysis.

The inputs and outputs in supervised learning training datasets and test datasets often take the form of a piece of data (input) and the tag corresponding to the piece of data (output), from a pre-defined set of tags. For example, in a text-based dataset for a sentiment analysis task with a set of two tags: *positive* and *negative*, the input could be a sentence such as "That was a great match!" and the output would be the tag that corresponds to the sentiment expressed in the sentence, in this case the *positive* tag. For a word classification task with a set of four tags: *event*, *organisation*, *person* and *other*, the input could be the same sentence and the output would consist of the *other* tag for the first four words in the input sentence and the *event* tag for the fifth word in the input sentence. Tags can also be referred to as annotations, labels or classes. A dataset complete with tags is known as an annotated dataset.

5.2.2 Entity and Relation Extraction

5.2.2.1 Flat Spatial Entity Extraction

Current work on flat spatial entity extraction is dominated by the use of BiLSTM and Transformer (Vaswani et al. 2017) models. Unless explicitly stated otherwise, all approaches presented here aim to identify flat spatial entities in text as locations or place names without identifying their type.

Berragan et al. (2023) develop and train five novel BiLSTM- and BERT-based models specifically for the extraction of flat named spatial entities from English-language text. They train their models via supervised learning on a dataset of Wikipedia extracts in CoNLL-03 NER format that they manually annotated with the BIOUL tagging scheme using the Doccano tool [1]. All five of their models that they train on this dataset are shown to outperform three off-the-shelf models for entity extraction: Stanza [2] (Qi et al. 2020), spaCy (large) [3] and spaCy (small) [4], on a test dataset also composed of Wikipedia extracts.

A new approach for flat spatial entity extraction from English-language text is also presented by X. Hu et al. (2022), for named as well as unnamed spatial entities. They train a LSTM-based ANN on worldwide place names from two gazetteers: OSMNames [5] and GeoNames [6]. Their approach consists of using this model to select candidate place names from text and then using two pretrained BERT models to identify whether or not the candidates are indeed place names. They test their approach on 19 different public tweet datasets and compare its performance with that of 11 competing approaches, including Google NLP [7], DBpedia Spotlight [8] and Stanza. The performance of the approach developed by X. Hu et al. is shown to give better results than each of the 11 other approaches on all 19 of the datasets.

Tao et al. (2022) present a new model for flat named spatial entity extraction from Chinese text composed of a BERT-based model that feeds into a BiLSTM network. They also introduce a dataset of 2 million Chinese named spatial entities extracted from two encyclopaedias that they annotate with a set of seven coarse-grained tags such as *residential land and facilities* ("a place where human beings live or engage in productive life") and *landforms* ("includes natural and artificial landforms"). The an-

1. https://doccano.github.io/doccano/
2. https://stanfordnlp.github.io/stanza/
3. https://spacy.io/models/en#en_core_web_lg
4. https://spacy.io/models/en#en_core_web_sm
5. https://osmnames.org/
6. https://www.geonames.org/
7. https://cloud.google.com/natural-language/
8. https://www.dbpedia-spotlight.org/

notation is carried out using ChineseNERAnno [9], an automatic annotation tool with a manual validation step that they developed during this project. Tao et al. train and test their model on their own annotated dataset. They evaluate its performance on this dataset and on three other public datasets, and show that it outperforms 12 other models.

5.2.2.2 Nested Generic Entity Extraction

Nested entity extraction has been tackled using different methods that include layered approaches and joint labelling in particular.

Layered approaches consist of training one model for each label that can appear in the nested entity structure, effectively resulting in stacked flat entity extraction modules. The models can be trained independently, thereby avoiding error propagation, or information can be passed between them to improve context representations. This is the case in the model presented by Ju et al. (2018), in which fine-grained entities are extracted first and each subsequent layer extracts more complex entities, using information encoded in previous layers, until no more entities are found.

To avoid training multiple models, the joint labelling approach presented by Agrawal et al. (2022) can be used. It requires training only one model as all the labels that correspond to a single token at different levels of nesting are concatenated to form one label. For example, using the three labels that we introduced in section 5.1, the three words in the spatial entity name "Port of Liverpool" would each be tagged with just one label: *geographic_feature+geographic_name*, *geographic_name* and *name+geographic_name* respectively. This technique has been improved by Tual et al. (2023) by the addition of class-based weights in the loss function that penalise semantically-distant classes more severely.

5.2.2.3 Binary Spatial Relation Extraction

Spatial relation extraction has been little studied independently of spatial entity extraction.

A convolutional neural network (CNN) (a type of FNN) modified to deal specifically with spatial relations in Chinese is presented by Qiu et al. (2022). The training and test datasets are composed of sentences that contain exactly two spatial entities and exactly one spatial relation each. The spatial relations are classed as *topological*, *distance* or *directional*. Qiu et al. generate their datasets by first manually annotating a small number of sentences and storing them in a knowledge base, and then using them to automatically identify and label new sentences in a corpus of Wikipedia.

9. https://github.com/guojson/ChineseNERAnno

To improve the semantic and contextual understanding of each sentence by the network, Yang et al. (2022) present a method for spatial relation extraction from Chinese text that combines a Transformer model with a BiGRU and an attention mechanism. As with the previous study, it has the disadvantage of relying on training and test datasets composed of sentences that contain exactly one spatial relation between exactly one pair of spatial entities.

5.2.2.4 Combined Entity and Relation Extraction

Deep learning approaches that aim to tackle both entity and relation extraction can either separate the two tasks and dedicate an independent neural network to each, keep the two tasks separate with dedicated neural networks but allow information to be shared between them, or model the two tasks together and have a single neural network perform both tasks together. We refer to the former as a *pipelined* approach whilst the latter two are known as *joint modelling* approaches. It is shown by Wu et al. (2022) that for spatial entity and relation extraction, a pipelined approach is the most effective of the two. This is attributed to the fact that the detection of spatial entities benefits from context that is different to the context that benefits the identification of spatial relations, meaning that sharing the same context between both tasks less beneficial than allowing each task to calculate its own context.

A pipelined approach for generic flat entity and generic binary relation extraction from English text is presented by Zhong and D. Chen (2021). Known as the Princeton University Relation Extraction system (PURE), it trains two separate encoders, one for each task, from existing pretrained deep language models. It is shown that cross-sentence information should be taken into account during the training of both the entity model and the relation model, as well as during the prediction phases, to maximise results. This pipelined approach with cross-sentence context outperforms joint modelling systems on three standard benchmark datasets: ACE04, ACE05 and SciERC using BERT models.

Flat spatial entity and binary spatial relation extraction from French text are carried out using a pipelined approach by Cadorel et al. (2021). A BiLSTM neural network coupled with a BERT model is trained for spatial entity extraction whilst spatial relation extraction is based on a dependency parsing method using a Stanza BiLSTM model. Cadorel et al. created a French-language dataset of online housing advertisements and manually annotated it using the BIESO tagging scheme to train and test their model. Their annotation scheme includes flat named spatial enti-

ties, unnamed spatial entities, unclassified topological spatial relations and distance relations. They experiment with combinations of different models and and find that the most successful is an approach that combines Flair [10], the pretrained French-language BERT model CamemBERT [11] and Word2Vec [12].

5.2.3 Comparing Monolingual and Multilingual Models

Many experiments have been done to determine whether the multilingual BERT language model first presented by Devlin et al. (2019) performs better than monolingual language models for monolingual texts. Rust et al. (2021) conducted such experiments for Arabic, English, Finnish, Indonesian, Japanese, Korean, Russian, Turkish and Chinese, Martin et al. (2020) for French, Delobelle et al. (2020) for Dutch, To et al. (2021) for Vietnamese and Velankar et al. (2022) for Marathi. All of these studies show better results using the monolingual language models for most if not all of the tasks evaluated.

5.2.4 Summary

Whilst nested generic entity extraction has been widely studied, there is little work specific to nested spatial entities. Although experiments have been carried out to explore and improve upon binary spatial relation extraction, there is too much reliance on datasets that are not representative of the corpora from which we may really want to extract geographic information.

We chose to implement PURE (Zhong and D. Chen 2021) thanks to it dealing with both entities and relations, thanks to its pipelined approach and thanks to the availability of the code. PURE is made for *flat*, *generic* entity extraction and *generic* binary relation extraction but we wish to adapt it to *nested*, *spatial* entities and the binary *spatial* relations between them. The authors of PURE do not attempt nested entity extraction nor do they specifically target spatial entities and relations.

Our review shows that monolingual BERT language models perform better than the multilingual model on monolingual texts. We decided to investigate the effects of a dataset containing words in multiple languages other than the main language of the text, as is the case in our dataset, on the performances of monolingual and multilingual models. This is espe-

10. https://github.com/flairNLP/flair
11. https://camembert-model.fr/
12. https://www.tensorflow.org/text/tutorials/word2vec

cially important given that the words in other languages are almost always part of a spatial entity name in our corpus.

5.3 Our Approach

In this section we demonstrate how we implemented PURE (Zhong and D. Chen 2021) on our corpus of *Instructions nautiques* and present the modification we made to adapt it to the extraction of *nested* as opposed to *flat* entities. We first discuss the way in which we prepared our annotated training and test datasets before explaining the different experiments that we carried out during the model training.

5.3.1 Dataset Preparation

Our dataset is made up of extracts from each one of the 15 volumes of the *Instructions nautiques* that we had at our disposal. The volumes, which are written in French, cover coastal areas in Africa, Europe, North and South America, as well as in the Indian and Pacific Oceans. We extracted the text from the PDF documents using pdfminer.six [13]. The resulting plain text files then needed to be cleaned because of excess empty lines and double spaces, which we did using regular expressions.

We annotated our dataset by hand using the brat rapid annotation tool [14], which allows creating nested labelled annotations and creating directed labelled links between them (Stenetorp et al. 2012). The source text was split on whitespace by brat, giving a dataset of 101,400 tokens.

Given that we wished to perform *nested* spatial entity extraction to simultaneously capture the full name of the spatial entity as well as its type and name, we implemented a nested labelling approach. We use the labels introduced by Moncla (2015), which we presented in section 5.1. Their definitions are as follows:

— geographic feature (geogFeat) refers to common nouns that represent *types* of spatial entities

— name refers to pure proper nouns

— geographic name (geogName) refers to the full name associated with a geographic feature

A token can be annotated with multiple labels, and labels can be associated to one token or span multiple tokens. We define that any token can be annotated with zero or one geographic feature or name label. A token

13. https://github.com/pdfminer/pdfminer.six
14. http://brat.nlplab.org

cannot be annotated with both a geographic feature and a name label. A token cannot be annotated only with a name label. A token annotated with a name label must also be annotated one or more times with a geographic name label. Any token, already labelled or not, can be annotated zero or more times with a geographic name label.

An extremely large number of different topological and directional spatial relations are used in our corpus so we decided to limit ourselves to extracting only those that would be the most useful during the disambiguation process.

The cardinal directions are heavily relied upon in navigation because spatial relations that employ them are constructed using an absolute frame of reference, which means that no viewpoint is involved (Levinson 1996). We chose to extract the directional spatial relations that employ the cardinal directions because of their frequent use and unambiguity. This amounts to 16 relation types in total: four that use the cardinal directions (N, E, S, W), four that use the intercardinal directions (NE, SE, SW, NW) and eight that use the secondary intercardinal directions (NNE, ENE, ESE, SSE, SSW, WSW, WNW, NNW). In our corpus, these spatial relations are always referred to by using these 16 one-, two- and three-letter abbreviations, for example "*le port est au NW de la ville*" ("the port is to the NW of the town") or "*la tour est à l'ESE du château*" ("the tower is to the ESE of the castle"). The 16 labels that correspond to these spatial relations are of the format "is XYZ of", where "XYZ" is one of the 16 cardinal direction abbreviations.

We identified three common topological spatial relations to capture more information about domain-specific spatial entities that are often unnamed in the corpus or are likely to be absent from reference geographic resources such as navigation marks (buoys, beacons, etc.), rocks or sandbanks. First, the "is off the coast of" label is used when it is indicated that one spatial entity is located off the coast of, or in the coastal waters of, another. This disjoint topological spatial relation is therefore frequently used to locate isolated spatial entities. This type of spatial relation, which is also constructed using an absolute frame of reference, is always referred to using the same three words in our corpus: "*au large de*" ("is off the coast of"). Second, the "is marked by" label is used for any spatial entity that is marked or pointed out by another deliberately-placed entity, often a navigation mark, either when the former poses a danger to navigators or when it allows a safe passage: "*Son musoir est marqué par un feu.*" ("Its pierhead is marked by a light.") (Shom 2021d). This disjoint topological spatial relation indicates a proximity between the two entities and is expressed in

a number of different ways in our corpus: "*est marqué par*" ("is marked by") can alternatively be expressed as "*est signalé par*" ("is flagged by") or "*est indiqué par*" ("is indicated by"). Third, the "is an element of" label indicates a topological relation that includes entities that are situated *on* or *in* another such that a bird's eye view shows the spatial footprint of one as being within or partly within the other. This enclosing topological spatial relation is expressed in a wide variety of different ways in our corpus, including implicitly, and rarely includes the word "*élément*" ("element"). For example, "*l'île porte un phare*" ("the island boasts a lighthouse"), "*le feu établi sur le quai*" ("the light located on the quay") and "*les haut-fonds de la baie*" ("the sandbanks of the bay") all indicate a "is an element of" relation.

All relation annotations must link two entity annotations, either geographic feature or geographic name labels. All relation annotations must have a direction. Instead of duplicating the relation labels to account for their inverses and create directed relation annotations that always go in the direction of the text ("A →is marked by→ B" and "C →marks→ D"), we created one version for each label and allow directed relation annotations that go in either direction ("A →is marked by→ B" and "C ←is marked by← D"). This has the advantage of reducing the number of labels in the annotation scheme and therefore facilitating the manual annotation process.

After having annotated one section from each of the 15 volumes of the *Instructions nautiques*, our dataset was considerably lacking in some relation labels, in particular those using the secondary intercardinal directions. This became evident during the initial tests that we carried out when training our models. To increase the number of examples of these relations we semi-automatically extracted random sentences containing the keywords NNE, ENE, ESE, SSE, SSW, WSW, WNW and NNW from the remaining text of each volume, manually annotated them and added them to our dataset.

Figure 5.3.1 shows how our nested spatial entity and spatial relation annotation scheme was applied to extract 5.1.1. The specific labelling of the geographic feature "Ras" within the geographic name combined with multilingual label values in an ontology means that this spatial entity could automatically be instantiated in the correct class regardless of the language in which the geographic feature is written, which in this case is romanised Arabic. Figure 5.3.2 shows an extract of our ontology, in which the instance `tse:Cape` has multilingual labels. Listing 5.3.1 shows a set of Resource Description Framework (RDF) triples that could automatically be

constructed from the information extracted from this sentence by following the ontological model presented in figure 5.3.2.

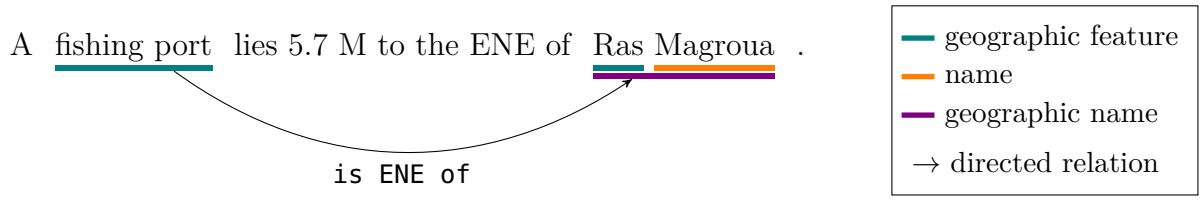


Figure 5.3.1 – Extract 5.1.1 annotated according to our nested spatial entity and binary spatial relation annotation scheme.

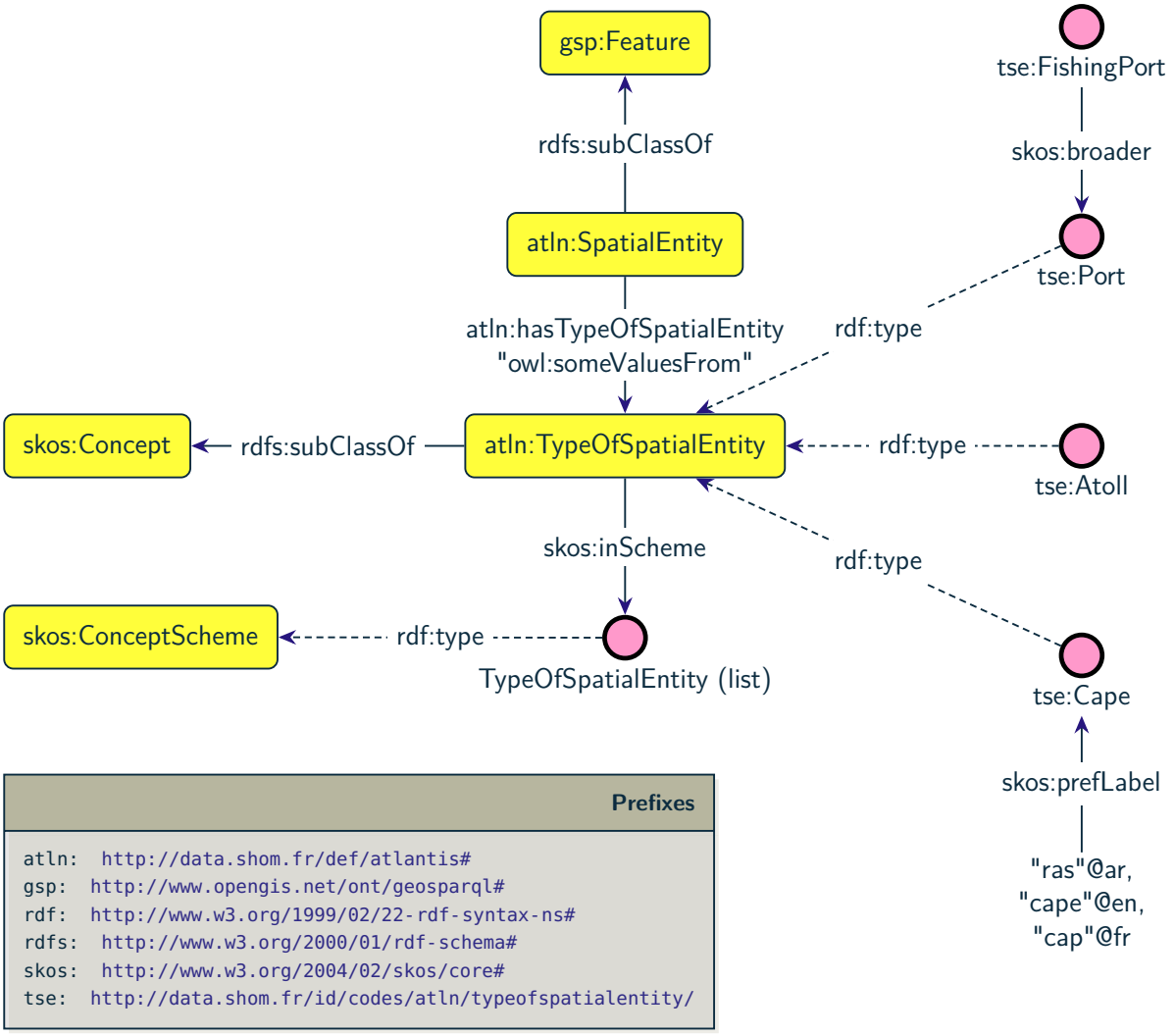


Figure 5.3.2 – Extract of the ATLANTIS ontology (Rawsthorne et al. 2022a) presented as a Graffoo diagram (Peroni 2013).

We split our annotated dataset into three parts: train, development and test, aiming to keep a 80:10:10 ratio of overall number of tokens and of numbers of entity labels. We also ensured homogeneity over the three parts of the dataset in terms of the geographic areas covered, the authors of the text and the types of content (linear description of the coast, description of a bay, description of an island or a group of islands, etc.). Our dataset of

```
ent:0001 atln:hasTypeOfSpatialEntity tse:FishingPort ; # entity number 1 is a
    fishing port
    atln:isENEof ent:0002 . # entity number 1 is ENE of entity number 2

ent:0002 atln:hasTypeOfSpatialEntity tse:Cape ; # entity number 2 is a cape
    rdfs:label "Ras Magroua" . # entity number 2 is called "Ras Magroua"
```

Listing 5.3.1 – RDF triples written in Turtle syntax constructed from the information annotated in figure 5.3.1 according to the ontological model presented in figure 5.3.2.

101,400 tokens contains 16,777 entity labels (which can span one or more tokens) and 3,051 relation labels (which connect exactly two entity labels in a given direction). In total, 18,030 tokens are annotated with at least one entity label, which corresponds to almost one in five tokens. We will refer to these manual annotations in our dataset as *gold annotations*. The dataset composition is summarised in table 5.3.1 and table 5.3.2 gives the label distribution. This dataset has been published as the ATLANTIS Dataset [15] under the Etalab Open License Version 2.0. We converted our dataset from the brat standoff format to the JSON Lines (JSONL) [16] format required for PURE using a Python script [17]. Listing 5.3.2 shows two annotated sentences converted to a JSON value in this format. The first of the two sentences is the one displayed in figure 5.3.1. In our case, each JSON value corresponds to one paragraph and contains a list of sentences (each of which is a list of tokens, see line 3), a list of label and span combinations that correspond to the entity annotations (boundary token pair + label, see line 4), and a list of label and span pair combinations that correspond to the relation annotations (ordered pair of boundary token pairs + label, see line 5). This nested annotation format that allows any token to be annotated with zero or more labels makes it possible to perform nested entity extraction without using joint labelling.

	Train nb.	Dev nb.	Test nb.	Total nb.
Tokens	83,851	8,156	9,393	101,400
Unlabelled tokens	69,200	6,507	7,663	83,370
Spatial entity-labelled tokens	14,651	1,649	1,730	18,030
Spatial entity labels	13,582	1,476	1,719	16,777
Spatial relation labels	2,507	222	322	3,051

Table 5.3.1 – Number of tokens and labels per split in the dataset. A single entity label can span one or more tokens.

15. https://github.com/umrlastig/atlantis-dataset
16. A JSONL file contains one valid JSON value on each line.
17. https://github.com/dwadden/dygiepp/blob/master/scripts/new-dataset/brat_to_input.py

Label	Train nb.	Dev nb.	Test nb.	Total nb.
all entity labels	13,582	1,476	1,719	16,777
geographic feature	6,602	692	801	8,095
name	3,486	391	462	4,339
geographic name	3,494	393	456	4,343
all relation labels	2,507	222	322	3,051
is an element of	1,300	109	190	1,599
is marked by	143	13	17	173
is off the coast of	21	1	1	23
is N of	84	9	8	101
is NNE of	46	1	3	50
is NE of	47	1	5	53
is ENE of	72	11	6	89
is E of	92	6	11	109
is ESE of	73	8	13	94
is SE of	42	4	1	47
is SSE of	51	10	3	64
is S of	84	8	12	104
is SSW of	86	6	7	99
is SW of	45	2	5	52
is WSW of	75	10	6	91
is W of	75	10	11	96
is WNW of	76	4	5	85
is NW of	32	2	8	42
is NNW of	63	7	10	80

Table 5.3.2 – Number of each type of entity and relation label per dataset split.

5.3.1.1 Reuse of the ATLANTIS Dataset: the TextMine'24 Dataset

The TextMine'24 Dataset is a subset of the ATLANTIS Dataset and contains nested spatial entity annotations (Rawsthorne et al. 2024a). It was used as the benchmark dataset for the *Défi TextMine 2024* [18], a spatial entity recognition challenge hosted by the TextMine working group [19], which is part of the *Association Internationale Francophone d'Extraction et de Gestion des Connaissances* (*Association EGC*) [20], the International Francophone Association for Knowledge Extraction and Management.

5.3.2 Model Training and Testing

PURE (Zhong and D. Chen 2021) independently trains two base encoders from existing pretrained deep language models: one to identify and label entity spans, and one to identify related pairs of entity spans and classify the relation between them. We will refer to the former as the

18. https://www.kaggle.com/competitions/defi-textmine-2024
19. https://textmine.sciencesconf.org/
20. https://www.egc.asso.fr/

```
{ "doc_key": "d6_sec_4-6_para_4-6-3_extract",
  "dataset": "atlantis_example",
  "sentences": [["A", "fishing", "port", "lies", "5.7", "M", "to", "the", "ENE"
      , "of", "Ras", "Magroua", "."], ["2", "M", "further", "north", ",", "
      Hadjrat", "Nadji", "(", "Colombi", "islet", ")", "is", "a", "rocky", "
      islet", "," "28", "m", "high", ",", "0.3", "M", "NW", "of", "the", "
      coastline", "." ]],
  "ner": [[[1, 2, "geogFeat"], [10, 10, "geogFeat"], [11, 11, "name"], [10, 11,
       "geogName"]], [[18, 18, "geogFeat"], [18, 19, "geogName"], [19, 19, "
      name"], [22, 22, "geogFeat"], [21, 22, "geogName"], [21, 21, "name"], [27
      , 27, "geogFeat"], [38, 38, "geogFeat"]]],
  "relations": [[[1, 2, 10, 11, "EastNorthEastOf"]], [[27, 27, 38, 38, "
      NorthWestOf"]]] }
```


Listing 5.3.2 – One line from a JSONL file formatted as required for PURE. It contains the text displayed in extract 5.3.1 and its corresponding nested entity and binary relation annotations. The first of the two sentences is illustrated with its corresponding annotations in figure 5.3.1.

> "A[0] fishing[1] port[2] lies[3] 5.7[4] M[5] to[6] the[7] ENE[8] of[9] Ras[10] Magroua[11].[12] 2[13] M[14] further[15] north[16],[17] Hadjrat[18] Nadji[19] ([20]Colombi[21] islet[22])[23] is[24] a[25] rocky[26] islet[27],[28] 28[29] m[30] high[31],[32] 0.3[33] M[34] NW[35] of[36] the[37] coastline[38].[39]"

Extract 5.3.1 – Translated from the original French text: "Un port de pêche est établi à 5,7 M à l'ENE de Ras Magroua. À 2 M plus au Nord, Hadjrat Nadji (îlot Colombi) est un îlot rocheux, haut de 28 m, à 0,3 M au NW du rivage." (Shom 2021d, p. 181). The index assigned to each token in listing 5.3.2 figures in superscript after each token in this extract.

entity model and the latter as the *relation model*. PURE also allows the regulation of the size of the context window W, that is to say the amount of cross-sentence context that is made available for the model. The context made available during the processing of a given sentence spans from $(W - n)/2$ words to the left of the sentence to $(W - n)/2$ words to the right, where n is the number of words in the sentence. A cross-entropy loss is used for both models.

For the base encoders we used *bert-base-french-europeana-cased*[21] as our pretrained monolingual French BERT model[22] and *bert-base-multilingual-cased*[23] as our pretrained multilingual BERT model. We used the default hyperparameters provided by Zhong and D. Chen (2021), which are presented in table 5.3.3, and experimented over multiple context window sizes, within the ranges of default values, for both the entity model and the rela-

21. https://huggingface.co/dbmdz/bert-base-french-europeana-cased
22. The popular pretrained monolingual French BERT model *CamemBERT* presented by Martin et al. (2020) is not compatible with PURE. In order to keep an identical workflow for the training of the monolingual and multilingual models we chose to use *bert-base-french-europeana-cased*, which is compatible with PURE. This model is pretrained primarily on 20th-century texts. We judge that its pretraining is well suited to our corpus of the *Instructions nautiques*, which are written in formal language.
23. https://huggingface.co/bert-base-multilingual-cased

tion model (for the entity model the default values are 0, 100 and 300, and for the relation model the default values are 0 and 100) to be able to study its effect on the results. For the entity model we used context windows of 0, 50, 100, 150, 200 and 248 ($W > 248$ exceeded our available GPU memory usage) and for the relation model we used context windows of 0, 50 and 100. We trained and evaluated the monolingual and multilingual base BERT encoders for nested spatial entity extraction and then separately trained and evaluated the same two base encoders for relation extraction. During training, the models had access to the gold entity annotations. We performed two different evaluations on the relation models: one with the gold entity annotations and one with predicted entities. The relations predicted from gold entity annotations give solely an evaluation of the relation extraction process. The relations predicted from predicted entities give an evaluation of the end-to-end entity and relation extraction process. For each configuration, we trained and evaluated five individual models using different seed values (to initialise the connection weights differently each time) and calculated the arithmetic mean and the standard deviation of the micro F1-scores obtained. A predicted annotation is considered to be correct if its span and its label is correctly assigned.

Hyperparameter	Entity Model	Relation Model
learning rate	1e-5	2e-5
task learning rate	5e-4	-
train batch size	16	32
training epochs	100	10

Table 5.3.3 – Values of hyperparameters used for all experiments. The learning rate is the learning rate for the BERT encoder parameters and the task learning rate is the learning rate for the classifier head after the encoder.

5.4 Results and Evaluation

Listings 5.4.1 and 5.4.2 show the JSONL files outputted by PURE with the predictions for the sentences shown in extract 5.3.1 on page 123 made by the monolingual and multilingual models respectively. The gold annotations for these sentences can be seen in listing 5.3.2, also on page 123 (these gold annotations were not used to train either model). The predictions made by both models are very similar. The only difference is that the multilingual model was able to identify the word “Hadjrat” (written in romanised Arabic, meaning *rock* or *islet*) as being a geographic feature whilst the monolingual model was not.

```
{ "doc_key": "d6_sec_4-6_para_4-6-3_extract",
  "dataset": "atlantis_example",
  "sentences": [["A", "fishing", "port", "lies", "5.7", "M", "to", "the", "ENE"
      , "of", "Ras", "Magroua", "."], ["2", "M", "further", "north", ",", "
      Hadjrat", "Nadji", "(", "Colombi", "islet", ")", "is", "a", "rocky", "
      islet", "," "28", "m", "high", ",", "0.3", "M", "NW", "of", "the", "
      coastline", "." ]],
  "ner": [[], []],
  "relations": [[], []],
  "predicted_ner": [[[1, 2, "geogFeat"], [10, 10, "geogFeat"], [11, 11, "name"]
      , [10, 11, "geogName"]], [[18, 19, "geogName"], [19, 19, "name"], [22, 22
      , "geogFeat"], [21, 22, "geogName"], [21, 21, "name"], [27, 27, "geogFeat
      "], [38, 38, "geogFeat"]]],
  "predicted_relations": [[[1, 2, 10, 11, "EastNorthEastOf"]], [[27, 27, 38, 38
      , "NorthWestOf"]]]}
```


Listing 5.4.1 – One line from a JSONL file outputted by PURE. It contains the annotation predictions made by the monolingual French BERT model that we trained when given the line from a JSONL file shown in listing 5.3.2 as input.

The overall F1-scores for the three tasks carried out with varying context window sizes are displayed in table 5.4.1. Tables 5.4.2, 5.4.3 and 5.4.4 give the detailed entity, relation and end-to-end entity and relation extraction results respectively for all labels.

Context Window	F1 Entity		F1 Relation		F1 e2e	
	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.
$W = 0$	91.1 ± 0.3	91.9 ± 0.2	$\underline{\mathbf{64.2}} \pm 2.2$	$\mathbf{63.2} \pm 1.0$	$\underline{\mathbf{63.9}} \pm 2.2$	$\mathbf{63.2} \pm 1.2$
$W = 50$	92.1 ± 0.2	92.3 ± 0.3	64.2 ± 1.4	63.0 ± 1.7	63.8 ± 1.4	63.1 ± 1.7
$W = 100$	91.9 ± 0.2	92.3 ± 0.2	63.7 ± 0.7	62.9 ± 0.7	63.6 ± 0.7	62.9 ± 0.8
$W = 150$	91.9 ± 0.2	92.2 ± 0.2	-	-	-	-
$W = 200$	92.0 ± 0.2	$\underline{\mathbf{92.3}} \pm 0.2$	-	-	-	-
$W = 248$	$\mathbf{92.2} \pm 0.2$	92.3 ± 0.2	-	-	-	-

Table 5.4.1 – Mean micro F1-score with standard deviation over five runs for varying context window (W) sizes for: entity extraction, relation extraction from gold entity annotations, and end-to-end entity and relation extraction (e2e) from best predicted entity annotations ($W = 248$ for monolingual, $W = 200$ for multilingual). For each task, the highest F1-score over all context window sizes for each base encoder is in bold, and the overall highest F1-score over all context window sizes and both base encoders is underlined.

Our experiments show that PURE (Zhong and D. Chen 2021) is capable of extracting nested spatial entities, and that it can do so via nested annotations. This dispenses with the need for joint labelling as introduced by Agrawal et al. (2022), which is more costly than producing nested annotations due to the additional pre-processing of the annotated dataset (either to merge the annotation labels, or to directly carry out the less intuitive

```
{ "doc_key": "d6_sec_4-6_para_4-6-3_extract",
  "dataset": "atlantis_example",
  "sentences": [["A", "fishing", "port", "lies", "5.7", "M", "to", "the", "ENE"
      , "of", "Ras", "Magroua", "."], ["2", "M", "further", "north", ",", "
      Hadjrat", "Nadji", "(", "Colombi", "islet", ")", "is", "a", "rocky", "
      islet", "," "28", "m", "high", ",", "0.3", "M", "NW", "of", "the", "
      coastline", "." ]],
  "ner": [[], []],
  "relations": [[], []],
  "predicted_ner": [[[1, 2, "geogFeat"], [10, 10, "geogFeat"], [11, 11, "name"]
      , [10, 11, "geogName"]], [[18, 18, "geogFeat"], [18, 19, "geogName"], [19
      , 19, "name"], [22, 22, "geogFeat"], [21, 22, "geogName"], [21, 21, "name
      "], [27, 27, "geogFeat"], [38, 38, "geogFeat"]]],
  "predicted_relations": [[[1, 2, 10, 11, "EastNorthEastOf"]], [[27, 27, 38, 38
      , "NorthWestOf"]]]}
```


Listing 5.4.2 – One line from a JSONL file outputted by PURE. It contains the annotation predictions made by the multilingual BERT model that we trained when given the line from a JSONL file shown in listing 5.3.2 as input.

Label	Precision		Recall		F1-Score	
	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.
all entity labels	94.6	**95.2**	**89.8**	89.6	92.2	**92.3**
geographic feature	94.1	94.4	95.8	95.1	95.0	94.8
name	97.7	97.4	78.0	78.4	86.7	86.9
geographic name	92.9	94.9	91.3	91.4	92.1	93.1

Table 5.4.2 – [Precision|Recall|F1-Score] [Mono.|Multi.] gives the mean [precision|recall|micro F1-score] for the [monolingual|multilingual] model over five runs for entity extraction for each entity label using the context window that gives the best overall results ($W = 248$ for monolingual, $W = 200$ for multilingual). For each task, the highest precision, recall and micro F1-score for all entity labels over both base encoders is in **bold**.

joint labelling annotation task) and post-processing of the predictions (to separate the joint labels).

For entity extraction our experiments show that making cross-sentence context available during training and prediction improves micro F1-scores for both models, and that the multilingual BERT model slightly outperforms the monolingual French BERT model for all context window sizes, with its highest mean micro F1-score being 92.3 when $W = 200$ or $W = 248$ (table 5.4.1). We attribute this contrast in results compared to those in the literature reviewed in section 5.2, where monolingual models outperform multilingual models, to a characterising feature of our dataset: although the main language of the text is French, it contains words from a large number of other languages. The words in question are primarily geographic features that are part of geographic names, meaning that they

Label	Precision		Recall		F1-Score	
	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.
all relation labels	**70.8**	67.2	58.8	**59.9**	**64.2**	63.2
is an element of	70.5	67.9	60.2	60.3	64.9	63.8
is marked by	64.9	55.1	51.8	49.4	57.5	50.8
is off the coast of	100.0	100.0	100.0	100.0	100.0	100.0
is N of	48.6	39.8	52.5	50.0	49.9	44.2
is NNE of	76.7	37.3	73.3	53.3	74.1	43.3
is NE of	95.0	100.0	64.0	60.0	76.1	75.0
is ENE of	83.0	93.3	63.3	83.3	71.6	87.9
is E of	62.5	57.1	45.5	41.8	52.6	48.1
is ESE of	73.4	71.1	55.4	66.2	62.6	67.8
is SE of	90.0	60.0	100.0	100.0	93.3	73.3
is SSE of	73.7	81.3	73.3	66.7	68.8	69.3
is S of	84.7	82.1	71.7	81.7	77.3	81.1
is SSW of	87.4	74.1	68.6	85.7	76.0	79.3
is SW of	0.0	0.0	0.0	0.0	0.0	0.0
is WSW of	92.0	83.6	66.7	70.0	77.1	75.3
is W of	79.3	80.4	65.5	60.0	71.3	68.0
is WNW of	100.0	89.3	88.0	88.0	93.3	87.9
is NW of	67.3	87.4	35.0	50.0	45.7	63.0
is NNW of	61.7	64.1	44.0	40.0	51.1	49.0

Table 5.4.3 – [Precision|Recall|F1-Score] [Mono.|Multi.] gives the mean [precision|recall|micro F1-score] for the [monolingual|multilingual] model over five runs for relation extraction for each relation label from gold entity annotations using the context window that gives the best overall results ($W = 0$ for monolingual and for multilingual). For each task, the highest precision, recall and micro F1-score for all relation labels over both base encoders is in **bold**.

must be identified and correctly labelled by the entity model. The multilingual model has an advantage over the monolingual model in these cases as it is able to understand the semantic meaning of a larger proportion of the words in the dataset.

For relation extraction and end-to-end extraction our experiments show that the monolingual French BERT model slightly outperforms the multilingual BERT model for all context window sizes, with its highest mean micro F1-scores being 64.2 and 63.9 respectively when $W = 0$ (table 5.4.1), which means that the monolingual model performs better at relation prediction whether provided with perfect or imperfect entity labels. The monolingual French BERT model achieves higher precision scores for relation extraction (table 5.4.3) and end-to-end extraction (table 5.4.4) than the multilingual BERT model, but the inverse is true of the recall scores. These results reflect the fact that relations are always expressed in French in the dataset, and sometimes require intricate semantic information to be understood. Taking a closer look at the results for the individual relation labels, we can see that the "is an element of" and the "is marked

Label	Precision		Recall		F1-Score	
	Mono.	Multi.	Mono.	Multi.	Mono.	Multi.
all relation labels	**70.2**	67.3	58.8	**59.8**	**63.9**	63.2
is an element of	70.2	68.2	60.2	60.2	64.8	63.9
is marked by	61.3	55.1	51.8	49.4	56.1	50.8
is off the coast of	100.0	100.0	100.0	100.0	100.0	100.0
is N of	47.7	41.8	52.5	50.0	49.7	45.2
is NNE of	76.7	39.3	73.3	53.3	74.1	44.8
is NE of	95.0	100.0	64.0	60.0	76.1	75.0
is ENE of	96.0	93.3	63.3	83.3	75.9	87.9
is E of	67.9	57.1	45.5	41.8	54.4	48.1
is ESE of	70.9	71.1	55.4	66.2	61.6	67.8
is SE of	90.0	60.0	100.0	100.0	93.3	73.3
is SSE of	71.7	81.3	73.3	66.7	67.1	69.3
is S of	84.7	82.1	71.7	81.7	77.3	81.1
is SSW of	87.4	74.1	68.6	85.7	76.0	79.3
is SW of	0.0	0.0	0.0	0.0	0.0	0.0
is WSW of	92.0	83.6	66.7	70.0	77.1	75.3
is W of	75.8	78.9	65.5	60.0	69.5	67.3
is WNW of	100.0	89.3	88.0	88.0	93.3	87.9
is NW of	80.0	87.4	35.0	50.0	47.0	63.0
is NNW of	63.5	64.1	44.0	40.0	51.8	49.0

Table 5.4.4 – [Precision|Recall|F1-Score] [Mono.|Multi.] gives the mean [precision|recall|micro F1-score] for the [monolingual|multilingual] model over five runs for end-to-end entity and relation extraction for each relation label from the best predicted entity annotations using the context window that gives the best overall results ($W = 0$ for monolingual and for multilingual). For each task, the highest precision, recall and micro F1-score for all relation labels over both base encoders is in **bold**.

by" labels have overall lower results than many of the relation labels that involve the cardinal directions. This may be explained by the numerous ways in which these two relations are expressed in our corpus, in comparison with all the other relations that are always expressed using the same key words. The results for both relation models decrease slightly as the size of the context window increases. This may be attributed to the fact that all the information that categorises one relation is generally included in one sentence, meaning that cross-sentence context may not contribute useful information. Both models give results that are less stable than for entity extraction. This lack of stability may be attributed to the relatively small number of examples of certain relation types in our dataset.

5.5 Conclusion

In this chapter we discussed and emphasised the importance of reliable nested spatial entity and spatial relation extraction to the construction of

geospatial knowledge graphs from text and the disambiguation of spatial entities. We then presented our approach for automatically extracting nested entities and binary relations from text. Based on PURE (Zhong and D. Chen 2021), it requires a dataset with nested annotations to train deep language models. As it uses deep learning techniques, our approach can adapt to text containing yet-unknown vocabulary. This means that there is no need to provide an exhaustive gazetteer, unlike for rule-based approaches that are very dependent on the pre-defined terminology and would therefore have required a complete ontology to work well.

We demonstrated how we implemented our adapted PURE approach to train multilingual and monolingual French deep language models to extract nested spatial entities and binary topological and directional spatial relations from our corpus of the *Instructions nautiques*. To do so, we created the annotated French-language ATLANTIS Dataset, which we introduced in this chapter, from extracts of the *Instructions nautiques*. We provided benchmark results for our own dataset and thereby demonstrated that PURE (Zhong and D. Chen 2021), an existing approach for generic entity and binary relation extraction from text, can be used to extract *nested* entities. This was achieved by training a BERT encoder with nested annotations, without using joint labelling. We also showed that PURE is a suitable baseline approach for the extraction of domain-specific *spatial* entities and *spatial* relations. Our results reveal that the multilingual BERT model outperforms the monolingual French BERT model for entity extraction, with a mean micro F1-score of 92.3, whilst for relation extraction and end-to-end entity and relation extraction the monolingual French BERT model performs best, with mean micro F1-scores of 64.2 and 63.9 respectively. Our results show that making cross-sentence context information available during training and prediction favours entity extraction but hinders relation extraction.

The end-to-end extraction results could potentially be improved by combining the training of a multilingual BERT model for the entity extraction task with that of a monolingual French BERT model for the relation extraction task. Results could also be improved thanks to the addition of class-based weights in the loss function to penalise impossible label combinations. Future work with respect to our application context includes extending this baseline approach to the extraction of intrinsic and relative directional spatial relations as well as distance spatial relations, such as the one in figure 5.3.1. We would like to investigate the possibility of expanding the set of relation types to include non-spatial relations to describe characteristics of the spatial entities such as their colour. Our ap-

proach would also need to be applied to the extraction of the information required to build the triplets that correspond to the other subdomains as defined in chapter 4. We believe that our approach could be directly applied to the entities and relations in the Temporalities, Meteorological and Oceanographic Phenomena and Maritime Vessels subdomains as they are constructed in a very similar way to spatial entities and relations, but that it may need to be modified for the Maritime Navigation Guidelines subdomain. Finally, it would be interesting to apply our approach to Sailing Directions written in other languages using different BERT models.

The following chapter presents the third and final component that makes up the ATONTE Methodology, which involves structuring geographic information in the form of a knowledge graph. We will apply this work to the spatial entities extracted using the approach presented in this chapter.

Chapter 6

Information Structuring and Entity Disambiguation

6.1 Introduction

In this chapter we present the third and final component of the ATlantis Ontology and kNowledge graph development from Texts and Experts (ATONTE) Methodology, which deals with the automated structuring of information extracted from text as a knowledge graph according to an ontology, and disambiguating entities via entity linking to a reference resource. We present a proof of concept of this stage, using off-the-shelf tools to first structure the spatial entities and relations extracted from the *Instructions nautiques* in chapter 5 as Resource Description Framework (RDF) triples to populate and enrich the coAsTaL mAritime NavigaTion InstructionS (ATLANTIS) Ontology that we developed in chapter 4, and then to link the spatial entities to their corresponding entries in the BD TOPO®. The result is an operational basis for the geospatial ATLANTIS Knowledge Graph of the *Instructions nautiques*. Although we present the construction of a geospatial knowledge graph and deal with the special case of associating the spatial entities within it to their geographic coordinates from a reference geographic resource, the same steps could be followed to construct a non-geospatial knowledge graph and associate the other types of entities within it to other pieces of information from other reference resources.

During geospatial knowledge graph population, spatial entities become instances of ontological classes and spatial relations become assertions of object properties. To be able to correctly assign spatial entities to their corresponding ontological class, it is necessary to know their *type*. By extracting *nested* as opposed to *flat* spatial entities, the entity type is already known and an instance of the right class can be created automatically. This extra piece of information, the entity *type*, facilitates the disambiguation

task of linking the instance to the correct entry in a reference geographic resource (Southall et al. 2011). Before being able to perform entity disambiguation on the named spatial entities extracted from the *Instructions nautiques*, we need to structure the extracted information as RDF triples. This includes creating an instance for each spatial entity, named and unnamed, and associating each one with its geographic feature type via a dedicated triple.

In section 6.2 we review related work in terms of structuring information as RDF triples (section 6.2.1) and in terms of entity disambiguation (section 6.2.2). Section 6.3 is dedicated to presenting the third component of ATONTE, which consists of an automated approach for the population of a knowledge graph from information extracted from text. It involves first structuring the extracted information as RDF triples and then disambiguating the entities by linking them with a reference resource. We also demonstrate how we applied our approach to automatically structure the nested spatial entities and binary spatial relations extracted from the *Instructions nautiques* in chapter 5 as RDF triples and then automatically disambiguate the structured spatial entities. In section 6.4 we present our results for the population task, which we then evaluate in section 6.5 before concluding in section 6.6.

6.2 Related Work

6.2.1 Structuring Information as RDF

There exist many ways of converting semi-structured data into RDF and vice versa. We wish to find a flexible approach that is suited to our available data formats and is easy to implement. The World Wide Web Consortium (W3C) community maintains a list of tools, frameworks and applications for converting many different formats of data into RDF [1]. Of note are the RDF Mapping Language (RML), SPARQL-Generate and XSPARQL.

RML is a mapping language that can be used to write rules to map various data sources such as CSV, TSV, XML and JSON to RDF (Dimou et al. 2014).

SPARQL-Generate [2], an extension of SPARQL 1.1, is an expressive language for generating RDF from heterogeneous data sources including SQL, XML, JSON, CSV, Geographic JSON (GeoJSON), HTML and plain text (Lefrançois et al. 2017).

The XSPARQL 1.1 query language combines languages including XQuery,

1. https://www.w3.org/wiki/ConverterToRdf
2. https://ci.mines-stetienne.fr/sparql-generate/

SPARQL and SQL to provide a way to map from XML, JSON or relational data to RDF (Dell'Aglio et al. 2014).

We chose to use SPARQL-Generate because of its adaptability to multiple data formats and thanks to its similarity to the SPARQL language, making it easy to learn.

6.2.2 Entity Disambiguation

A first study concerning the automatic disambiguation of named spatial entities extracted from the *Instructions nautiques* was carried out by Loynes and Ruiz (2020). To test the various algorithms that they create, they use an RDF dataset of flat named spatial entities extracted from the *Instructions nautiques* by Lamotte et al. (2020). To produce this dataset, Lamotte et al. developed a method that combines a lexical approach with a linguistic pattern-based approach for indirect spatial information extraction and applied it to the *Instructions nautiques*.

Entity disambiguation involves associating each entity within a set of ambiguous [3] entities with its corresponding entry in a reference resource. For Loynes and Ruiz (2020), the set of ambiguous entities is the RDF dataset of flat named spatial entities extracted from the *Instructions nautiques* and the reference resource is the BD TOPO®. The spatial entities contained within the RDF dataset are dispersed along the north coast of France, from the border with Belgium to south Brittany. For the disambiguation of these entities, Loynes and Ruiz use only an extract of the BD TOPO®, keeping entries that lie within 10 km of the coastline indicated, to limit noise. To simplify their calculations, they approximate all line and polygon geometries in the BD TOPO® extract as point geometries at their centre. One entry from the BD TOPO® extract is displayed in listing 6.2.1. Each entry has numerous attribute values including its unique identifier (line 13), its coordinates in the Lambert 93 projection (lines 6 to 9), its toponym (line 18) and two attributes that can be used to describe its type (lines 16 and 17). By associating an ambiguous spatial entity from the *Instructions nautiques* with its corresponding entry in the BD TOPO® extract, it is attributed direct spatial referencing in the form of geographic coordinates that allow locating the entity in the real world.

The entity disambiguation process involves two main steps. First, candidate entries for each entity needing to be disambiguated are selected from the reference resource by an algorithm according to given criteria. Second,

3. An entity mention is ambiguous if its corresponding real-world entity is not explicit. For example, "Montreuil" could refer to multiple different places in France including communes in Centre-Val de Loire, Île-de-France and Pays de la Loire.

```
{
        "_source": {
                "type": "Feature",
                "geometry": {
                        "type": "Point",
                        "coordinates": [
                                580025.196113,
                                7034336.384778
                        ]
                },
                "properties": {
                        "id": 119549,
                        "cleabs": "SURFPARC0000000322465458",
                        "gazetier": "bdtopo",
                        "tablesource": "parc_ou_reserve_in",
                        "nature": "Parc naturel marin",
                        "naturedetaillee": null,
                        "toponyme": "Parc Naturel Marin des Estuaires Picards
                           et de la Mer d'Opale"
                }
        }
}
```

Listing 6.2.1 – One entry from the extract of the BD TOPO® as produced by Loynes and Ruiz (2020) as indexed in Elasticsearch.

another algorithm assigns a similarity score to each candidate entry, which allows the set of candidate entries for each ambiguous entity to be ranked in terms of relevance. They are ranked from most relevant in first place to least probable in last place. Once these two steps have been carried out, Loynes and Ruiz associate the top candidate entry to the ambiguous spatial entity via an `owl:sameAs` property pointing from the IRI of the ambiguous entity to the IRI of the entry in the BD TOPO® extract.

In their study, Loynes and Ruiz store the RDF dataset of flat spatial entity names extracted from the *Instructions nautiques* in a GraphDB [4] triplestore. Saved as a GeoJSON file, the BD TOPO® extract is loaded into a local Elasticsearch [5] search engine server, which allows configuring the indexation of the database and fetching data from the triplestore. They implement their various candidate selection, scoring and ranking methods using Python scripts that fetch data from the triplestore and execute Elasticsearch queries on the database.

4. https://graphdb.ontotext.com/
5. https://www.elastic.co/

6.2.2.1 Candidate Selection

Loynes and Ruiz (2020) present three different methods for the candidate selection step. They rely on three different measurements of the similarity between two strings, which are: the name of the flat spatial entity extracted from the *Instructions nautiques*, and the value of the `"toponyme"` attribute (listing 6.2.1, line 18) of entries from the BD TOPO® extract. The first method simply assesses whether or not the two strings are identical. The second method uses the Levenshtein distance, which defines the distance between two strings as being equal to the minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into the other. The third method is based on n-grams, which are series of n adjacent characters or spaces within a string. It compares the first string with the second string divided into n-grams. The three methods and their specific criteria as implemented by Loynes and Ruiz are:

— Strict string equality, the criterion of which requires both strings to be identical

— Levenshtein distance, the criterion of which requires the Levenshtein distance between the two strings to be equal or inferior to 2

— N-gram, the criterion of which requires either all of the words[6] (if the 'AND' operator is used) or at least one of the words (if the 'OR' operator is used) in the entity name string to have a match within the n-grams present in the index of n-grams from $n = 2$ to $n = 10$[7] for a given database entry

6.2.2.2 Candidate Scoring and Ranking

Loynes and Ruiz present two scoring methods for the candidate ranking step. The first is based on the similarity of the two strings according to the candidate selection method used. The second is taken from work by Keller (2016) who implemented a method based on the median geographic location of spatial entities mentioned in the same document. It uses the working hypothesis that if a text is organised geographically then the entities that are close together in the text are also close together geographically. The method calculates and compares the geographic distance between each individual candidate selected from the database for the entity in question with the median location of all the candidates selected for all the other

6. The entity name string is split into words on whitespace.

7. For the string "chausée de sein", the 2-grams would be: ch, ha, au, us, ss, sé, ée, e␣, ␣d, de, e␣, ␣s, se, ei, in. The 10-grams would be: chaussée␣d, haussée␣de, aussée␣de␣, ussée␣de␣s, ssée␣de␣se, sée␣de␣sei, ée␣de␣sein.

entities mentioned in the same document. The individual candidate with the smallest distance to the median location is assumed to be the most likely corresponding entry.

In section 6.3.2 we present our proof of concept of the entity disambiguation task by building on the work done by Loynes and Ruiz (2020) and adding in particular the entity *type* to the similarity measurements.

6.3 Proof of Concept

In this section we present our approach for automatically populating a knowledge graph from information extracted from text as a proof of concept. The first step involves structuring the extracted information as RDF triples, which we demonstrate in section 6.3.1, and the second step involves disambiguating the entities by linking them to a reference resource, which we demonstrate in section 6.3.2.

This third stage of the ATONTE Methodology is designed to be carried out using the output of the deep neural network trained in chapter 5: a JSON Lines (JSONL) file such as the one shown in listing 5.4.2 on page 126. However, to avoid error propagation in this proof of concept we use part of the manually-annotated gold dataset presented in section 5.3.1 on pages 117 to 121, which we have formatted as if it were the output of the network (we moved our manual annotations from the `"ner"` and `"relations"` keys to the `"predicted_ner"` and `"predicted_relations"` keys) as can be seen in listing 6.3.1. The specific part of the manually-annotated gold dataset that we use in this stage includes all the paragraphs that we extracted from the volume of the *Instructions nautiques* that covers the coast of France from the Cap de la Hague in Normandy to Point Penmarc'h in Brittany (Shom 2021a).

6.3.1 Structuring Information as RDF

The deep neural network used in chapter 5 outputs nested entity and binary relation label predictions on the input text in a JSONL file such as the one shown in listing 5.4.2 on page 126.

Thanks to the fact that we performed *nested* as opposed to *flat* named spatial entity extraction on the *Instructions nautiques*, the geographic feature type of each named spatial entity is isolated and labelled independently of the name, when the noun is included in the name. This makes it possible to automatically assign the correct geographic feature type to most of the named spatial entities extracted.

```
{ "doc_key": "C22_sec_2-2_para_2-2-3-6_extract",
  "dataset": "atlantis_example",
  "sentences": [["La", "chaussée", "de", "Sein", "est", "un", "vaste", "
     ensemble", "d'", "îles", ",", "de", "rochers", "et", "de", "roches", ",",
      "qui", "englobe", "l'", "île", "de", "Sein", "et", "s'", "étire", "sur",
      "12", "M", "entre", "le", "Pont", "des", "Chats", ",", "à", "l'", "Est",
      ",", "et", "la", "bouée", "«", "Chaussée", "de", "Sein", "»", ",", "
     cardinale", "Ouest", "lumineuse", "à", "Racon", "et", "AIS", ",", "à", "l
     '", "Ouest", "."]],
  "ner": [[]],
  "relations": [[]],
  "predicted_ner": [[[1, 1, "geogFeat"], [3, 3, "name"], [1, 3, "geogName"], [9
     , 9, "geogFeat"], [12, 12, "geogFeat"], [15, 15, "geogFeat"], [20, 20, "
     geogFeat"], [22, 22, "name"], [20, 22, "geogName"], [31, 31, "geogFeat"],
      [33, 33, "name"], [31, 33, "geogName"], [41, 41, "geogFeat"], [43, 45, "
     name"], [41, 46, "geogName"], [52, 52, "geogFeat"], [54, 54, "geogFeat"]]
     ],
  "predicted_relations": [[[9, 9, 1, 3, "ElementOf"], [12, 12, 1, 3, "ElementOf
     "], [15, 15, 1, 3, "ElementOf"], [20, 22, 1, 3, "ElementOf"], [31, 33, 1,
      3, "EastOf"], [41, 46, 1, 3, "WestOf"], [52, 52, 41, 46, "ElementOf"], [
     54, 54, 41, 46, "ElementOf"]]] }
{ "doc_key": "C22_sec_4-3_para_4-3-1-1_extract",
  "dataset": "atlantis_example",
  "sentences": [["Au", "SSE", "de", "la", "pointe", "de", "la", "Varde", ",", "
     le", "phare", "de", "Rochebonne", ",", "tour", "carrée", "grise", ",", "
     blanche", "sur", "sa", "face", "Ouest", "et", "à", "sommet", "rouge", ","
     , "haute", "de", "20", "m", ",", "domine", "la", "plage", "du", "Minihic"
     , "devant", "Paramé", "."]],
  "ner": [[]],
  "relations": [[]],
  "predicted_ner": [[[4, 4, "geogFeat"], [7, 7, "name"], [4, 7, "geogName"], [1
     0, 10, "geogFeat"], [12, 12, "name"], [10, 12, "geogName"], [14, 14, "
     geogFeat"], [21, 21, "geogFeat"], [22, 22, "name"], [21, 22, "geogName"],
      [25, 25, "geogFeat"], [35, 35, "geogFeat"], [37, 37, "name"], [35, 37, "
     geogName"], [39, 39, "name"], [39, 39, "geogName"]]],
  "predicted_relations": [[[10, 12, 4, 7, "SouthSouthEastOf"], [14, 14, 10, 12,
      "ElementOf"], [21, 22, 14, 14, "ElementOf"], [25, 25, 14, 14, "ElementOf
     "]]] }
```


Listing 6.3.1 – Two lines from a JSONL file formatted as the output from the Princeton University Relation Extraction system (PURE). The values of the `"predicted_ner"` and `"predicted_relations"` keys are in fact the manually-annotated labels from the gold dataset prepared in section 5.3.1.

Given the specific formatting of the JSONL output files, which is not directly compatible with SPARQL-Generate, we first convert the files to XML using a Jupyter Notebook [8]. This notebook transforms the predicted labels to inline XML tags [9] and produces one XML file with the entity tags and the text such as the one shown in listing 6.3.2, and one XML file with the relation tags such as the one shown in listing 6.3.3. Each child element of the `root` element is a `doc` element that corresponds to one paragraph from the *Instructions nautiques*.

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
    <doc id="C22_sec_2-2_para_2-2-3-6_extract">La <geogName id="1-3"><geogFeat
        id="1-1">chaussée</geogFeat> de <name id="3-3">Sein</name></geogName>
        est un vaste ensemble d' <geogFeat id="9-9">îles</geogFeat> , de <
        geogFeat id="12-12">rochers</geogFeat> et de <geogFeat id="15-15">
        roches</geogFeat> , qui englobe l' <geogName id="20-22"><geogFeat id="
        20-20">île</geogFeat> de <name id="22-22">Sein</name></geogName> et s'
        étire sur 12 M entre le <geogName id="31-33"><geogFeat id="31-31">Pont<
        /geogFeat> des <name id="33-33">Chats</name></geogName> , à l' Est , et
         la <geogName id="41-46"><geogFeat id="41-41">bouée</geogFeat> « <name
        id="43-45">Chaussée de Sein</name> »</geogName> , cardinale Ouest
        lumineuse à <geogFeat id="52-52">Racon</geogFeat> et <geogFeat id="
        54-54">AIS</geogFeat> , à l' Ouest .</doc>
    <doc id="C22_sec_4-3_para_4-3-1-1_extract">Au SSE de la <geogName id="4-7">
        <geogFeat id="4-4">pointe</geogFeat> de la <name id="7-7">Varde</name><
        /geogName> , le <geogName id="10-12"><geogFeat id="10-10">phare</
        geogFeat> de <name id="12-12">Rochebonne</name></geogName> , <geogFeat
        id="14-14">tour</geogFeat> carrée grise , blanche sur sa <geogName id="
        21-22"><geogFeat id="21-21">face</geogFeat> <name id="22-22">Ouest</
        name></geogName> et à <geogFeat id="25-25">sommet</geogFeat> rouge ,
        haute de 20 m , domine la <geogName id="35-37"><geogFeat id="35-35">
        plage</geogFeat> du <name id="37-37">Minihic</name></geogName> devant <
        geogName id="39-39"><name id="39-39">Paramé</name></geogName> .</doc>
</root>
```

Listing 6.3.2 – XML file with inline entity tags after conversion from the JSONL file shown in listing 6.3.1.

We wrote three SPARQL-Generate queries [10], two of which must be run on the XML file containing the inline entity tags and one that must be run on the XML file containing the relation tags. Each query outputs a file of RDF triples written in Turtle syntax, such as those that can be seen in

8. https://github.com/umrlastig/atonte-structure-and-disambiguate/blob/main/structuring/jsonl-to-xml_entities-and-relations.ipynb
9. One pair of XML tags corresponds to one label from the JSONL file.
10. https://github.com/umrlastig/atonte-structure-and-disambiguate/tree/main/structuring/SPARQL-Generate_queries

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <root>
3     <doc id="C22_sec_2-2_para_2-2-3-6_extract"> <relation L="9-9" R="1-3">
          ElementOf</relation> <relation L="12-12" R="1-3">ElementOf</relation> <
          relation L="15-15" R="1-3">ElementOf</relation> <relation L="20-22" R="
          1-3">ElementOf</relation> <relation L="31-33" R="1-3">EastOf</relation>
           <relation L="41-46" R="1-3">WestOf</relation> <relation L="52-52" R="
          41-46">ElementOf</relation> <relation L="54-54" R="41-46">ElementOf</
          relation> </doc>
4     <doc id="C22_sec_4-3_para_4-3-1-1_extract"> <relation L="10-12" R="4-7">
          SouthSouthEastOf</relation> <relation L="14-14" R="10-12">ElementOf</
          relation> <relation L="21-22" R="14-14">ElementOf</relation> <relation
          L="25-25" R="14-14">ElementOf</relation> </doc>
5 </root>
```

Listing 6.3.3 – XML file with relation tags after conversion from the JSONL file shown in listing 6.3.1.

listing 6.3.4. The first query [11], to be ran on the XML file containing the inline entity tags, performs the following functions:

- Creates an instance for each `doc` element (one paragraph from the *Instructions nautiques*) and assigns it a unique IRI based on its `id` attribute (the name of the paragraph), such as in line 1 of listing 6.3.4
  - Creates a `rdf:type` property pointing towards the `atln:INPara` class
- Creates an instance for each `geogFeat` element (one unnamed spatial entity) and assigns it a unique IRI based on its `id` attribute (its span) and the `id` attribute of the `doc` element in which it is contained, such as in lines 8 to 11 of listing 6.3.4
  - Creates an `rdf:type` property pointing towards the `atln:SpatialEntity` class
  - Creates an `atln:hasTypeOfSpatialEntity` property pointing towards the `tse:X` instance, where `X` is the text contained within the `geogFeat` element if there is one
  - Creates an `atln:hasSource` property pointing towards the instance that represents the `doc` element in which it is contained
- Creates an instance for each `geogName` element (one named spatial entity) and assigns it a unique IRI based on its `id` attribute (its span) and the `id` attribute of the `doc` element in which it is contained, such as in lines 3 to 6 of listing 6.3.4

11. https://github.com/umrlastig/atonte-structure-and-disambiguate/blob/main/structuring/SPARQL-Generate_queries/query1_features-and-named-entities.rqg

— Creates an `rdf:type` property pointing towards the `atln:SpatialEntity` class

— Creates an `atln:hasTypeOfSpatialEntity` property pointing towards the `tse:X` instance, where `X` is the text contained within the `geogFeat` element that is contained within the `geogName` element

— Creates an `atln:hasSource` property pointing towards the instance that represents the `doc` element in which it is contained

Using regular expressions, we remove the `geogFeat` and `name` tags (but not the text contained within them) from the XML file containing the inline entity tags and then run the second SPARQL-Generate query on the modified file. The second query [12] performs the following functions:

— For each instance of a named spatial entity created by the previous query:

— Creates an `rdfs:label` property pointing towards a string that corresponds to the text contained within the corresponding `geogName` element

The third and final query [13], to be ran on the XML file containing the relation tags, performs the following functions:

— Creates an `atln:isY` property for each `relation` element pointing from and to the instances whose IRI can be deduced from the `id` attribute of the `doc` element in which the `relation` element is contained combined with the `L` and `R` attributes of the `relation` element respectively, where `Y` is the text contained within the `relation` element, such as in line 15 of listing 6.3.4

We merge the contents of the three files outputted by the queries to create our final file of RDF triples written in Turtle syntax. The final file contains 431 declarations of instances of spatial entities, 260 named and 171 unnamed, and 14 declarations of instances of *Instructions nautiques* paragraphs. When the information was present in the text, the spatial entity instances are associated to their name (only if it is a named spatial entity), their geographic feature type (inevitable if it is an unnamed entity but not if it is a named entity), and their spatial relations with other named and unnamed spatial entities. Each instance of a spatial entity is associated to the paragraph from which it was extracted. An extract of this file is

12. https://github.com/umrlastig/atonte-structure-and-disambiguate/blob/main/structuring/SPARQL-Generate_queries/query2_named-entity-names.rqg

13. https://github.com/umrlastig/atonte-structure-and-disambiguate/blob/main/structuring/SPARQL-Generate_queries/query3_relations.rqg

shown in listing 6.3.4. It contains the triples corresponding to six instances of named spatial entities, one instance of an unnamed spatial entity and two instances of paragraphs from the *Instructions nautiques*, structured via our SPARQL-Generate queries from the XML files in listings 6.3.2 and 6.3.3. One of the named spatial entities, shown on lines 38 to 40 of listing 6.3.4, is not associated with its type because the information does not feature in the entity name, as is the case for the five other named spatial entities displayed in the listing.

6.3.2 Entity Disambiguation

The output of the information structuring step in section 6.3.1 is an RDF dataset of 431 instances of spatial entities and 14 instances of paragraphs from the *Instructions nautiques*.

Thanks to the fact that we performed *nested* as opposed to *flat* named spatial entity extraction on the *Instructions nautiques*, the geographic feature type of each named spatial entity was isolated and labelled independently of the name, when the noun was included in the name. This made it possible to automatically assign the correct geographic feature type to most of the named spatial entities extracted in the form of a dedicate RDF triple, such as in line 4 of listing 6.3.4.

For our spatial entity disambiguation dataset we conserve only the 260 named spatial entities outputted by the information structuring process, as we focus on improving the string similarity matching methods introduced by Loynes and Ruiz (2020) and presented in section 6.2.2. We implement two principal improvements to their algorithms, which we present below.

To have easy access to the RDF disambiguation dataset via SPARQL queries, we store it in a GraphDB triplestore like Loynes and Ruiz. The advantage of our dataset over the one used by Loynes and Ruiz lies in the presence of triples indicating the geographic feature type of most of the named spatial entities. Out of the 260 instances of named spatial entities in our RDF dataset, 201 are associated to their geographic feature type via an `atln:hasTypeOfSpatialEntity` property.

To be able to evaluate the results of our entity disambiguation process, we manually create a gold dataset. This file contains an `owl:sameAs` property pointing from the IRI of each of the 260 instances of spatial entities from the RDF dataset to its corresponding entry in the BD TOPO®. The corresponding entries in the BD TOPO® were found by importing the extract of the BD TOPO® used by Loynes and Ruiz (2020) that covers the north coast of France into QGIS [14]. We were then able to use key-

14. https://qgis.org/

```
inp:C22_sec_2-2_para_2-2-3-6_extract rdf:type atln:INPara .

ent:C22_sec_2-2_para_2-2-3-6_extract_span_1-3 rdf:type atln:SpatialEntity ;
    atln:hasTypeOfSpatialEntity tse:chausée ;
    rdfs:label "chausée de Sein" ;
    atln:hasSource inp:C22_sec_2-2_para_2-2-3-6_extract .

ent:C22_sec_2-2_para_2-2-3-6_extract_span_9-9 rdf:type atln:SpatialEntity ;
    atln:hasTypeOfSpatialEntity tse:îles ;
    atln:isElementOf ent:C22_sec_2-2_para_2-2-3-6_extract_span_1-3 ;
    atln:hasSource inp:C22_sec_2-2_para_2-2-3-6_extract .

ent:C22_sec_2-2_para_2-2-3-6_extract_span_31-33 rdf:type atln:SpatialEntity ;
    atln:hasTypeOfSpatialEntity tse:Pont ;
    atln:isEastOf ent:C22_sec_2-2_para_2-2-3-6_extract_span_1-3 ;
    rdfs:label "Pont des Chats" ;
    atln:hasSource inp:C22_sec_2-2_para_2-2-3-6_extract .

ent:C22_sec_2-2_para_2-2-3-6_extract_span_41-46  rdf:type atln:SpatialEntity ;
    atln:hasTypeOfSpatialEntity tse:bouée ;
    atln:isWestOf ent:C22_sec_2-2_para_2-2-3-6_extract_span_1-3 ;
    rdfs:label "bouée « Chaussée de Sein »" ;
    atln:hasSource inp:C22_sec_2-2_para_2-2-3-6_extract .

inp:C22_sec_4-3_para_4-3-1-1_extract rdf:type atln:INPara .

ent:C22_sec_4-3_para_4-3-1-1_extract_span_4-7 rdf:type atln:SpatialEntity ;
    atln:hasTypeOfSpatialEntity tse:pointe ;
    rdfs:label "pointe de Varde" ;
    atln:hasSource inp:C22_sec_4-3_para_4-3-1-1_extract .

ent:C22_sec_4-3_para_4-3-1-1_extract_span_10-12 rdf:type atln:SpatialEntity ;
    atln:hasTypeOfSpatialEntity tse:phare ;
    atln:isSouthSouthEastOf ent:C22_sec_4-3_para_4-3-1-1_extract_span_4-7 ;
    rdfs:label "phare de Rochebonne" ;
    atln:hasSource inp:C22_sec_4-3_para_4-3-1-1_extract .

ent:C22_sec_4-3_para_4-3-1-1_extract_span_39-39 rdf:type atln:SpatialEntity ;
    rdfs:label "Paramé" ;
    atln:hasSource inp:C22_sec_4-3_para_4-3-1-1_extract .
```

Listing 6.3.4 – RDF triples written in Turtle syntax produced from the annotated XML files shown in listings 6.3.2 and 6.3.3 using SPARQL-Generate.

word searches to find the corresponding database entry for each ambiguous named spatial entity in the gold dataset, and visually verify that it was correct. When a corresponding database entry did not exist, we created an `owl:sameAs` property pointing from the IRI of the instance of the spatial entity to the string `"nil"`. Six instances from the gold dataset can be seen in listing 6.3.5, including one that does not have a corresponding entry in the BD TOPO® extract on line 5.

```
1  ent:C22_sec_2-2_para_2-2-3-6_extract_span_1-3 owl:sameAs
       <http://data.ign.fr/id/topo/PAIHYDRO0000000023222126> .
2
3  ent:C22_sec_2-2_para_2-2-3-6_extract_span_31-33 owl:sameAs
       <http://data.ign.fr/id/topo/PAIHYDRO0000000023222143> .
4
5  ent:C22_sec_2-2_para_2-2-3-6_extract_span_41-46 owl:sameAs "nil" .
6
7  ent:C22_sec_4-3_para_4-3-1-1_extract_span_4-7 owl:sameAs
       <http://data.ign.fr/id/topo/PAIOROGR0000000046888945> .
8
9  ent:C22_sec_4-3_para_4-3-1-1_extract_span_10-12 owl:sameAs
       <http://data.ign.fr/id/topo/CONSPONC0000002202833122> .
10
11 ent:C22_sec_4-3_para_4-3-1-1_extract_span_39-39 owl:sameAs
       <http://data.ign.fr/id/topo/PAIHABIT0000000046957735> .
```

Listing 6.3.5 – RDF triples written in Turtle syntax corresponding to the six named spatial entities shown in listing 6.3.4.

Out of the 260 entities in the gold dataset, 100 have a corresponding entry in the BD TOPO® extract and 160 do not. This is due to the high proportion of long-tail entities (Weikum et al. 2021) in our dataset such as buoys and beacons. These are geographic feature types specific to the maritime environment and often not inventoried in the BD TOPO®, which focuses on terrestrial entities.

Like Loynes and Ruiz, we loaded the BD TOPO® extract into a local Elasticsearch search engine server and implemented our various candidate selection, scoring and ranking methods using Python scripts that fetch data from our triplestore and execute Elasticsearch queries on the database.

6.3.2.1 Candidate Selection

Loynes and Ruiz (2020) presented three candidate selection methods based on a strict string equality, the Levenshtein distance and n-grams respectively. We decided to work on improving only the Levenshtein distance and n-gram methods, as the strict string equality method is noted as not yielding satisfactory results because of its rigidity.

Our first objective is to leverage the extra piece of information in our dataset: the spatial entity type, to improve the results of the two methods. Instead of only measuring the similarity between the name of the spatial entity extracted from the *Instructions nautiques* and the value of the `"toponyme"` attribute of entries from the BD TOPO®, we also wish to measure the similarity between the spatial entity type extracted from the *Instructions nautiques* and the type attribute of the entries as registered in the BD TOPO®. As indicated in section 6.2.2 on page 133 onwards and shown in listing 6.2.1 on page 134, each entry in our extract of the BD TOPO® has two different attributes that can be used to describe its type: `"nature"` and `"naturedetaillee"`. We are obliged to work with both the `"nature"` and the `"naturedetaillee"` attributes as their relevance to the spatial entity type as extracted from the *Instructions nautiques* is not consistent. In listing 6.2.1 on page 134, for example, only the `"nature"` attribute on line 16 has a value associated to it. Let us now take a closer look at some more specific examples. Listing 6.3.6 shows the three attributes with which we will be working (`"nature"`, `"naturedetaillee"` and `"toponyme"`) for five BD TOPO® entries that correspond to five of the spatial entities that feature in listing 6.3.4 on page 142: `"chausée de Sein"`, `"Pont des Chats"`, `"pointe de Varde"`, `"phare de Rochebonne"` and `"Paramé"`. Neither the `"nature"` nor the `"naturedetaillee"` attributes (lines 4, 5, 13 and 14 of listing 6.3.6) are useful for the disambiguation of the first and second entities `"chausée de Sein"` and `"Pont des Chats"`. For the third entity `"pointe de Varde"`, the `"nature"` attribute on line 22 is not useful but the `"naturedetaillee"` attribute on line 23 could be useful if we made use of a thesaurus of synonyms [15]. For the fourth entity `"phare de Rochebonne"`, only the `"naturedetaillee"` attribute on line 32 is of interest. For the fifth and final entity, neither the `"nature"` nor the `"naturedetaillee"` attribute is useful because the geographic feature type is not mentioned in the entity name `"Paramé"`.

6.3.2.2 Candidate Scoring and Ranking

Loynes and Ruiz (2020) presented two candidate scoring and ranking methods.

The first scoring method measures the similarity of the two strings according to the candidate selection method used. Like Loynes and Ruiz, we implement it by using the default relevance score calculated by Elasticsearch [16]. The higher the score, the better the result: the candidate with

15. *Cap* is a synonym of *pointe*.
16. https://www.elastic.co/guide/en/elasticsearch/reference/8.10/query-filter-context.html#

```
{
        "_source": {
                "properties": {
                        "nature": "Détail hydrographique",
                        "naturedetaillee": "Espace maritime",
                        "toponyme": "chaussée de sein"
                }
        }
}
{
        "_source": {
                "properties": {
                        "nature": "Détail hydrographique",
                        "naturedetaillee": "Espace maritime",
                        "toponyme": "pont des chats"
                }
        }
}
{
        "_source": {
                "properties": {
                        "nature": "Détail orographique",
                        "naturedetaillee": "Cap",
                        "toponyme": "pointe de la varde"
                }
        }
}
{
        "_source": {
                "properties": {
                        "nature": "Construction ponctuelle",
                        "naturedetaillee": "Phare",
                        "toponyme": "phare de rochebonne"
                }
        }
}
{
        "_source": {
                "properties": {
                        "nature": "Zone d'habitation",
                        "naturedetaillee": "Quartier",
                        "toponyme": "paramé"
                }
        }
}
```


Listing 6.3.6 – Extracts of five entries from the BD TOPO® cited in listing 6.3.5 as indexed in Elasticsearch.

the highest score becomes the top candidate.

The second scoring method is based on the geographic distance between each individual candidate selected from the database for the entity in question and the median location of all the candidates selected for all the other entities mentioned in the same paragraph. Thanks to the geographic organisation of the text of the *Instructions nautiques*, we are able to make the assumption that all the spatial entities mentioned in one paragraph are geographically located close to one another. To calculate the geographic distance it uses the values of the `"coordinates"` attribute, which are in the Lambert 93 projection. This means that the value of the score, which is equal to the geographic distance calculated, can be assimilated to be in metres. Therefore, the lower the score, the better the result: the candidate that has the lowest score (distance to the median location of all the candidates selected for all the other entities mentioned in the same paragraph) becomes the top candidate. If no candidate entries have been selected for any of the other entities mentioned in the same paragraph, the first scoring method is used to calculate a score for the entity in question and the candidate with the highest score becomes the top candidate.

Our second objective is to refine the candidate selection process by introducing a minimum or maximum (depending on the scoring method) score that an entry must have to be selected as a candidate. This is to eliminate the recurrent problem in the work by Loynes and Ruiz (2020), which selected entries with very low scores as candidates. This resulted in very improbable candidate entries featuring in the sets of selected candidates when there were no better candidates to be selected, when it would have been more appropriate to return no candidates at all.

6.3.2.3 Our Candidate Selection, Scoring and Ranking Algorithms

The following sections present our implementation of the four different algorithms that we apply to our own dataset:

— v0_name_AND, the original algorithm as published by Loynes and Ruiz (2020)

— v1_name+type_AND, our first adaptation of v0 that leverages the spatial entity type

— v2_name+type_AND_ScrLmt, an adaptation of v1 that introduces score limits

— v6_name+type_AND_ScrLmt_Bst40, an adaptation of v2 that introduces score boosts

`relevance-scores`

For each of the four algorithms, we implement one version with the Levenshtein distance candidate selection method and and the string similarity scoring method, one version with the n-gram candidate selection method and the string similarity scoring method, and one version with the Levenshtein distance candidate selection method and the geographic scoring method.

6.3.2.3.1 v0_name_AND This is the algorithm created and used by Loynes and Ruiz (2020). It measures the similarity between the name of the spatial entity extracted from the *Instructions nautiques* (the query string) and the value of the `"toponyme"` attribute of each entry in the BD TOPO® extract, either using the Levenshtein distance method or using the n-gram method. For the Levenshtein distance method Loynes and Ruiz used a fuzzy query [17], which is a term-level query [18] and therefore does not analyse the query string. Fuzziness is how Elasticsearch measures similarity via the Levenshtein distance. For the n-gram method they used a match query [19], which is a full-text query [20] that therefore analyses the query string. The analyser splits the query string into words on whitespace. Loynes and Ruiz used the 'AND' operator for the n-gram method, which requires all of the words in the query string to have a match within the n-grams of a given database entry.

6.3.2.3.2 v1_name+type_AND In our first iteration of an improved candidate selection algorithm, the similarity between the name of the spatial entity extracted from the *Instructions nautiques* and the value of the `"toponyme"` attribute of entries from the BD TOPO® extract is measured, as well as the similarity between the spatial entity type extracted from the *Instructions nautiques* and the values of the `"nature"` and `"naturedetaillee"` attributes. The three string similarities are measured either using the Levenshtein distance method or using the n-gram method. For the Levenshtein distance method we chose to change from a fuzzy term-level query to a match full-text query to be able to analyse the query strings. The query is composed of three subqueries, one for the `"toponyme"` attribute, one for the `"nature"` attribute and one for the `"naturedetaillee"` attribute. To manage the subqueries we create a compound Boolean query [21]. We assign the 'must' occurrence type to the subquery that measures the similarity

17. https://www.elastic.co/guide/en/elasticsearch/reference/8.10/query-dsl-fuzzy-query.html
18. https://www.elastic.co/guide/en/elasticsearch/reference/8.10/term-level-queries.html
19. https://www.elastic.co/guide/en/elasticsearch/reference/8.10/query-dsl-match-query.html
20. https://www.elastic.co/guide/en/elasticsearch/reference/8.10/full-text-queries.html
21. https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-bool-query.html

between the name of the spatial entity extracted from the *Instructions nautiques* and the value of the `"toponyme"` attribute of each entry in the database, which means that the subquery must be successful for the entry to be selected as a candidate. We assign the 'should' occurrence type to the two other subqueries, which measure the similarity between the spatial entity type extracted from the *Instructions nautiques* and the values of the `"nature"` and `"naturedetaillee"` attributes of each entry in the database. This means that neither of these two subqueries have to be successful for the entry to be selected as a candidate, but if either or both are then their string similarity scores will contribute to the overall string similarity score. The scores from the three subqueries are added together by the string similarity scoring method to give the final score for a candidate entry. The score is weighted evenly across the three subqueries. For the Levenshtein distance method we use the Elasticsearch fuzziness parameter [22] with the value 'AUTO', which generates a maximum Levenshtein distance based on the length of the word. We use the 'AND' operator in each subquery for both the Levenshtein distance method and the n-gram method.

6.3.2.3.3 v2_name+type_AND_ScrLmt Our second iteration is identical to the previous one but involves only selecting entries as candidates if their similarity score is superior or inferior (depending on the scoring method) to a given limit. For the Levenshtein distance and n-gram string similarity scoring methods we only keep entries whose score is equal or greater than 10. For the geographic scoring method we only keep entries whose score is inferior to 100,000 (which is roughly equal to a median distance of 100 kilometres). We also tested the effect of using the 'OR' operator, which requires only one of the words in the query string to have a match within the n-grams of a given database entry, but it did not yield interesting results.

6.3.2.3.4 v6_name+type_AND_ScrLmt_Bst40 The third improvement we make is to add weighted scoring to the subqueries. We tested the effect of multiplying the score of the `"nature"` and `"naturedetaillee"` attributes by 0.2, 0.4, 0.6 and 0.8 to decrease the impact of the scores of these two subqueries with respect to the score of the more important `"toponyme"` subquery. We did this by adding the 'boost' parameter to the two subqueries in question. For both the Levenshtein distance and the n-gram methods, a 'boost' parameter value of 0.4 gave the best results.

22. https://www.elastic.co/guide/en/elasticsearch/reference/8.10/common-options.html#fuzziness

6.4 Results

6.4.1 Structuring Information as RDF

Using our approach as described in section 6.3.1, implemented with SPARQL-Generate, we were able to structure all of the information extracted from the *Instructions nautiques* in chapter 5: the named and unnamed spatial entities, their type and the spatial relations between them. We also created instances to represent the provenance of each piece of information, and constructed triples to link each instance of a spatial entity to its source. The result is a set of 2258 RDF triples constructed according to the ATLANTIS Ontology.

The next step would be to clean the Simple Knowledge Organization System (SKOS) concepts created during the construction of the RDF triples and that are destined to be stored in the `atln:TypeOfSpatialEntity` SKOS thesaurus, which we described in section 4.4.4.4. We would need to indicate the language of each new concept (French, English, Arabic, etc.) and link synonyms to one another. This is an example of how the seed ontology created in chapter 4 can be enriched thanks to the extraction process carried out in chapter 5.

6.4.2 Entity Disambiguation

In this section we give the results of our implementation of the four entity disambiguation algorithms presented in section 6.3.2 on our dataset of RDF triples constructed according to the ATLANTIS Ontology. We implemented each of the four algorithms with the Levenshtein distance candidate selection method and and the string similarity scoring method, with the n-gram candidate selection method and the string similarity scoring method, and with the Levenshtein distance candidate selection method and the geographic scoring method. We evaluate the algorithms using the evaluation metrics presented below.

6.4.2.1 Evaluation Metrics

Like Loynes and Ruiz (2020), we adopt the eight entity disambiguation evaluation metrics introduced by Brando et al. (2016), which we present below.

6.4.2.1.1 CardM Average cardinality of the sets of candidates (CardM) is the total number of candidates returned divided by the total number of entities. It gives an indication of the ability of the database to provide

candidates for the candidate selection step. If the value is very low, it indicates that the database does not contain enough relevant entries with respect to the queries. If the value is very high, it indicates that the database contains too many homonymous entries.

6.4.2.1.2 PrecCand Candidate precision (PrecCand) is the number of entities for which candidates have been found and for which the correct candidate is present in the selection, divided by the number of entities for which candidates have been found.

6.4.2.1.3 RapCand Candidate recall (RapCand) is the number of entities for which candidates have been found and for which the correct candidate is present in the selection, divided by the number of entities for which a non-nil gold reference exists.

6.4.2.1.4 PrecN "nil" precision (PrecN) is the number of entities for which zero candidates have been found and for which no gold reference exists, divided by the number of entities for which zero candidates have been found.

6.4.2.1.5 RapN "nil" recall (RapN) is the number of entities for which zero candidates have been found and for which no gold reference exists, divided by the number of entities for which no gold reference exists.

6.4.2.1.6 ExactD Disambiguation accuracy (ExactD) is the number of entities for which a non-nil gold reference exists and for which the correct candidate has been selected as the top candidate, divided by the number of entities for which candidates have been found and for which the correct candidate is present in the selection. It gives an indication of the accuracy of the candidate scoring and ranking method that is independent of the accuracy of the candidate selection method.

6.4.2.1.7 ExactG Global accuracy (ExactG) is the number of entities for which a non-nil gold reference exists and for which the correct candidate has been selected as the top candidate, divided by the number of entities for which a non-nil gold reference exists. It gives an indication of the overall accuracy of the candidate selection method together with the candidate scoring and ranking method to select the correct candidate as the top candidate, without taking into account entities whose gold reference is "nil".

6.4.2.1.8 ExactN Global accuracy including "nil" (ExactN) is the number of entities for which the correct candidate (including "nil") has been selected as the top candidate, divided by the total number of entities. It gives an indication of the overall accuracy of the candidate selection method together with the candidate scoring and ranking method to select the correct candidate (including "nil") as the top candidate. We consider this to be the most important evaluation metric.

6.4.2.1.9 F1-scores We also calculate a candidate F1-score (CandF1) from PrecCand and RapCand, and a "nil" F1-score (NilF1) from PrecN and RapN.

6.4.2.2 Levenshtein Distance Method with Default Scoring

Tables 6.4.1, 6.4.2 and 6.4.3 show the results for each of the four algorithms implemented with the Levenshtein distance candidate selection method and the string similarity scoring method.The first row in each table shows the results that we obtained on our own dataset using the algorithm developed by Loynes and Ruiz (2020), which we presented in section 6.3.2.3.1. Rows two, three and four in each table show the results that we obtained on our own dataset using our algorithms, which we presented in sections 6.3.2.3.2, 6.3.2.3.3 and 6.3.2.3.4 respectively.

Levenshtein distance + default scoring	CardM	PrecCand	RapCand	CandF1
v0_name_AND	1.94	0.22	0.08	0.12
v1_name+type_AND	3.04	0.47	**0.62**	0.54
v2_name+type_AND_ScrLmt	1.43	0.52	0.53	0.53
v6_name+type_AND_ScrLmt_Bst40	1.28	**0.56**	0.53	**0.55**

Table 6.4.1 – Candidate cardinality, precision, recall and F1-score results for the Levenshtein distance candidate selection method with the string similarity scoring method. The highest score obtained on our dataset for the latter three evaluation metrics is in **bold**.

Levenshtein distance + default scoring	PrecN	RapN	NilF1
v0_name_AND	0.61	0.85	0.71
v1_name+type_AND	**0.79**	0.64	0.71
v2_name+type_AND_ScrLmt	**0.79**	0.79	0.79
v6_name+type_AND_ScrLmt_Bst40	0.78	**0.81**	**0.80**

Table 6.4.2 – "nil" precision, recall and F1-score results for the Levenshtein distance candidate selection method with the string similarity scoring method. The highest score obtained on our dataset for each of the evaluation metrics is in **bold**.

Levenshtein distance + default scoring	ExactD	ExactG	ExactN
v0_name_AND	0.25	0.02	0.53
v1_name+type_AND	0.82	**0.51**	0.59
v2_name+type_AND_ScrLmt	**0.96**	**0.51**	0.68
v6_name+type_AND_ScrLmt_Bst40	**0.96**	**0.51**	**0.70**

Table 6.4.3 – Disambiguation, global and global including "nil" accuracy results for the Levenshtein distance candidate selection method with the string similarity scoring method. The highest score obtained on our dataset for each of the evaluation metrics is in **bold**.

We see a significant improvement in most results between the v0 algorithm and our v1 algorithm, particularly for the candidate precision, recall and F1-score results. This demonstrates the benefit of including a measurement of the similarity between the spatial entity type extracted from the *Instructions nautiques* and the values of the `"nature"` and `"naturedetaillee"` attributes of entries in the database.

For example, the v0 algorithm returns no candidates with the query string "port de Douarnenez" for the `"toponyme"` attribute. However, our v1 algorithm with the subquery string "port de Douarnenez" for the `"toponyme"` attribute and the subquery string "port" for the `"nature"` and `"naturedetaillee"` attributes returns three candidates:

```
{ "nature": "Equipement de transport",
  "naturedetaillee": "Port",
  "toponyme": "port de pêche de douarnenez" }

{ "nature": "Zone d'activité ou d'intérêt",
  "naturedetaillee": "Musée",
  "toponyme": "port-musée de douarnenez" }

{ "nature": "Zone d'activité ou d'intérêt",
  "naturedetaillee": "Capitainerie",
  "toponyme": "capitainerie du port de pêche de douarnenez" }
```

with scores of 16.31, 11.41 and 8.86 respectively. Our v1 algorithm was able to return candidates with these subqueries although the v0 algorithm was not thanks to our use of a term-level query in our v1 algorithm instead of a full-text query like in the v0 algorithm. This means that the v1 algorithm processes each word within the subquery string individually, trying to match it with the words in the corresponding attribute values, instead of processing the entire subquery string at once. Our v1 algorithm is therefore more lenient than the v0 algorithm. Our v1 algorithm was also able to assign a higher score to the correct entry thanks to the subquery that measures the similarity between the `"nature"` and `"naturedetaillee"`

attributes and the second subquery string “port”, despite the value of its `"toponyme"` attribute being closer to the first subquery string “port de Douarnenez”.

We see an overall improvement in most results between our v1 and v2 algorithms. Avoiding the selection of candidates with particularly low scores decreases the selection of improbable candidates as top candidates and therefore has a positive effect on the results. For example, our v1 algorithm returns only one entry when supplied with the subquery strings “îles Scilly” for the `"toponyme"` attribute and “îles” for the `"nature"` and `"naturedetaillee"` attributes:

```
{ "nature": "Zone d’activité ou d’intérêt",
  "naturedetaillee": "Zone industrielle",
  "toponyme": "parc d’activités les villes billy" }
```

The attribute values have little resemblance with the corresponding subquery strings and so this entry is given a score of 5.19. However, our v1 algorithm still selected it as the top candidate because it is the only candidate to have been selected from the database. This entry is not selected by our v2 algorithm because of its very low score (< 10), which means that the result is now correct as no gold reference exists for this entity in the database.

We see a slight improvement in most results between our v2 and v6 algorithms, making it the most successful algorithm on our dataset that uses the Levenshtein distance method and the default Elasticsearch relevance score method. It decreases the relative importance of the subqueries on the `"nature"` and `"naturedetaillee"` attributes compared to the subquery on the `"toponyme"` attribute. This is beneficial because it is more important for the names to be similar than for the types to be similar. For example, our v2 algorithm, which assigns equal importance to the three subqueries, returns multiple entries for the subquery strings “La Rocque Point” and “Point”, the top candidate being:

```
{ "nature": "Construction surfacique",
  "naturedetaillee": "Pont",
  "toponyme": "pont de la rocade sud" }
```

with a score of 10.82. Although the value of the `"naturedetaillee"` attribute is very similar to the subquery string “Point”, the value of the `"toponyme"` does not have much resemblance. Our v6 algorithm gives this entity a lower score of 8.33, meaning that it is not selected as a candidate and therefore that the result is now correct as no gold reference exists for this entity in the database.

6.4.2.3 N-Gram Method with Default Scoring

Tables 6.4.4, 6.4.5 and 6.4.6 show the results for the n-gram method with the string similarity scoring method.

N-gram + default scoring	CardM	PrecCand	RapCand	CandF1
v0_name_AND	1.95	**0.56**	**0.61**	**0.59**
v1_name+type_AND	1.95	**0.56**	**0.61**	**0.59**
v2_name+type_AND_ScrLmt	1.21	**0.56**	0.54	0.55
v6_name+type_AND_ScrLmt_Bst40	1.21	**0.56**	0.54	0.55

Table 6.4.4 – Candidate cardinality, precision, recall and F1-score results for the n-gram candidate selection method with the string similarity scoring method. The highest score obtained on our dataset for the latter three evaluation metrics is in **bold**.

N-gram + default scoring	PrecN	RapN	NilF1
v0_name_AND	**0.81**	0.77	**0.79**
v1_name+type_AND	**0.81**	0.77	**0.79**
v2_name+type_AND_ScrLmt	0.79	**0.80**	**0.79**
v6_name+type_AND_ScrLmt_Bst40	0.79	**0.80**	**0.79**

Table 6.4.5 – "nil" precision, recall and F1-score results for the n-gram candidate selection method with the string similarity scoring method. The highest score obtained on our dataset for each of the evaluation metrics is in **bold**.

N-gram + default scoring	ExactD	ExactG	ExactN
v0_name_AND	0.75	0.46	0.65
v1_name+type_AND	0.79	**0.48**	0.66
v2_name+type_AND_ScrLmt	**0.87**	0.47	**0.67**
v6_name+type_AND_ScrLmt_Bst40	**0.87**	0.47	**0.67**

Table 6.4.6 – Disambiguation, global and global including "nil" accuracy results for the n-gram candidate selection method with the string similarity scoring method. The highest score obtained on our dataset for each of the evaluation metrics is in **bold**.

Between the v0 algorithm and our v1 algorithm on our own dataset, we only see improvements in the three accuracy measurements. The other scores remain constant. This means that the set of candidates selected for each entity has not much changed, but that the algorithm is more successful at selecting the correct candidate as the top candidate. For example, the subquery string "phare de Rochebonne" for the `"toponyme"` attribute returned two candidates with identical scores of 16.94 with the v0 algorithm. Our v1 algorithm returned the same two candidates with the same string for the `"toponyme"` attribute subquery and the additional subquery string "phare" for the other two attributes, but this time with different scores. The correct entry:

```
{ "nature": "Construction ponctuelle",
  "naturedetaillee": "Phare",
  "toponyme": "phare de rochebonne" }
```


was returned with a score of 25.59, whilst the other almost identical entry:


```
{ "nature": "Détail hydrographique",
  "naturedetaillee": "Feu",
  "toponyme": "phare de rochebonne" }
```


was returned with a score of 16.94. As opposed to the v0 algorithm, our v1 algorithm is able to distinguish between the two very similar entries although thanks to the subquery that measures the similarity between the `"naturedetaillee"` attribute and the entity type extracted from the *Instructions nautiques*.

We see very little difference in most results between our v1, v2 and v6 algorithms for the n-gram method with the string similarity scoring method. The results are also very similar to those obtained with the Levenshtein distance method with the string similarity scoring method. To improve the performance of the n-gram method, the settings would need to be refined and adapted to the gram composition of the spatial entity names and types extracted from the *Instructions nautiques*.

6.4.2.4 Levenshtein Distance Method with Geographic Scoring

Tables 6.4.7, 6.4.8 and 6.4.9 show the results for the Levenshtein distance method with the geographic scoring method.

Levenshtein distance + geographic scoring	CardM	PrecCand	RapCand	CandF1
v0_name_AND	1.03	0.00	0.00	0.00
v1_name+type_AND	3.04	0.47	**0.62**	**0.54**
v2_name+type_AND_ScrLmt	1.20	**0.52**	0.22	0.31
v6_name+type_AND_ScrLmt_Bst40	1.23	0.48	0.23	0.31

Table 6.4.7 – Candidate cardinality, precision, recall and F1-score results for the Levenshtein distance candidate selection method with the geographic scoring method. The highest score obtained on our dataset for the latter three evaluation metrics is in **bold**.

On the whole, our v1 and v2 algorithms achieve lower results than those obtained with the Levenshtein distance or n-gram methods with the string similarity scoring method. However, the "nil" F1-score and the global accuracy including "nil" score obtained with our v2 algorithm are very close to those obtained with the Levenshtein distance and n-gram methods with the string similarity scoring method. To be in a position to improve the performance of the geographic scoring method, it would be useful to

Levenshtein distance + geographic scoring	PrecN	RapN	NilF1
v0_name_AND	0.61	**0.98**	0.75
v1_name+type_AND	**0.79**	0.64	0.71
v2_name+type_AND_ScrLmt	0.68	0.93	**0.78**
v6_name+type_AND_ScrLmt_Bst40	0.67	0.89	0.77

Table 6.4.8 – "nil" precision, recall and F1-score results for the Levenshtein distance candidate selection method with the geographic scoring method. The highest score obtained on our dataset for each of the evaluation metrics is in **bold**.

Levenshtein distance + geographic scoring	ExactD	ExactG	ExactN
v0_name_AND	0.00	0.00	0.60
v1_name+type_AND	0.52	**0.32**	0.52
v2_name+type_AND_ScrLmt	**0.55**	0.12	**0.62**
v6_name+type_AND_ScrLmt_Bst40	0.52	0.12	0.60

Table 6.4.9 – Disambiguation, global and global including "nil" accuracy results for the Levenshtein distance candidate selection method with the geographic scoring method. The highest score obtained on our dataset for each of the evaluation metrics is in **bold**.

have a gold dataset with a higher proportion of non-nil gold references. If the results of this algorithm were sufficiently improved, it could be applied to the unnamed spatial entities extracted from the text in an effort to disambiguate them thanks to the geographic location of the named spatial entities mentioned in the same paragraph.

6.5 Exploitation of the Results

We have demonstrated that it is possible to use off-the-shelf tools to structure nested information extracted from text using our implementation of PURE, and then to link instances of entities to their corresponding entries in a reference database resource, to populate an ontology and construct a knowledge graph.

By refining these two processes and applying them, as well as our extraction approach, to the entire collection of *Instructions nautiques*, it would be possible to populate the ATLANTIS Ontology with all of the spatial entities contained within the *Instructions nautiques* and thereby construct a basis for the ATLANTIS Knowledge Graph.

6.5.1 Nereus Web Platform

As discussed in section 2.5, the *Service hydrographique et océanographique de la Marine* (Shom) would like to provide users of the *Instructions nautiques* with access to their content in novel ways. Also, as discussed in

section 4.4.3.1, most of the current users of the *Instructions nautiques* that we interviewed expressed an interest in having access to the content of the *Instructions nautiques* through a digital tool, and that based on an interactive nautical chart. However, they also mentioned wanting to maintain the possibility to consult the full text of the *Instructions nautiques*.

In this thesis we have presented a method that could be used to construct a geospatial knowledge graph of the *Instructions nautiques*, the content of such a knowledge graph would only be directly accessible via SPARQL queries. The average user of the *Instructions nautiques* is not expected to be familiar with the SPARQL query language, meaning that there would need to be a user-friendly tool that allowed easy access to the content of the knowledge graph without the use of SPARQL for it to be useful for users of the *Instructions nautiques*.

In 2022 we supervised a development project carried out by Alla et al. (2022) as part of their Master degree. They developed a prototype of a Web platform called Nereus [23] that allows access to the content of the ATLANTIS Knowledge Graph [24] via a graphical user interface (GUI) without having to write SPARQL queries from scratch. The background of the GUI is a nautical chart on which the disambiguated spatial entities contained within the knowledge graph can be displayed and consulted. It also features a panel that can display a PDF of an *Instructions nautiques* volume. Finally, there is a component that allows users to visually construct SPARQL queries using buttons and drop-down menus without using the SPARQL language. This component allows users to carry out searches for elements such as specific spatial entity types, specific named spatial entities and navigation instructions. Figure 6.5.1 shows the visual query component on the nautical chart background, which is obtained from `data.shom.fr` via a Web Map Service (WMS) flux. Figure 6.5.2 shows the PDF panel on the right, the results of a query on the left, and in the middle is a pop-up that is displayed when a spatial entity on the chart is clicked on and that presents the knowledge contained within selected RDF triples associated to the entity in question in the knowledge graph.

The query component was developed using Sparnatural [25], an adaptable JavaScript component that allows visually navigating and querying a knowledge graph. Users can specify the geographic boundaries of a query by drawing a bounding box on the chart. When a user clicks on a piece of information such as the name of a spatial entity or an instruction, the PDF

23. https://github.com/mcharzat/SHOM_IN

24. They carried out their work on a sample ATLANTIS Knowledge Graph, constructed by manually populating the ATLANTIS Ontology with RDF triples including geographic coordinates, like those produced in this chapter.

25. https://sparnatural.eu/

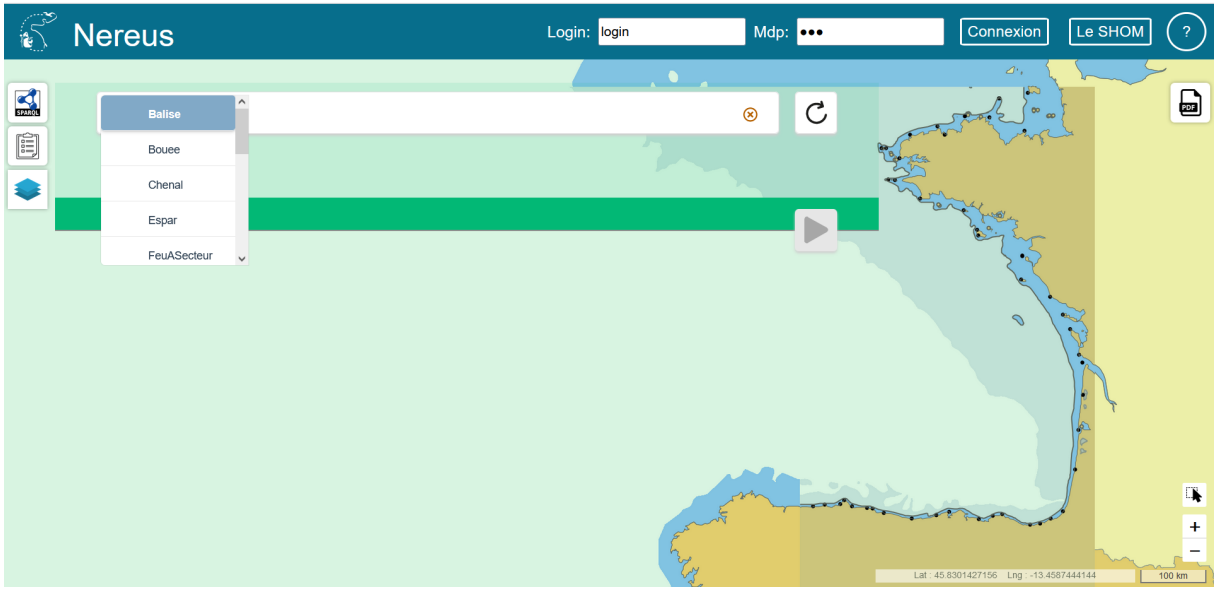


Figure 6.5.1 – Screenshot of the Nereus Web platform showing a SPARQL query being constructed using the Sparnatural component (Alla et al. 2022).

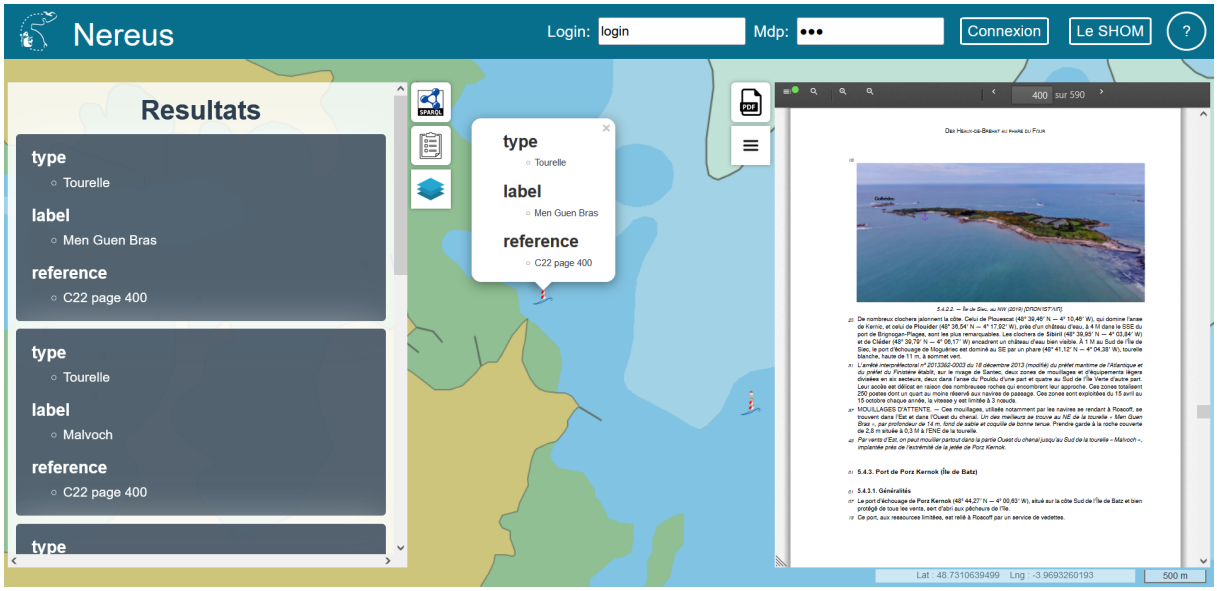


Figure 6.5.2 – Screenshot of the Nereus Web platform showing the results of a query and the corresponding page in a PDF of the *Instructions nautiques* (Alla et al. 2022).

of the corresponding *Instructions nautiques* volume automatically appears at the page from which the information was extracted.

This functional prototype of a Web platform that allows visually accessing the content of a geospatial knowledge graph demonstrates one of the advantages of structuring the content of the *Instructions nautiques* for users.

6.6 Conclusion

In this chapter we presented a proof of concept for automatically populating a geospatial knowledge graph from information that has been extracted from text using our extraction approach based on PURE (Zhong and D. Chen 2021), which we presented in chapter 5. This third and final stage of the ATONTE Methodology involves two main steps: automatically structuring information as RDF triples and then automatically disambiguating the instances of entities by linking them to their corresponding entries in reference resources. We implemented both of these steps using dedicated off-the-shelf tools.

For the first step of the proof of concept we used the JSONL-formatted spatial entities and relations extracted from the *Instructions nautiques* using our extraction approach based on PURE (Zhong and D. Chen 2021) as demonstrated in the previous chapter. By executing queries written with SPARQL-Generate we were able to structure the spatial entities and their relations as RDF triples according to the ATLANTIS Ontology that we developed in chapter 4. Each spatial entity is assigned a unique IRI and is associated to its paragraph of origin within the *Instructions nautiques*, and where the information was available in the text it is also associated to its name, its geographic feature type, and its spatial relations with other spatial entities.

For the second step of the proof of concept we built upon the work of Loynes and Ruiz (2020) on entity disambiguation using Elasticsearch. We applied their entity disambiguation algorithms, which involve Elasticsearch queries, to the named spatial entity RDF triples that we constructed in the previous step to link them to their corresponding entries in the BD TOPO® extract. We make two principal modifications to Loynes and Ruiz's algorithms. First, we add a measurement of the similarity between the entity type extracted from the *Instructions nautiques* and the feature type of the entries contained within the BD TOPO® instead of relying entirely on the similarity between their names. Second, we add limits to the candidate selection process to avoid selecting very improbable entries

as candidates. These two changes significantly improve the ExactN score, which evaluates the global ability of the algorithm to select the correct reference entry, or no reference entry at all if none correspond, of each of the three combinations of candidate selection methods and scoring methods compared to the original versions presented by Loynes and Ruiz. The ExactN score for the Levenshtein distance method with default scoring increased from 0.53 to 0.70, for the n-gram method with default scoring it increased from 0.65 to 0.67 and for the Levenshtein distance method with geographic scoring it increased from 0.52 to 0.62 on our own dataset.

The prototype Nereus Web platform developed by Alla et al. (2022) demonstrates that the results of our structuring and disambiguation approach are directly exploitable within a query and visualisation tool that can be used by novices.

In the following final chapter we will conclude our work on the creation of the ATONTE Methodology and our application of it to the *Instructions nautiques*. We will discuss and evaluate our methodology, the ATLANTIS Ontology that we developed thanks to it and the future ATLANTIS Knowledge Graph for which we provided a proof of concept in this chapter.

Chapter 7

Conclusion and Future Work

7.1 Overview and Contributions

The objective of this thesis was to provide a functional approach for constructing geospatial knowledge graphs from heterogeneous text-based sources that is suited to applications that require performing direct and indirect spatial reasoning. In view of the objective, the three main scientific challenges that we set ourselves were:

1. How can we to acquire a domain ontology suited to the text corpus to serve as a structure for the geospatial knowledge graph?
2. What techniques for spatial information extraction can we apply to heterogeneous text-based sources and do they need to be refined for this use case?
3. How can we automatically structure the extracted spatial information according to the domain ontology to populate it and therefore construct the geospatial knowledge graph, and how can we disambiguate the spatial entities and link them to a reference geographic resource?

The main contributions of this thesis are:

— The ATlantis Ontology and kNowledge graph development from Texts and Experts (ATONTE) Methodology, which is a methodology for constructing knowledge graphs, geospatial or not, from heterogeneous textual sources, expert knowledge and reference data (Rawsthorne et al. 2022b)

— The coAsTaL mAritime NavigaTion InstructionS (ATLANTIS) Ontology [1], which is a multilingual geospatial domain seed ontology that covers the scope of the *Instructions nautiques* (Rawsthorne et al. 2022a)

1. https://github.com/umrlastig/atlantis-ontology

— The ATLANTIS Dataset [2], which is a manually-annotated French-language benchmark dataset on the maritime domain that can be used to train algorithms for nested spatial entity and binary spatial relation extraction from text, with benchmark results (Rawsthorne et al. 2023)

— The TextMine'24 Dataset [3], which is a subset of the ATLANTIS Dataset and contains nested spatial entity annotations (Rawsthorne et al. 2024a)

We developed the ATONTE Methodology empirically. It is the result of our efforts to construct a geospatial knowledge graph of the contents of the *Instructions nautiques* with the help of domain experts and reference geographic resources. The ATONTE Methodology, of which we presented an overview in chapter 3, is composed of three stages:

1. Ontology Development from Text and Experts, which we presented in chapter 4
2. Entity and Relation Extraction from Text, which we presented in chapter 5
3. Information Structuring and Entity Disambiguation, which we presented in chapter 6

For each stage, in its corresponding chapter, we also demonstrated how we implemented it in practice on our corpus of *Instructions nautiques* to create the ATLANTIS Ontology and an operational basis for the ATLANTIS Knowledge Graph. The entire ATONTE Methodology has therefore been shown to be operational for the task of creating a knowledge graph from text and experts, and in this case a *geospatial* knowledge graph.

The first stage of the ATONTE Methodology is dedicated to the manual development of domain ontologies, geospatial or not, from text and experts. It is suitable for use in cases where the content of a text corpus needs to be modelled ontologically according to a structure defined in conjunction with domain experts. ATONTE is loosely based on Simple Agile Methodology for Ontology Development (SAMOD) (Peroni 2016a) and incorporates elements from Modular Ontology Modeling (MOMo) (Shimizu et al. 2022) and Networked Ontologies (NeOn) (Suárez-Figueroa et al. 2012). One of the main characteristics of ATONTE is that it requires working directly with the text corpus from the beginning and informally modelling its content as semantic triples before the implementation process, meaning that the ontology is built from the text itself. The programmed involvement of

2. https://github.com/umrlastig/atlantis-dataset
3. https://www.kaggle.com/competitions/defi-textmine-2024

domain experts and final users provide the ontology developers with targeted assistance for the refinement of the ontological model whilst taking into account the needs of the users.

By applying the first stage of the ATONTE Methodology to our corpus of *Instructions nautiques* we developed the ATLANTIS Ontology, a geospatial seed ontology that covers the content of the *Instructions nautiques*. It can be used to represent the spatial entities present in the coastal maritime environment, the spatial relations between them, coastal navigation guidelines, the vessels used for coastal navigation, temporalities, and meteorological and oceanographic phenomena.

Benoit and Kergus (2022) carried out a study to determine whether the ATLANTIS Ontology could be used to model other Sailing Directions, as presented in section 4.6.2 on page 101 onwards. Using the ATLANTIS Ontology as a basis for their work, they applied the first component of the ATONTE Methodology, which is our ontology development methodology that we described in section 4.3 on page 54 onwards, to three English-language Sailing Directions publications: the *United States Coast Pilot*, the *Sailing Directions (Enroute)* and the *Canadian Sailing Directions*. Their work demonstrates the relevance of the ATLANTIS Ontology to international Sailing Directions, and is also a first demonstration that our ontology development methodology is repeatable and that it can be applied to manually enrich existing ontologies as well as creating ontologies from scratch.

The second stage of the ATONTE Methodology is dedicated to the automatic extraction of nested entities and binary relations from text using a supervised deep learning approach. It is an adaptation of the Princeton University Relation Extraction system (PURE) (Zhong and D. Chen 2021), which deals with the extraction of flat entities and binary relations. The modification that we made to PURE to adapt it to the extraction of nested as opposed to flat entities resides in the nested labelling that we applied to the annotated dataset.

By applying the second stage of the ATONTE Methodology to our corpus of *Instructions nautiques*, we demonstrated that PURE is suited to the extraction of *spatial* entities and relations. This involved creating the ATLANTIS Dataset, a French-language dataset composed of extracts from the *Instructions nautiques* that we annotated manually with nested labelling, and using it to train multilingual and monolingual French Bidirectional Encoder Representations from Transformers (BERT) deep language models for the tasks of nested spatial entity and binary spatial relation extraction from text. Our benchmark results for this approach on our ATLANTIS Dataset show that the multilingual BERT model outperforms the

monolingual French BERT model for entity extraction, whilst for relation extraction and end-to-end entity and relation extraction the monolingual French BERT model performs best. The results of our experiments reveal that making cross-sentence context information available during training and prediction favours entity extraction but hinders relation extraction.

The third stage of the ATONTE Methodology is dedicated to the automatic structuring of the information extracted from text during the previous stage of the ATONTE Methodology as Resource Description Framework (RDF) triples, and the disambiguation of the entities by linking them to their corresponding entries in a reference resource.

By applying the third stage of the ATONTE Methodology to the spatial entities and relations extracted from our corpus of *Instructions nautiques*, we produced an operational basis for the ATLANTIS Knowledge Graph. The first step involved extracting the spatial entities and relations from the JSON Lines (JSONL) files and structuring them according to the ATLANTIS Ontology, taking advantage of the nested labelling of the entities to associate them to their correct type. We implemented this step using SPARQL-Generate. The second step is based on the work by Loynes and Ruiz (2020) and involved matching the RDF-structured spatial entities to their corresponding entries in the BD TOPO®, taking advantage of not only the entity name but also its type to refine the search results. We implemented this step using Python scripts and Elasticsearch.

Alla et al. (2022) carried out a project to develop a prototype Web platform that allows access to the content of the ATLANTIS Knowledge Graph via a graphical user interface (GUI) without having to write SPARQL queries from scratch. The result is a platform that allows navigating and querying the ATLANTIS Knowledge Graph using drop-down menus and an interactive nautical chart.

7.2 Integration of ATONTE and ATLANTIS within the Shom

During the interviews that we carried out with users of the *Instructions nautiques*, which we described in detail in section 4.4.3.1.6 from page 75 onwards, we were told that they would find it useful to be able watch to videos of other vessels entering ports, to gain a better understanding of the environment and be better equipped to plan the manoeuvre using their own vessel. Following reception of a report of the results of the interviews that we carried out, the *Service hydrographique et océanographique de la Marine* (Shom) has recorded videos from onboard a vessel of the ports of Roscoff,

Trégastel and Saint-Malo in Brittany [4]. In each video they have integrated the names of the landmarks that feature on the route, within the ports they have included the names of the quays and significant buildings in the surrounding area, and they highlight the leading lines (the landmarks involved and the angle) used to navigate. At the bottom of the video is a nautical chart that follows the trajectory of the vessel and displays its course. The Shom is also currently experimenting with the possibility of making the photos in the *Instructions nautiques* interactive, giving users the possibility to click on objects within the photo to get more information about them. They have the advantage of requiring less storage space than the videos.

The Shom has recruited a full-time data scientist on a long-term contract to continue the research carried out in this thesis. The role of the data scientist is to develop our work to create a more complete tool for populating the knowledge graph, test it on a sufficiently large study area and assess the quality of the result, in order to validate the ATLANTIS Ontology and the ATONTE Methodology more generally. The Shom would then like the data scientist to develop a prototype system that will enable a nautical knowledge management and processing chain to be consolidated and validated in a representative environment, for the agile production and linking of other publications, products and services.

7.3 Future Work

Regarding the ATONTE Methodology, future work includes applying it to other corpora that cover other domains, to assess its generalisability. In addition, evaluation steps should be added to each stage of the methodology. These steps would allow ensuring that the model built and the data produced to populate it are coherent and appropriate for the context to which they are to be applied.

Regarding the second stage of the ATONTE Methodology in particular, future work includes reducing the time required to annotate the training dataset. To achieve this, an annotation tool that offers either machine learning-assisted labelling or automatic annotation according to a user-defined set of keywords could be used. The annotations produced with different tools would need to be evaluated and compared to verify that their quality does not drop below that of the annotations that we produced entirely by hand. With a view to increasing the reliability of the annotated dataset, we could adopt a multi-annotator approach and calculate inter-

4. These videos have not yet been published by the Shom.

annotator agreement. We could also automatically produce synthetic data in the form of annotated text via regular expressions. Alternatively, an unsupervised learning step could be combined with a supervised learning step using a smaller manually-annotated dataset in the extraction approach. If the results are equivalent to or better than those obtained with our current approach then the unsupervised learning step could be integrated to our methodology and the size of manually-annotated dataset required could be reduced. Another idea for future work is the extension of the approach for extracting information from text to take into account several types of entity simultaneously and to deal with n-ary relationships (ternary or higher). This would allow extracting relations that involve more than two entities like the spatial relation *between*, such as in the following sentence: "A beacon is located between the Green Rock and the island." It could also be extended to manage intrinsic and relative directional spatial relations, as well as distance spatial relations.

The third stage of the ATONTE Methodology could be improved by making use of the knowledge contained in the triples that define a spatial relation between entities to aid the disambiguation process. The possible location of an entity according to its position relative to other entities could be integrated to the evaluation of candidate entities. This would influence the entity ranking and could be set to improve the score of candidate entities whose geographic position is close to the possible location calculated. The disambiguation of unnamed entities should be added to the ATONTE Methodology. This could be achieved by exploiting the entity type, the advantage of which for named entities we demonstrated in chapter 6, as well as analysing its spatial relations with other entities. The measurement of the similarity between the entity type extracted from the corpus and the type of the entries contained within the reference resource could be improved in two ways. A thesaurus of entity type synonyms, could be integrated to the comparison process, or a calculation of the semantic similarity between the entity types (Mustière et al. 2011). Both have the advantage of not penalising the use of alternative but equivalent terms, such as *cape* and *headland*.

Regarding the implementation of the ATONTE Methodology on the *Instructions nautiques*, future work includes applying the second and third stages to the remaining subdomains of the ATLANTIS ontology: Maritime Navigation Guidelines, Temporalities, Meteorological and Oceanographic Phenomena and Maritime Vessels, and completing their application to the Maritime Spatial Entities and Spatial Relations subdomain. Achieving this would require adapting both stages to other types of entities and rela-

tions. This would allow extending the operational basis of the ATLANTIS Knowledge Graph that we produced by extracting only the spatial entities and relations from the Maritime Spatial Entities and Spatial Relations subdomain to cover the entire domain of the *Instructions nautiques*. The ATLANTIS Ontology could in turn be enriched with new concepts, classes and properties as they are extracted from the text. The ATLANTIS Knowledge Graph could later be enriched with information from sources other than the *Instructions nautiques*: either other publications by the Shom or external sources such as the weather forecast.

Appendix A

Interview Questionnaire

In total we interviewed 12 people with different levels of experience of using the *Instructions nautiques* in their studies or for their work, in the military or civilian domain. We held 10 interviews of 30 to 60 minutes long with individuals or small groups, in person where possible and otherwise virtually. We spoke with five students and four military instructors at the *École navale*, the French Naval Academy, and three civilian instructors from the *École nationale supérieure maritime* (ENSM), the French National Maritime Academy where merchant navy officers are trained.

The French-language questionnaire that we used during these interviews is included below.

Présentation du projet

Helen Rawsthorne a commencé une thèse en novembre 2020 portant sur la création de bases de connaissances topographiques à partir de sources hétérogènes. Ces travaux impliquent de proposer une approche pour l'extraction (semi-)automatique de connaissances géographiques de textes. Les textes principaux sur lesquels nous travaillons sont les ouvrages *Instructions nautiques* du Service hydrographique et océanographique de la Marine (Shom). Le but est donc d'extraire, d'organiser et de stocker numériquement les informations contenues dans ces ouvrages, de manière à pouvoir les réutiliser différemment, en dehors de leur format habituel. Par exemple, on pourrait envisager de proposer aux utilisateurs des moyens plus efficaces et rapides pour accéder aux informations contenues dans les ouvrages sans prendre le temps nécessaire pour lire l'intégralité des *Instructions nautiques*.

Dans le cadre de ces travaux, nous sommes actuellement en train d'analyser les besoins des utilisateurs des *Instructions nautiques* afin de pouvoir orienter notre travail vers des solutions qui seront réellement utiles. C'est dans ce cadre que cette enquête se situe.

Déroulé de l'entretien auprès des instructeurs

Partie présentation

1. Pouvez-vous vous présenter brièvement ?
2. Depuis combien d'années êtes-vous instructeur ?

Pratiques professionnelles (hors enseignement)

3. Dans votre expérience professionnelle (hors enseignements), comment et pourquoi (pour quels besoins) utilisez-vous les *Instructions nautiques* en appui aux outils de navigation que vous avez ? Que recherchez-vous comme informations dans les *Instructions nautiques* ?
4. Identifiez-vous des cas ou les *Instructions nautiques* posent problème ? (Informations erronées, périmées, ambigües, manquantes, etc.)
5. Avez-vous l'impression que votre usage des *Instructions nautiques* a changé au cours du temps, avec l'expérience et/ou avec l'apparition de nouveaux outils ? Par exemple, y'a-t-il des informations que vous n'allez plus rechercher dans les *Instructions nautiques*, et si oui, pourquoi ? Ou au contraire, certaines informations vous semblent-elles primordiales aujourd'hui ?
6. Quelle documentation utilisez-vous à côté des *Instructions nautiques* ? Énumérer les noms, sources, types, et pourquoi elles viennent compléter les *Instructions nautiques*.
7. Identifiez-vous des cas où les *Instructions nautiques* ne suffisent pas ?
8. Utilisez-vous d'autres logiciels de navigation dans un cadre plus privé ? Si oui, pourquoi ? Meilleure interface, données plus précises, autre ?
   (a) Éventuellement : L'information proposée est-elle différente ? Que trouvez-vous mieux côté interfaces ? Utilisez-vous également *Instructions nautiques* ? Si y'a-t'il malgré tout des manques et que proposeriez-vous pour les combler ?

Pratiques professionnelles pour l'enseignement

9. Dans le cadre de vos enseignements, comment utilisez-vous les *Instructions nautiques* ? Pour répondre à cette question, pourriez-vous détailler un scénario d'usage typique des cartes et des IN, ou ceux scénarios d'usage si vous estimez que c'est bien différent lors de la formation théorique et en mer ?

10. Dans vos enseignements, vous utilisez un « navigateur électronique ». Quels sont les besoins supplémentaires nécessitant l'usage des *Instructions nautiques* ?
11. Identifiez-vous chez les élèves des difficultés récurrentes, ou plus occasionnelles, à la compréhension des *Instructions nautiques*, ou à leurs usages ?
12. Pour vous, y'a-t-il des manques dans les *Instructions nautiques*, que ce soit dans vos enseignements théoriques et/ou pratiques ?
13. Quels sont les retours des élèves lorsque vous leur donnez à utiliser des *Instructions nautiques* (questions, commentaires, manques ou dans ce que vous pouvez voir dans leurs pratiques) ?
14. Lors de vos sorties, vous utilisez des outils numériques (navigateurs, etc.) et quelles sont les données qui vous manquent et qui sont présentes dans les *Instructions nautiques* ? Utilisez vous d'autres sources de données ?
15. Décrivez l'outil idéal.

Déroulé de l'entretien auprès des élèves

Partie présentation

1. Pouvez-vous vous présenter brièvement ?
2. En quelle année de formation êtes-vous ?
3. Depuis combien d'années manipulez-vous les *Instructions nautiques* ? Dans quel cadre ? (Cela peut remonter à avant la formation.)

Pratiques dans le cadre des études

4. Dans le cadre de vos études, comment utilisez-vous les *Instructions nautiques* ? Pour répondre à cette question, pourriez-vous détailler un scénario d'usage des cartes et des *Instructions nautiques* ou deux scénarios d'usage si vous estimez que c'est bien différent lors de la formation théorique et en mer ?
5. Dans la pratique, vous utilisez un « navigateur électronique ». Quels sont les besoins supplémentaires nécessitant l'usage des *Instructions nautiques* ?
6. Rencontrez-vous des difficultés récurrentes, ou plus occasionnelles, à la compréhension des *Instructions nautiques*, ou à leurs usages ?
7. Pour vous, y'a-t-il des manques dans les *Instructions nautiques*, que ce soit dans vos cours théoriques et/ou pratiques ?

8. Lors de vos sorties, vous utilisez des outils numériques (navigateurs, etc.). Quelles sont les données qui vous manquent et qui sont présentes dans les *Instructions nautiques* ? Utilisez vous d'autres sources de données ?
9. Décrivez l'outil idéal.

Appendix B

Ontology Subdomain Documentation: Maritime Navigation Guidelines

This appendix contains all the documentation produced for the Maritime Navigation Guidelines subdomain.

B.1 Motivating Scenario

Name

Maritime Navigation Guidelines

Description

The *Instructions nautiques* contain many maritime navigation guidelines. Maritime navigation guidelines are pieces of information, instructions or prohibitions that concern all possible actions in the maritime domain. Such actions are most commonly navigating or remaining stationary on the water. Maritime navigation guidelines can also come in the form of contact information. An instruction can be advisory or obligatory, in which case a decree is cited. A prohibition is necessarily obligatory and cites a decree. A piece of information can be linked to a decree. Maritime navigation guidelines can be dependent on local conditions such as temporality (time of day, season, etc.) or meteorological and oceanographic conditions. They can also be targeted at maritime vessels of a specific type, such as fishing boat or cruise ship, and/or with specific material characteristics such as size, cargo or origin.

Extracts

a. “La grande passe de l’Ouest (12° 47,90’ S — 44° 58,00’ E) est malsaine et non balisée. Il est déconseillé d’emprunter cette passe.” (Shom 2021g, p. 231)

b. “INSTRUCTIONS. — De jour, la route d’approche de l’entrée de la passe est orientée à environ 114° vers l’extrémité Sud du sommet du mont Mahinia, ou vers le versant Nord du mont de la Selle (§ 5.5.2.). Dès que les balises sont identifiées, suivre l’alignement (114,5°) indiqué par la carte.” (Shom 2021g, p. 309)

c. “INSTRUCTIONS. — En venant de l’Est, on prend le chenal en suivant l’alignement à 293,3° du clocher de l’Île de Batz (chapelle Notre-Dame de Bon Secours) [48° 44,65’ N — 4° 00,58’ W], sur la côte Sud de l’île, par la pyramide blanche de l’Île Pigued (48° 43,98’ N — 3° 58,22’ W). Cet alignement n’est visible par les petits navires que jusqu’à environ 0,6 M à l’Est de la tourelle « Le Menk » (à mi-marée) et, par les navires à passerelle plus haute, jusqu’au Nord de la tourelle. Cet alignement se situe dans le secteur blanc (289,5° – 293°) du feu de la tourelle « Ar Chaden ». La route à 293,3° laisse au Nord le plateau des Duons et au Sud la tourelle « Le Menk » (48° 43,29’ N — 3° 56,70’ W), cardinale Ouest lumineuse, et la Basse de Bloscon.” (Shom 2021a, p. 399)

Main Concepts and Characteristics

A maritime navigation guideline must be of one of the following types:

— information

— instruction

— prohibition

— contact information

— decree

Maritime navigation guidelines typically have the following optional or obligatory characteristics:

— type of guideline [obligatory]

— region to which guideline applies [obligatory: information|prohibition, optional: instruction]

— spatial entity to be followed according to guideline [optional: instruction]

— action to be carried out according to guideline [obligatory: instruction]

— action prohibited by guideline [obligatory: prohibition]

— decree at origin of guideline [optional: information|instruction|prohibition]

— name of decree [obligatory: decree]

— local condition under which guideline is valid [optional: information|instruction|prohibition|contact information]

— target of guideline [optional: information|instruction|prohibition|contact information]

— exception to guideline [optional: information|instruction|prohibition]

— complementary information [optional]

B.2 Informal Competency Questions

1. What are the navigation instructions for The Great Western Pass? (Extract a.)
   — It is not recommended to take The Great Western Pass.
2. How can the North Channel be accessed? (Extract b.)
   — During the day, the channel entry access route is oriented at approximately 114° towards the southern extremity of the summit of Mont Mahinia, or towards the northern slope of Mont de la Selle.

B.3 Glossary

Term	Definition
Action	Any possible action that can be executed in the maritime domain. The most common are 'navigating' and 'remaining stationary' (on the water). Other examples include 'dragging', 'fishing' and 'swimming'.
Contact information	A type of maritime navigation guideline that gives one or more ways of contacting a service that may need to be reached before or during navigation.
Information (piece of)	A type of maritime navigation guideline that is purely informative and can be linked to a decree.

Continued on next page

Term	Definition
Instruction	A type of maritime navigation guideline that indicates how an activity is to be performed. It can be advisory or obligatory, in which case a decree is cited.
Local condition	*See glossary for Temporalities, Meteorological and Oceanographic Phenomena subdomain*
Maritime domain	*See glossary for Maritime Spatial Entities and Spatial Relations subdomain*
Maritime navigation guideline	A piece of information, an instruction or a prohibition that concerns any possible action that can be executed in the maritime domain.
Maritime navigation guideline type	A category of maritime navigation guideline. There are five categories of maritime navigation guideline: information, instruction, prohibition, contact information, decree.
Maritime vessel	*See glossary for Maritime Vessels subdomain*
Material characteristic	*See glossary for Maritime Vessels subdomain*
Navigating	An action that involves the deliberate movement of a vessel on the water.
Prohibition	A type of maritime navigation guideline that indicates an action that may not performed. It is necessarily obligatory and cites a decree.
Remaining stationary	An action that involves avoiding the movement of a vessel at a given position on the water by mooring or by dropping the anchor. Places where the action of remaining stationary can be executed are called 'stopping places'.
Target	The type of maritime vessel for which a maritime navigation guideline applies.

Appendix C

Ontology Subdomain Documentation: Maritime Spatial Entities and Spatial Relations

This appendix contains all the documentation produced for the Maritime Spatial Entities and Spatial Relations subdomain.

C.1 Motivating Scenario

Name

Maritime Spatial Entities and Spatial Relations

Description

The *Instructions nautiques* contain references to many spatial entities and spatial relations in the maritime environment. Spatial entities and spatial relations are cited in the maritime navigation guidelines, in the descriptions of the maritime environment and in the descriptions of meteorological and oceanographic phenomena given in the *Instructions nautiques*. Spatial entities are macroscopic things located in space. The location of a spatial entity can be expressed via a spatial reference, which can either be direct (geographic coordinates) or indirect (name). When referenced in natural language text, a spatial entity can either be named or unnamed. Spatial entities must necessarily be defined at least either their name or their type. Spatial entities can be physical (lighthouse) or virtual (leading line) whilst still occupying a location in space. Their nature can be natural (sandbank), artificial (beacon) or administrative (country). They can be located in the terrestrial domain (house), the maritime domain (buoy), either (rock) or both (foreshore). The position of a spatial entity can be described by geographic coordinates or by its spatial relations with other

entities. A spatial entity can be described by its visual characteristics such as its height or colour.

Extracts

a. “La pointe Barn Hill (23° 33,3’ S — 43° 44,6’ E) est l’extrémité d’une étroite péninsule escarpée qui s’avance à 1 M au SSW de Taliokoaka, promontoire haut de 60 m. Cette péninsule, connue sous le nom de Ny Andrea (Nosy Andrea), est bordée de falaises calcaires de couleur blanche, très apparentes lorsqu’elles sont éclairées par le soleil.” (Shom 2021g, p. 309)

b. “La grande passe de l’Ouest (12° 47,90’ S — 44° 58,00’ E) est malsaine et non balisée. Il est déconseillé d’emprunter cette passe.” (Shom 2021g, p. 231)

c. “INSTRUCTIONS. — De jour, la route d’approche de l’entrée de la passe est orientée à environ 114° vers l’extrémité Sud du sommet du mont Mahinia, ou vers le versant Nord du mont de la Selle (§ 5.5.2.). Dès que les balises sont identifiées, suivre l’alignement (114,5°) indiqué par la carte.” (Shom 2021g, p. 309)

d. “INSTRUCTIONS. — En venant de l’Est, on prend le chenal en suivant l’alignement à 293,3° du clocher de l’Île de Batz (chapelle Notre-Dame de Bon Secours) [48° 44,65’ N — 4° 00,58’ W], sur la côte Sud de l’île, par la pyramide blanche de l’Île Pigued (48° 43,98’ N — 3° 58,22’ W). Cet alignement n’est visible par les petits navires que jusqu’à environ 0,6 M à l’Est de la tourelle « Le Menk » (à mi-marée) et, par les navires à passerelle plus haute, jusqu’au Nord de la tourelle. Cet alignement se situe dans le secteur blanc (289,5° – 293°) du feu de la tourelle « Ar Chaden ». La route à 293,3° laisse au Nord le plateau des Duons et au Sud la tourelle « Le Menk » (48° 43,29’ N — 3° 56,70’ W), cardinale Ouest lumineuse, et la Basse de Bloscon.” (Shom 2021a, p. 399)

e. “Vue du Nord, l’Île de Batz montre la tour du sémaphore (48° 44,78’ N — 4° 00,69’ W) et surtout le phare (48° 44,72’ N — 4° 01,61’ W), tour grise haute de 43 m, entourée de maisons.” (Shom 2021a, p. 398)

Main Concepts and Characteristics

Spatial entities typically have the following optional or obligatory characteristics:

— type of entity [obligatory unless name is known]

— name of entity [obligatory unless type is known]

— geographic coordinates of entity [optional]

— spatial relations in which entity is involved [optional]

— visual characteristics of entity [optional]

Spatial relations typically have the following optional or obligatory characteristics:

— type of spatial relation [obligatory]

— spatial entities involved in spatial relation [obligatory]

C.2 Informal Competency Questions

1. What does Barn Hill Point look like from a boat on the water? (Extract a.)

   — Barn Hill Point is the extremity of a narrow craggy peninsula that reaches 1 M SSW of Taliokoaka, a headland 60 m tall. This peninsula is lined with white limestone cliffs that stand out when illuminated by the sun.

2. Is The Great Western Pass marked? (Extract b.)

   — No, The Great Western Pass is unmarked.

3. What is the orientation of the North Channel entry access route? (Extract c.)

   — During the day, the channel entry access route is oriented at approximately 114° towards the southern extremity of the summit of Mont Mahinia, or towards the northern slope of Mont de la Selle.

4. What colour is the pyramid of Île Pigued? (Extract d.)

   — The colour of the pyramid of Île Pigued is white.

5. What landmarks are there on the Île de Batz? (Extracts d. and e.)

   — On the Île de Batz, Île de Batz bell tower, Notre-Dame de Bon Secours chapel, a semaphore, a semaphore tower and a lighthouse serve as landmarks.

C.3 Glossary

Term	Definition
Maritime domain	The space on Earth that is occupied by a sea or an ocean.
Maritime navigation guideline	*See glossary for Maritime Navigation Guidelines subdomain*
Meteorological phenomenon	*See glossary for Temporalities, Meteorological and Oceanographic Phenomena subdomain*
Oceanographic phenomenon	*See glossary for Temporalities, Meteorological and Oceanographic Phenomena subdomain*
Spatial entity	A macroscopic thing located in space.
Spatial entity type	A category of spatial entity such as 'island'.
Spatial relation	A description of the relative position of two or more spatial entities, or one or more spatial entities and one or more meteorological or oceanographic phenomena.
Spatial relation type	A category of spatial relation.
Terrestrial domain	The space on Earth that is not occupied by a sea or an ocean.
Visual characteristic	A characteristic that describes a visual quality of a spatial entity in a qualitative or quantitative manner, such as its colour or its size.

Appendix D

Ontology Subdomain Documentation: Temporalities, Meteorological and Oceanographic Phenomena

This appendix contains all the documentation produced for the Temporalities, Meteorological and Oceanographic Phenomena subdomain.

D.1 Motivating Scenario

Name

Temporalities, Meteorological and Oceanographic Phenomena

Description

The *Instructions nautiques* contain references to many temporalities as well as meteorological and oceanographic phenomena. They give indications of the typical local conditions or of the local conditions under which given guidelines are valid. Temporalities refer to time-related local conditions such as the time of day, the month of the year or the season, which can affect typical meteorological and oceanographic conditions and therefore light levels and opening hours. Temporalities are cited in the maritime navigation guidelines, the descriptions of spatial entities and spatial relations, and in the descriptions of meteorological and oceanographic phenomena given in the *Instructions nautiques*. Meteorological and oceanographic phenomena are cited in the navigation guidelines, and in the descriptions of spatial entities and spatial relations. Meteorological and oceanographic phenomena must necessarily be associated to their type and they can be associated to a temporality. The position of a meteorological or oceanographic phenomena can be described by geographic coordinates or by its

spatial relations with spatial entities. A meteorological or oceanographic phenomenon can be described by its physical characteristics such as its direction or its intensity.

Extracts

a. “Le climat est froid, humide et très venteux. Sur les plaines côtières, la neige peut tomber à toute époque de l’année mais subsiste rarement plus de quelques jours.” (Shom 2021g, p. 458)

Main Concepts and Characteristics

Temporalities typically have the following optional or obligatory characteristics:

— type of temporality [obligatory]

— guideline to which temporality applies [optional]

— meteorological or oceanographic phenomenon to which temporality applies [optional]

Meteorological and oceanographic phenomena typically have the following optional or obligatory characteristics:

— type of phenomenon [obligatory]

— guideline to which phenomenon applies [optional]

— temporality during which phenomenon applies [optional]

— physical characteristic of phenomenon [optional]

D.2 Informal Competency Questions

1. What is the climate like on the Kerguelen Islands? (Extract a.)
   — The climate is cold, humid and very windy.
2. Does it snow on the Kergeulen Islands? (Extract a.)
   — On the coastal plains, snow can fall at any time of year but rarely lasts more than a few days.

D.3 Glossary

Term	Definition
Local condition	A temporal, meteorological or oceanographic condition that holds locally.
Maritime navigation guideline	*See glossary for Maritime Navigation Guidelines subdomain*
Meteorological phenomenon	A natural phenomenon that takes place in the Earth's atmosphere.
Meteorological phenomenon type	A category of meteorological phenomenon such as 'wind'.
Oceanographic phenomenon	A natural phenomenon that takes place in the sea or the ocean.
Oceanographic phenomenon type	A category of oceanographic phenomenon such as 'current'.
Spatial entity	*See glossary for Maritime Spatial Entities and Spatial Relations subdomain*
Spatial relation	*See glossary for Maritime Spatial Entities and Spatial Relations subdomain*
Temporality	A time-related local condition such as the time of day, the month of the year or the season.
Physical characteristic	A characteristic that describes a physical quality of a meteorological or oceanographic phenomenon in a qualitative or quantitative manner such as its intensity or its direction.

Appendix E

Ontology Subdomain Documentation: Maritime Vessels

This appendix contains all the documentation produced for the Maritime Vessels subdomain.

E.1 Motivating Scenario

Name

Maritime Vessels

Description

The *Instructions nautiques* contain references to many maritime vessels. Maritime vessels are cited in the maritime navigation guidelines because some guidelines are targeted at maritime vessels of a specific type, such as fishing boat or cruise ship, and/or with specific material characteristics such as size, cargo or origin.

Extracts

a. "INSTRUCTIONS. — En venant de l'Est, on prend le chenal en suivant l'alignement à 293,3° du clocher de l'Île de Batz (chapelle Notre-Dame de Bon Secours) [48° 44,65' N — 4° 00,58' W], sur la côte Sud de l'île, par la pyramide blanche de l'Île Pigued (48° 43,98' N — 3° 58,22' W). Cet alignement n'est visible par les petits navires que jusqu'à environ 0,6 M à l'Est de la tourelle « Le Menk » (à mi-marée) et, par les navires à passerelle plus haute, jusqu'au Nord de la tourelle. Cet alignement se situe dans le secteur blanc (289,5° – 293°) du feu de la tourelle « Ar Chaden ». La route à 293,3° laisse au Nord le plateau des Duons et au Sud la tourelle « Le Menk » (48° 43,29' N — 3° 56,70'

W), cardinale Ouest lumineuse, et la Basse de Bloscon." (Shom 2021a, p. 399)

Main Concepts and Characteristics

Maritime vessels typically have the following optional characteristics:

— type of vessel [optional]

— material characteristics of vessel [optional]

E.2 Informal Competency Questions

1. Is the alignment of Île de Batz clock tower and the pyramid of Île Pigued visible to all vessels? (Extract a.)
   — The alignment of Île de Batz clock tower and the pyramid of Île Pigued is visible to small vessels up to around 0.6 M to the east of 'Le Menk' turret (at half tide) and, for vessels with higher bridges, up to the north of the turret.

E.3 Glossary

Term	Definition
Maritime navigation guidelines	*See glossary for Maritime Navigation Guidelines subdomain*
Maritime domain	*See glossary for Maritime Spatial Entities and Spatial Relations subdomain*
Maritime vessel	A means of transport that can be used to navigate in the maritime domain.
Maritime vessel type	A category of maritime vessel such as 'fishing boat', 'cruise ship' or 'submarine'.
Material characteristic	A characteristic that describes a material quality of a maritime vessel in a qualitative or quantitative manner such as its cargo or its length.

Résumé détaillé de la thèse en français

Création de graphes de connaissances géospatiaux à partir de sources hétérogènes

Mots clés : traitement automatique du langage naturel, ontologie, apprentissage profond, données géospatiales

Introduction

Certaines connaissances spatiales, actuelles ou historiques, n'existent que sous forme de texte. Les guides de voyage, les documents historiques et les publications sur les réseaux sociaux sont quelques exemples de sources de connaissances spatiales non structurées. Les sources textuelles contiennent des connaissances spatiales naturellement hétérogènes : elles peuvent être écrites par différents auteurs, en utilisant un vocabulaire différent, à partir d'un point de vue différent. Elles peuvent par ailleurs couvrir des zones géographiques larges et diverses et contenir des niveaux de détail variés (Ezeani et al. 2023 ; Jiménez–Badillo et al. 2020 ; Y. Hu et al. 2019 ; Kim et al. 2015 ; Beall 2010). Pour toutes ces raisons il est difficile d'intégrer dans les modèles de systèmes d'information géographique (SIG) l'information géographique provenant de sources textuelles. L'hypothèse du monde ouvert des technologies du Web sémantique induit que les graphes de connaissances sont une meilleure solution pour modéliser et stocker les connaissances géographiques extraites de textes hétérogènes, incomplets et imparfaits en langage naturel (Janowicz et al. 2022 ; H. Chen et al. 2018 ; Melo et Martins 2017 ; Stadler et al. 2012). Structurées en graphe de connaissances géospatial, les connaissances spatiales ambiguës peuvent être désambiguïsées et liées formellement à des ressources

géographiques de référence (telles que DBpedia [1] ou BD TOPO® [2]), ce qui les enrichit de références spatiales directes lorsque c'est possible et facilite considérablement leur accessibilité et réutilisation (Janowicz et al. 2022 ; Melo et Martins 2017).

Objectif et verrous

L'objectif de cette thèse est de développer une approche opérationnelle pour la construction de graphes de connaissances à partir de texte et des données géographiques de référence. Cette approche doit permettre d'intégrer à la fois des références spatiales directes et indirectes.

Afin d'atteindre cet objectif, il faut résoudre trois verrous scientifiques principaux :

1. Comment peut-on se doter d'une ontologie de domaine adaptée au corpus de texte qui structurera le graphe de connaissances géospatial ?
2. Quelles techniques pour l'extraction d'information spatiale peut-on appliquer aux sources de texte hétérogènes et faut-il les raffiner pour cette utilisation ?
3. Comment peut-on structurer l'information spatiale extraite automatiquement selon l'ontologie de domaine afin de la peupler et ainsi construire un graphe de connaissances, et comment peut-on désambiguïser les entités spatiales et les lier à une ressource géographique de référence ?

Contexte d'application

Nous appliquons nos recherches à un corpus de texte en langue française en collaboration avec le Service hydrographique et océanographique de la Marine (Shom). Ceci nous permet d'identifier et de valider empiriquement une méthodologie fonctionnelle pour la construction de graphes de connaissances géospatiales à partir de texte. Le corpus est constitué des *Instructions nautiques*, une série d'ouvrages publiés par le Shom qui décrivent l'environnement maritime côtier et donnent des instructions de navigation côtière.

Les *Instructions nautiques* font partie d'une gamme de produits diffusés par le Shom qui servent à la planification d'itinéraires de navigation maritime. D'autres ouvrages du Shom, plus spécialisés, viennent compléter les

1. DBpedia (https://www.dbpedia.org/) est une ressource géographique de référence mondiale.
2. La BD TOPO® (https://geoservices.ign.fr/bdtopo) est une ressource géographique de référence pour le territoire français.

connaissances sur l'environnement côtier et la navigation, parmi lesquels on trouve *Feux et signaux de brume*, *Radiosignaux*, *Courants de marée* ainsi que l'*Annuaire des marées*. Ils apportent des renseignements qui sont nécessaires à la préparation d'un itinéraire adapté et sûr. Le type de navire, l'expérience du navigateur, la temporalité [3], les conditions météorologiques et les conditions océanographiques sont également à prendre en considération lors de la planification. Les *Instructions nautiques* contiennent principalement trois types de renseignements (Shom 2020) :

1. elles donnent des informations complémentaires à celles qui sont affichées sur les cartes marines comme les caractéristiques physiques (couleur, forme, taille, etc.) d'un amer [4],
2. elles recensent les informations absentes des cartes marines telles que le climat typique de la zone décrite,
3. elles donnent des instructions ou des informations à propos de la navigation telles que les routes conseillées, les conditions d'accès aux ports ou encore les réglementations en place.

Les *Instructions nautiques* sont divisées en plusieurs volumes, un par zone de couverture. Une zone de couverture peut être définie soit comme une section de trait de côte entre deux positions sur la côte, soit comme l'ensemble du trait de côte d'une île ou d'un ensemble d'îles. Chaque volume commence avec un chapitre de renseignements généraux. Le plan général du reste de l'ouvrage suit linéairement le trait de côte, chaque chapitre étant dédié à une section du trait de côte. En lisant un chapitre, le lecteur a l'impression d'être emmené le long de la côte par le rédacteur ; chaque repère, danger et autre particularité de l'environnement est décrit, et chaque mouillage, accès de port et entrée de chenal est détaillé. Les consignes mentionnent également les spécificités de la météorologie, la courantologie et la réglementation locales. Des photographies montrant les amers et les ports notables sont intercalées dans le texte. Elles illustrent également le positionnement relatif des différentes entités géographiques et doivent conforter le lecteur dans la représentation qu'il se fait de son environnement.

Contributions

Dans cette section nous détaillons les contributions de cette thèse. Nous présentons d'abord la principale contribution, qui est méthodologique, et

3. Une temporalité est une condition locale qui est dépendante sur le temps, par exemple l'heure, le mois de l'année ou le saison.

4. Un amer est un « objet remarquable situé à un endroit fixe sur la terre et pouvant être utilisé pour déterminer un emplacement ou une direction. » Traduit de Hydrographic Dictionary Working Group (2019).

ensuite les contributions ressources.

La méthodologie ATONTE

La contribution principale de cette thèse est la méthodologie *ATlantis Ontology and kNowledge graph development from Texts and Experts* (ATONTE) pour la construction semi-automatique de graphes de connaissances, géospatiaux ou non, à partir de sources textuelles hétérogènes, des connaissances d'experts et des données de référence.

La première composante de la méthodologie ATONTE est une nouvelle méthodologie pour le développement d'ontologies de domaine à partir de texte et d'experts (Rawsthorne et al. 2022b). Elle peut être utilisée pour créer des ontologies *géospatiales* si nécessaire.

La deuxième composante est une approche pour l'extraction automatique d'entités imbriquées et de relations binaires de texte, qui peut être appliquée aux entités et aux relations *spatiales* si nécessaire (Rawsthorne et al. 2024b ; Rawsthorne et al. 2023). Notre approche est une adaptation du système *Princeton University Relation Extraction* (PURE) existant (Zhong et D. Chen 2021), qui a été conçu pour l'extraction d'entités génériques plates (non imbriquées) et de relations binaires génériques.

La troisième et dernière composante est dédiée à la structuration en triplets *Resource Description Framework* (RDF) de l'information extraite du texte afin de construire un graphe de connaissances, et la désambiguïsation des entités qu'elle contient [5]. Le graphe de connaissances est *géospatial* s'il contient des entités *spatiales*. Un schéma illustrant la méthodologie ATONTE est présenté en figure 1.

L'ontologie ATLANTIS

L'ontologie *coAsTaL mAritime NavigaTion InstructionS* (ATLANTIS) [6] est une ontologie noyau géospatiale et multilingue qui couvre le domaine des *Instructions nautiques* (Rawsthorne et al. 2022a). Elle a été publié sous la Licence Ouverte Version 2.0 Etalab.

Le jeu de données ATLANTIS avec résultats de référence

Le jeu de données ATLANTIS [7] est un jeu de données de référence en langue française, annoté manuellement, qui porte sur le domaine maritime (Rawsthorne et al. 2024b ; Rawsthorne et al. 2023). Il a été publié

5. https://github.com/umrlastig/atonte-structure-and-disambiguate
6. https://github.com/umrlastig/atlantis-ontology
7. https://github.com/umrlastig/atlantis-dataset

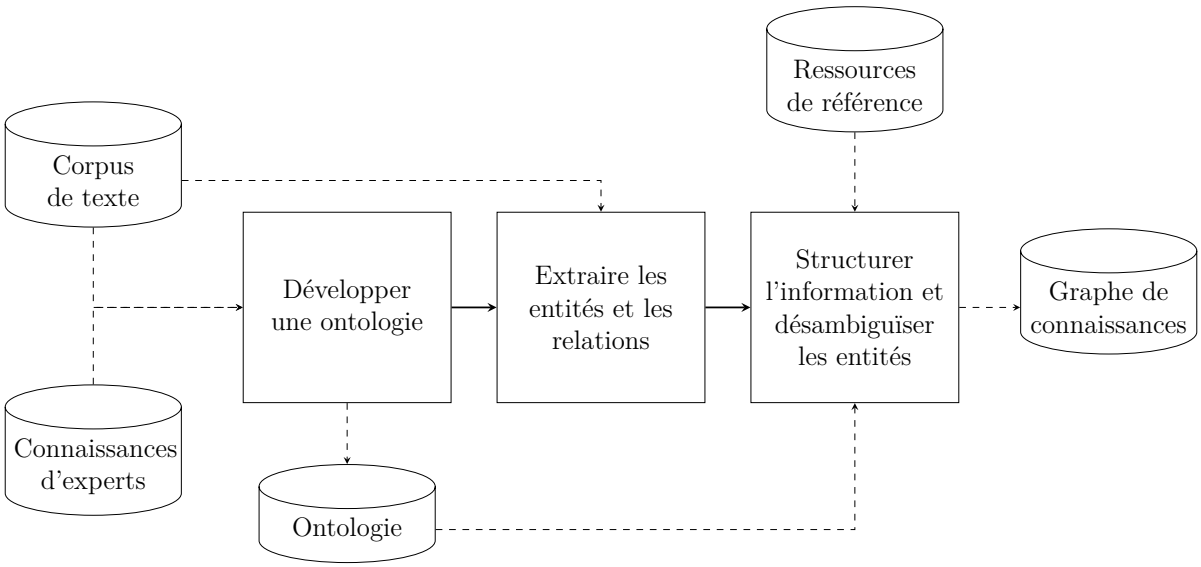


Figure 1 – Schéma de l'approche globale adoptée dans cette thèse. Les processus principaux sont représentés par des carrés tandis que les ressources entrantes et sortantes sont représentées par des cylindres. Les flèches avec une ligne continue représentent le flux des tâches et les flèches avec une ligne discontinue en pointillés représentent les flux d'informations ou de connaissances.

sous la Licence Ouverte Version 2.0 Etalab. Il peut être utilisé pour entraîner des algorithmes pour l'extraction d'entités spatiales imbriquées et de relations spatiales binaires à partir de texte. Nous proposons également des résultats de référence pour ce jeu de données, pour les tâches d'extraction d'entités spatiales imbriquées, d'extraction de relations spatiales binaires, et d'extraction combinée d'entités et de relations spatiales de bout en bout.

Le jeu de données TextMine'24

Le jeu de données TextMine'24 est un sous-ensemble du jeu de données ATLANTIS et contient des annotations d'entités spatiales imbriquées (Rawsthorne et al. 2024a). Il a été utilisé en tant que jeu de données de référence pour le Défi TextMine 2024 [8], un défi de reconnaissance d'entités spatiales organisé par le groupe de travail TextMine [9] qui fait partie de l'Association Internationale Francophone d'Extraction et de Gestion des Connaissances (EGC) [10].

La méthodologie ATONTE : pour la création de graphes de connaissances à partir de texte et d'experts

Dans cette section nous présentons le cœur du travail de cette thèse : la méthodologie ATONTE, ainsi que notre implémentation de la méthodolo-

8. https://www.kaggle.com/competitions/defi-textmine-2024
9. https://textmine.sciencesconf.org/
10. https://www.egc.asso.fr/

gie sur notre corpus d'*Instructions nautiques*.

Le développement d'ontologies à partir de texte et d'experts

La première des trois composantes qui constituent ATONTE est une nouvelle méthodologie pour le développement manuel d'ontologies de domaine à partir de corpus de texte et des connaissances d'experts du domaine. Elle est composée de quatre étapes principales, dont les deux dernières se réalisent de manière itérative, ainsi qu'une étape préliminaire qui doit être effectuée avant de commencer l'implémentation de la méthodologie. Cette étape préliminaire sert à vérifier que cette méthodologie est le bon choix de méthodologie de développement d'ontologie pour un projet donné. Voici un aperçu des étapes de la méthodologie et des tâches à réaliser à chaque étape :

0. Étude de faisabilité
   - Vérification que la méthodologie est adaptée
1. Travail de fondation
   - Familiarisation avec le corpus, identification et analyse de sources de connaissances du domaine, identification d'experts du domaine, définition de l'application de l'ontologie, création d'un jeu de données préliminaire composé de triplets sémantiques, division du domaine en sous-domaines
2. Production de documentation
   - Rédaction d'un argumentaire qui donne une description en langage naturel du sous-domaine ainsi qu'un ou plusieurs exemples illustratifs, d'une liste de questions informelles de compétence et d'un glossaire
3. Structurer, implémenter et tester les modèles de sous-domaine
   - Conceptualiser les modèles de sous-domaine, implémenter les modèles de sous-domaine en utilisant *Web Ontology Language* (OWL), créer des jeux de données RDF pour chaque sous-domaine, faire évoluer les modèles et les jeux de données de chaque sous-domaine de façon itérative, tester les modèles et les jeux de données
4. Fusionner, remanier et aligner
   - Fusionner les modèles de sous-domaine afin de créer le modèle complet, fusionner les jeux de données de chaque sous-domaine afin de créer le jeu de données exemple pour le modèle complet, remanier et aligner le modèle complet, tester le modèle complet et le jeu de données exemple

Nous avons implémenté notre méthodologie de développement d'ontologies sur notre corpus des *Instructions nautiques* avec l'aide de deux types d'experts du domaine : l'équipe de rédacteurs des *Instructions nautiques* et des utilisateurs des ouvrages. Le résultat est l'ontologie ATLANTIS, qui permet de modéliser les entités spatiales présentes dans l'environnement maritime côtier, les relations spatiales entre elles, les consignes de navigation maritime, les navires utilisés pour la navigation côtière, les temporalités, et les phénomènes météorologiques et océanographies.

En 2022 nous avons encadré le stage de Benoit et Kergus (2022), qui ont réalisé une étude pour savoir si l'ontologie ATLANTIS est adaptée à la modélisation des ouvrages équivalents aux *Instructions nautiques* provenant d'autres pays. Pour ce faire, ils ont appliqué la première étape de la méthodologie ATONTE à trois séries d'*Instructions nautiques* écrites en anglais. Leurs résultats montrent que l'ontologie ATLANTIS est adaptée à la modélisation de la plupart de l'information contenue dans ces ouvrages, mais qu'elle doit être légèrement modifiée afin d'être intégralement adaptée. Ces travaux ont montré que notre méthodologie de développement d'ontologies est répétable et qu'elle peut être appliquée à l'enrichissement manuel d'ontologies existantes.

L'extraction d'entités et de relations de texte

La deuxième des trois composantes qui constituent ATONTE est une approche automatique pour l'extraction d'entités imbriquées et de relations binaires à partir de texte en utilisant un réseau de neurones profond, basée sur PURE (Zhong et D. Chen 2021). Il implique l'entraînement de deux modèles de langage profonds pré-entraînés existants : un pour la tâche d'extraction d'entités et l'autre pour l'extraction de relations. Les modèles sont entraînés sur un jeu de données annoté manuellement, spécifique au domaine. Ce jeu de données contient notamment des annotations d'entités imbriquées.

Nous avons implémenté cette approche afin d'extraire les entités spatiales et les relations spatiales de notre corpus, ce qui a exigé la création d'un jeu de données d'entraînement en français, annoté à la main. Nous donnons des résultats de référence pour ce jeu de données pour trois tâches : l'extraction d'entités spatiales imbriquées, l'extraction de relations spatiales binaires, et l'extraction combinée d'entités et de relations spatiales de bout en bout.

La structuration d'information et la désambiguïsation d'entités

La troisième et dernière composante qui constitue ATONTE est dédiée à la structuration de l'information extraite lors de l'étape précédente sous la forme d'un graphe de connaissances, et la désambiguïsation des entités en les liant à une ressource de référence.

Nous présentons une preuve de concept, en utilisant des outils disponibles afin de structurer les entités et relations spatiales extraites des *Instructions nautiques* selon l'ontologie ATLANTIS dans un premier temps, et de lier les entités à leurs entrées correspondantes dans la BD TOPO® [11] dans un second temps. Pour cette deuxième tâche, nous nous sommes basés sur le travail réalisé par Loynes et Ruiz (2020). Nous l'avons amélioré notamment en ajoutant une comparaison du type de l'entité avec le type de l'entrée dans la ressource de référence, au lieu d'utiliser uniquement une comparaison de leurs noms respectifs. Le résultat est une base opérationnelle du graphe de connaissances géospatial des *Instructions nautiques*.

En 2022 nous avons encadré le projet de développement d'Alla et al. (2022) dans le cadre de leurs études de Master. Ils ont développé un prototype d'une plateforme Web nommée Nereus [12] qui permet d'accéder au contenu du graphe de connaissances ATLANTIS [13] via une interface graphique basée sur une carte marine, sans avoir à écrire des requêtes SPARQL. Ce prototype fonctionnel d'une plateforme Web qui permet d'accéder visuellement au contenu d'un graphe de connaissances géospatial démontre un des avantages de la structuration du contenu des *Instructions nautiques* pour ses utilisateurs.

Conclusion

Nous avons développé la méthodologie ATONTE de manière empirique. Elle est le résultat de notre volonté de construire un graphe de connaissances géospatial du contenu des *Instructions nautiques* (que nous présentons dans le chapitre 2) avec l'aide d'experts du domaine et des ressources géographiques de référence. La méthodologie ATONTE, dont nous présentons un aperçu dans le chapitre 3, est composée de trois étapes :

1. Le développement d'ontologies à partir de texte et d'experts, présenté dans le chapitre 4

11. La BD TOPO® est une base de données qui couvre l'ensemble des entités géographiques et administratives du territoire national français.
12. https://github.com/mcharzat/SHOM_IN
13. Ils ont réalisé leur travail en utilisant un échantillon teste du graphe de connaissances ATLANTIS, construit en peuplant manuellement l'ontologie ATLANTIS avec des triplets RDF géospatiales.

2. L'extraction d'entités et de relations à partir de texte, présentée dans le chapitre 5
3. La structuration d'information et la désambiguïsation d'entités, présentées dans le chapitre 6

Dans chacun de ces trois chapitres, nous présentons également l'implémentation de l'étape correspondante sur notre corpus d'*Instructions nautiques*. Le résultat de l'application de l'ensemble de cette méthodologie sur les *Instructions nautiques* est l'ontologie ATLANTIS ainsi qu'une base opérationnelle pour le graphe de connaissances géospatial ATLANTIS. Nous avons donc montré que la méthodologie ATONTE est opérationnelle pour la tâche de création d'un graphe de connaissances à partir de texte et d'experts.

Le Shom a recruté un scientifique des données pour un contrat à durée indéterminée afin de continuer le travail réalisé pendant cette thèse. Le rôle de cette personne est de concevoir un outil de peuplement de graphe de connaissances plus complet, le tester sur une zone d'étude suffisamment étendue et évaluer la qualité du résultat dans le but de valider l'ontologie ATLANTIS et de manière plus large la méthodologie ATONTE.

Perspectives

Concernant la méthodologie ATONTE, une première perspective est de l'appliquer à d'autres corpus de texte qui couvrent d'autres domaines afin d'évaluer l'étendue de son domaine d'application. De plus, il faudrait ajouter des évaluations à chaque étape de la méthodologie qui permettront de valider au fur et à mesure de son implémentation que le modèle construit et les données produites pour le peupler sont cohérents et adaptés à leur contexte d'application.

Concernant la deuxième composante en particulier, une des perspectives concerne la réduction du temps nécessaire pour annoter le jeu de données d'entraînement. Pour ce faire, on pourrait utiliser un outil d'annotation qui propose soit un étiquetage assisté par apprentissage machine, soit une annotation automatisée en fonction d'un ensemble de mots clés défini par l'utilisateur. Il faudrait évaluer et comparer les annotations produites à l'aide des différents outils afin de vérifier que la qualité ne diminue pas par rapport aux annotations que nous avons réalisées de manière entièrement manuelle. Dans le but d'augmenter la fiabilité du jeu de données annoté, on pourrait adopter une approche d'annotation à plusieurs et calculer l'accord inter-annotateurs. On pourrait aussi produire automatiquement des données synthétiques sous la forme d'un texte annoté par le biais d'expres-

sions régulières. Il serait également possible de combiner un apprentissage non supervisé avec un apprentissage supervisé en utilisant un plus petit jeu de données annoté manuellement lors de l'étape d'extraction. Si les résultats sont identiques ou meilleurs que ceux obtenus avec notre approche actuelle, l'apprentissage non supervisé pourrait être intégré à notre méthodologie et le volume de données annotées manuellement nécessaire pourrait être réduit. Une autre perspective concerne l'élargissement de l'approche d'extraction d'information à partir de texte afin de permettre la prise en compte de plusieurs types d'entités simultanément et la gestion de relations n-aires (ternaires ou plus). Ceci permettrait d'extraire des relations impliquant plus de deux entités telles que la relation spatiale *entre*, comme par exemple dans la phrase « la bouée est entre l'île Est et l'île Ouest ».

La troisième composante de la méthodologie ATONTE pourrait être améliorée en tirant profit des connaissances contenues dans les triplets qui définissent une relation spatiale entre entités afin d'aider le processus de désambiguïsation. Le calcul de la possible localisation d'une entité selon sa position relative à d'autres entités pourrait être intégré à l'évaluation des entités candidates. Ceci influencerait le classement des entités et pourrait être réglé afin d'améliorer le score d'entités candidates dont la position géographique est proche de la possible localisation calculée. Enfin, il serait souhaitable d'intégrer à la méthodologie ATONTE la désambiguïsation d'entités non-nommées. Pour ce faire, l'exploitation du type de l'entité, en complément de l'analyse de ses relations spatiales avec d'autres entités, semble être une piste prometteuse comme nous l'avons démontré dans le cas des entités nommées dans le chapitre 6.

References

Agrawal, Ankit, Sarsij Tripathi, Manu Vardhan, Vikas Sihag, Gaurav Choudhary, and Nicola Dragoni (2022). "BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling". In: *Applied Sciences* 12.3, p. 976. DOI: 10.3390/app12030976.

Alla, Claire-Marie, Maxime Charzat, Théo Hermann, Lucie Jeannest, and Paul Miancien (2022). "Développement d'une plateforme pour l'interrogation d'une base de connaissances géoréférencées sur l'environnement maritime côtier". Rapport de projet de 3ème année de cycle ingénieur. Champs-sur-Marne, France: École nationale des sciences géographiques. 25 pp.

Barbe, Arnaud, Molka Tounsi Dhouib, Catherine Faron, Marco Corneli, and Arnaud Zucker (2023). "Construction d'un graphe de connaissance à partir des annotations manuelles de textes de zoologie antique". In: *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA 2022)*. Ingénierie des Connaissances. Strasbourg, France, pp. 15–20.

Beall, Jeffrey (2010). "Geographical research and the problem of variant place names in digitized books and other full-text resources". In: *Library Collections, Acquisitions, & Technical Services* 34.2, pp. 74–82. ISSN: 1464-9055. DOI: 10.1080/14649055.2010.10766263.

Benoit, Jean-Guillaume and Corentin Kergus (2022). "Base de connaissances générée automatiquement à partir de données textuelles". Projet de fin d'études Sciences de l'information et de la modélisation. Brest: École navale. 59 pp.

Berragan, Cillian, Alex Singleton, Alessia Calafiore, and Jeremy Morley (2023). "Transformer based named entity recognition for place name extraction from unstructured text". In: *International Journal of Geographical Information Science* 37.4, pp. 747–766. DOI: 10.1080/13658816.2022.2133125.

Biemann, Chris (2005). "Ontology Learning from Text: A Survey of Methods". In: *LDV-Forum* 20.2, pp. 75–93.

Black, Paul E. (1999). *Algorithms and Theory of Computation Handbook, CRC Press LLC, 1999, "decidable language"*. In: *Dictionary of Algorithms and Data Structures*. 9 August 2004. URL: https://www.nist.gov/dads/HTML/decidableLanguage.html.

Blomqvist, Eva, Karl Hammar, and Valentina Presutti (2016). "Engineering Ontologies with Patterns – The eXtreme Design Methodology". In: *Ontology Engineering with Ontology Design Patterns*. IOS Press, pp. 23–50. DOI: 10.3233/978-1-61499-676-7-23.

Bonatti, Piero A., Aidan Hogan, Axel Polleres, and Luigi Sauro (2011). "Robust and scalable Linked Data reasoning incorporating provenance and trust annotations". In: *Journal of Web Semantics*. Provenance in the Semantic Web 9.2, pp. 165–201. DOI: 10.1016/j.websem.2011.06.003.

Brageul, David and Hans W. Guesgen (2007). "A Model for Qualitative Spatial Reasoning Combining Topology, Orientation and Distance". In: *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference (FLAIRS*

2007). The Florida AI Research Society. Key West, FL, USA: AAAI Press. URL: https://aaai.org/papers/flairs-2007-128/.

Brando, Carmen, Nathalie Abadie, and Francesca Frontini (2016). "Évaluation de la qualité des sources du web de données pour la résolution d'entités nommées". In: *Ingénierie des Systèmes d'Information* 21.5, pp. 31–54. DOI: 10.3166/isi.21.5-6.31-54.

Bucher, Bénédicte, Esa Tiainen, Thomas Ellett, Elise Acheson, Dominique Laurent, and Sylvain Boissel (2019). *Data Linking by Indirect Spatial Referencing Systems*. EuroGeographics Seminar Report. EuroSDR, p. 25. URL: https://hal.science/hal-02536996.

Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*. Frontiers in Artificial Intelligence and Applications. Amsterdam, Netherlands: IOS Press.

Bunel, Mattia (2021). "Modélisation et raisonnement spatial flous pour l'aide à la localisation de victimes en montagne". PhD thesis. Université Gustave Eiffel. 333 pp. URL: https://tel.archives-ouvertes.fr/tel-03298717.

Cadorel, Lucie, Alicia Blanchi, and Andrea G. B. Tettamanzi (2021). "Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text". In: *Proceedings of the 11th on Knowledge Capture Conference*. K-CAP 2021 - International Conference on Knowledge Capture. USA: Association for Computing Machinery, pp. 41–48. DOI: 10.1145/3460210.3493547.

Casati, Roberto and Achille C. Varzi (1997). "Spatial Entities". In: *Spatial and Temporal Reasoning*. Ed. by Oliviero Stock. Dordrecht: Springer Netherlands, pp. 73–96. DOI: 10.1007/978-0-585-28322-7_3.

Čebirić, Šejla, François Goasdoué, Haridimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika (2019). "Summarizing semantic graphs: a survey". In: *The VLDB Journal* 28.3, pp. 295–327. DOI: 10.1007/s00778-018-0528-3.

Ceccarelli, Diego, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani (2013). "Dexter: an open source framework for entity linking". In: *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*. CIKM'13: 22nd ACM International Conference on Information and Knowledge Management. ESAIR '13. San Francisco, CA, USA: Association for Computing Machinery, pp. 17–20. DOI: 10.1145/2513204.2513212.

Chen, Hao, Maria Vasardani, Stephan Winter, and Martin Tomko (2018). "A Graph Database Model for Knowledge Extracted from Place Descriptions". In: *International Journal of Geo-Information* 7.6, p. 221. DOI: 10.3390/ijgi7060221.

Chessa, Alessandro, Gianni Fenu, Enrico Motta, Diego Reforgiato Recupero, Francesco Osborne, Angelo Salatino, and Luca Secchi (2022). "Enriching Data Lakes with Knowledge Graphs". In: *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022)*. First International Workshop on Knowledge Graph Generation From Text. Vol. 3184. Hersonissos, Crete, Greece: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-3184/TEXT2KG_Short_1.pdf.

Cimiano, Philipp (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. New York, NY, USA: Springer. 347 pp. DOI: 10.1007/978-0-387-39252-3.

Clementini, Eliseo and Paolino De Felice (1997). "A Global Framework for Qualitative Shape Description". In: *GeoInformatica* 1.1, pp. 11–27. DOI: 10.1023/A:1009790715467.

Clementini, Eliseo, Paolino Di Felice, and Daniel Hernández (1997). "Qualitative representation of positional information". In: *Artificial Intelligence* 95.2, pp. 317–356. DOI: 10.1016/S0004-3702(97)00046-5.

Conesa, Jordi, Antoni Olive, and Santi Caballé (2011). "Refactoring and its Application to Ontologies". In: *Semantic Web Personalization and Context Awareness: Management of Personal Identities and Social Networking*. Ed. by Miltiadis Lytras, Patricia Ordóñez de Pablos, and Ernesto Damiani. Hershey, PA, USA: IGI Global, pp. 107–136. DOI: 10.4018/978-1-61520-921-7.ch010.

Contarinis, Stelios, Athanasios Palikaris, and Byron Nakos (2020). "The Value of Marine Spatial Open Data Infrastructures—Potentials of IHO S-100 Standard to Become the Universal Marine Data Model". In: *Journal of Marine Science and Engineering* 8, p. 564. DOI: 10.3390/jmse8080564.

d'Aquin, Mathieu (2012). "Modularizing Ontologies". In: *Ontology Engineering in a Networked World*. Ed. by Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi. Berlin, Heidelberg: Springer, pp. 213–233. ISBN: 978-3-642-24794-1.

Dell'Aglio, Daniele, Axel Polleres, Nuno Lopes, and Stefan Bischof (2014). "Querying the web of data with XSPARQL 1.1". In: *Proceedings of the ISWC Developers Workshop 2014, co-located with the 13th International Semantic Web Conference (ISWC 2014)*. Vol. 1268. Riva del Garda, Italy: CEUR Workshop Proceedings, pp. 113–118. URL: https://ceur-ws.org/Vol-1268/paper19.pdf.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020). "RobBERT: a Dutch RoBERTa-based Language Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Findings 2020. Association for Computational Linguistics, pp. 3255–3265. DOI: 10.18653/v1/2020.findings-emnlp.292.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2019. Vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Dey, Rahul and Fathi M. Salem (2017). "Gate-variants of Gated Recurrent Unit (GRU) neural networks". In: *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1597–1600. DOI: 10.1109/MWSCAS.2017.8053243.

Dimou, Anastasia, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle (2014). "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data". In: *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014)*. Vol. 1184. Seoul, Korea: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

DNV (2023). *The Future of Seafarers 2030: A decade of transformation*. Hamburg, Germany, p. 65.

Dokic, Jérôme and Elisabeth Pacherie (2006). "On the very idea of a frame of reference". In: *Space in Languages: Linguistic Systems and Cognitive Categories*. Ed. by Maya Hickmann and Stéphane Robert. Typological Studies in Language. John Benjamins Publishing Company, pp. 259–280. DOI: 10.1075/tsl.66.16dok.

Ehrlinger, Lisa and Wolfram Wöß (2016). "Towards a Definition of Knowledge Graphs". In: *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop*

on Semantic Change & Evolving Semantics (SuCCESS'16). 12th International Conference on Semantic Systems (SEMANTiCS 2016). Vol. 1695. Leipzig, Germany: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-1695/paper4.pdf.

Elliott, Thomas Robert and Sean Gillies (2011). "Pleiades: the un-GIS for Ancient Geography". In: *Digital Humanities 2011: Conference Abstracts*. Poster abstract. Digital Humanities 2011. Stanford, CA, USA: Stanford University Library, pp. 311–313.

Ezeani, Ignatius, Paul Rayson, and Ian Gregory (2023). "Extracting Imprecise Geographical and Temporal References from Journey Narratives". In: *Proceedings of Text2Story — Sixth Workshop on Narrative Extraction From Texts*. Poster and short article. 45th European Conference on Information Retrieval. Vol. 3370. Dublin, Ireland: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-3370/paper11.pdf.

Fan, Runyu, Lizhe Wang, Jining Yan, Weijing Song, Yingqian Zhu, and Xiaodao Chen (2020). "Deep Learning-Based Named Entity Recognition and Knowledge Graph Construction for Geological Hazards". In: *ISPRS International Journal of Geo-Information* 9.1. Publisher: Multidisciplinary Digital Publishing Institute, p. 15. DOI: 10.3390/ijgi9010015.

Fang, Huang (2015). "Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem". In: *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. 5th Annual IEEE International Conference on CYBER Technology in Automation, Control, and Intelligent Systems. Shenyang, China: IEEE, pp. 820–824. DOI: 10.1109/CYBER.2015.7288049.

Feliachi, Abdelfettah, Nathalie Abadie, and Fayçal Hamdi (2014). "Intégration et visualisation de données liées thématiques sur un référentiel géographique". In: *Revue des Nouvelles Technologies de l'Information*. Extraction et Gestion des Connaissances. Vol. RNTI-E-26. Rennes, France: Éditions RNTI, pp. 35–46.

Ferndández, M., A. Gómez-Pérez, and N. Juristo (1997). *Methontology: From Ontological Art Towards Ontological Engineering*. SS-97-06. Spain: Laboratorio de Inteligencia Artificial, Facultad de Informáitica, Universidad Politécnica de Madrid, p. 8.

Fine, Terrence L. (1999). *Feedforward Neural Network Methodology*. Information Science and Statistics. New York, NY, USA: Springer New York. 340 pp. DOI: 10.1007/b97705.

Gaio, Mauro, Christian Sallaberry, and Van Tien Nguyen (2012). "Typage de noms toponymiques à des fins d'indexation géographique". In: *Traitement Automatique des Langues* 53.2. Publisher: ATALA (Association pour le Traitement Automatique des Langues), pp. 143–176. URL: https://aclanthology.org/2012.tal-2.6.

Garvie, Laurence A. J. (1995). "A semantic net representation for the classification of minerals". In: *Computers & Geosciences* 21.3, pp. 387–396. DOI: 10.1016/0098-3004(94)00083-7.

Gómez-Pérez, Astunción, Natalia Juristo, and Juan Pazos (1995). "Evaluation and Assessment of the Knowledge Sharing Technology". In: *Towards Very Large Knowledge Bases*. Ed. by N. J. I. Mars. Amsterdam: IOS Press, pp. 289–296.

Gouvernement de la République française (2020). *Instruction du Premier ministre du 08 avril 2020 relative au recueil, à la transmission, au traitement et à la diffusion de l'information nautique*. NOR : PRMM2002228A.

Greff, Klaus, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber (2017). "LSTM: A Search Space Odyssey". In: *IEEE Transactions on Neural Networks and Learning Systems* 28.10, pp. 2222–2232. DOI: 10.1109/TNNLS.2016.2582924.

Guarino, Nicola (1998). "Formal Ontology and Information Systems". In: *Formal Ontology in Information Systems*. 1st International Conference on Formal Ontology in Information Systems (FOIS'98). Vol. 46. Frontiers in Artificial Intelligence and Applications. Trento, Italy: IOS Press, pp. 3–15.

Hagaseth, M., L. Lohrmann, A. Ruiz, F. Oikonomou, D. Roythorne, and S. Rayot (2016). "An Ontology for Digital Maritime Regulations". In: *Journal of Maritime Research* 8.2, pp. 7–18.

Hannou, Fatma-Zohra, Victor Charpenay, Maxime Lefrançois, Catherine Roussey, Antoine Zimmermann, and Fabien Gandon (2023). "The ACIMOV Methodology: Agile and Continuous Integration for Modular Ontologies and Vocabularies". In: MK 2023 - 2nd Workshop on Modular Knowledge associated with FOIS 2023 - the 13th International Conference on Formal Ontology in Information Systems. URL: https://hal.laas.fr/hal-04187236.

Hart, Glen and Catherine Dolbear (2007). "What's So Special about Spatial?" In: *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. Ed. by Arno Scharl and Klaus Tochtermann. Advanced Information and Knowledge Processing. London: Springer, pp. 39–44. DOI: 10.1007/978-1-84628-827-2_4.

Hoffart, Johannes, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum (2013). "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia". In: *Artificial Intelligence*. Artificial Intelligence, Wikipedia and Semi-Structured Resources 194, pp. 28–61. DOI: 10.1016/j.artint.2012.06.001.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann (2021). "Knowledge Graphs". In: *ACM Computing Surveys* 54.4, pp. 1–37. DOI: 10.1145/3447772.

Hu, Xuke, Zhiyong Zhou, Yeran Sun, Jens Kersten, Friederike Klan, Hongchao Fan, and Matti Wiegmann (2022). "GazPNE2: A General Place Name Extractor for Microblogs Fusing Gazetteers and Pretrained Transformer Models". In: *IEEE Internet of Things Journal* 9.17. Conference Name: IEEE Internet of Things Journal, pp. 16259–16271. ISSN: 2327-4662. DOI: 10.1109/JIOT.2022.3150967.

Hu, Yingjie, Chengbin Deng, and Zhou Zhou (2019). "A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People toward Their Living Environments". In: *Annals of the American Association of Geographers* 109.4, pp. 1052–1073. DOI: 10.1080/24694452.2018.1535886.

Hydrographic Dictionary Working Group (2019). *S-32 IHO Hydrographic Dictionary*. URL: http://iho-ohi.net/S32/.

Iglesias, Enrique, Samaneh Jozashoori, and Maria-Esther Vidal (2023). "Scaling up knowledge graph creation to large and heterogeneous data sources". In: *Journal of Web Semantics* 75, p. 100755. DOI: 10.1016/j.websem.2022.100755.

International Association of Marine Aids to Navigation and Lighthouse Authorities (2018). *NAVGUIDE*.

International Hydrographic Organization (2000). *S-57 IHO Object Catalogue*. URL: http://www.s-57.com/.

— (2023). *IHO Yearbook*. Periodic Publication P-5. Monaco, p. 361. URL: https://iho.int/uploads/user/pubs/periodical/P5YEARBOOK_ANNUAIRE.pdf.

International Organization for Standardization (2014). *ISO 19101-1:2014. Geographic information — Reference model — Part 1: Fundamentals*. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:19101:-1:ed-1:v1.
— (2019). *ISO 19112:2019. Geographic information — Spatial referencing by geographic identifiers*. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:19112:ed-2:v1.
— (2020). *ISO 24617-7:2020. Language resource management — Semantic annotation framework — Part 7: Spatial information*. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:24617:-7:ed-2:v1.
International Organization for Standardization and International Electrotechnical Commission (2021). *ISO/IEC 21838-1:2021. Information technology — Top-level ontologies (TLO) — Part 1: Requirements*. URL: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:21838:-1:ed-1:v1.
Jain, Lakhmi C. and Larry R. Medsker, eds. (1999). *Recurrent Neural Networks: Design and Applications*. Boca Raton, FL, USA: CRC Press. 416 pp. DOI: 10.1201/9781420049176.
Janowicz, Krzysztof, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby K. Fisher, Ling Cai, Gengchen Mai, Joseph Zalewski, Lu Zhou, Shirly Stephen, Seila Gonzalez, Bryce Mecum, Anna Lopez-Carr, Andrew Schroeder, David Smith, Dawn Wright, Sizhe Wang, Yuanyuan Tian, Zilong Liu, Meilin Shi, Anthony D'Onofrio, Zhining Gu, and Kitty Currier (2022). "Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence". In: *AI Magazine* 43.1, pp. 30–39. DOI: 10.1002/aaai.12043.
Jarrar, M. and R. Meersman (2002). "Formal Ontology Engineering in the DOGMA Approach". In: *Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics*. ODBASE 2002. Vol. 2519. Lecture Notes in Computer Science. Irvine, CA, USA: Springer, pp. 1238–1254. DOI: 10.1007/3-540-36124-3_78.
— (2008). "Ontology Engineering - The DOGMA Approach". In: *Advances in Web Semantics I*. Vol. 4891. Lecture Notes in Computer Science. Berlin and Heidelberg: Springer.
Jiménez–Badillo, Diego, Patricia Murrieta–Flores, Bruno Martins, Ian Gregory, Mariana Favila-Vázquez, and Raquel Liceras-Garrido (2020). "Developing Geographically Oriented NLP Approaches to Sixteenth–Century Historical Documents: Digging into Early Colonial Mexico". In: *Digital Humanities Quarterly* 14.4. ISSN: 1938-4122.
Ju, Meizhi, Makoto Miwa, and Sophia Ananiadou (2018). "A Neural Layered Model for Nested Named Entity Recognition". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. NAACL-HLT 2018. New Orleans, LA, USA: Association for Computational Linguistics, pp. 1446–1459. DOI: 10.18653/v1/N18-1131.
Keller, Antoine (2016). "Extraction, structuration et analyse d'informations géo-historiques : Application aux décrets de l'Assemblée Nationale portant sur le redécoupage administratif lors de la Révolution française." Projet de fin d'études. Brest: École navale. 39 pp.
Keller, Antoine, Nathalie Abadie, Bertrand Dumenieu, Stéphane Baciocchi, and Eric Kergosien (2018). "Vers la construction d'une base de connaissances sur la réorganisation territoriale française à la Révolution". In: Atelier EXtraction de Connaissances à partir de donnéEs Spatialisées (EXCES), SAGEO 2018. Montpellier, France. URL: https://hal.science/hal-02399176.

Kertkeidkachorn, Natthawut and Ryutaro Ichise (2018). "An Automatic Knowledge Graph Creation Framework from Natural Language Text". In: *IEICE Transactions on Information and Systems* E101.D.1, pp. 90–98. DOI: 10.1587/transinf.2017SWP0006.

Kim, Junchul, Maria Vasardani, and Stephan Winter (2015). "Harvesting large corpora for generating place graphs". In: International Workshop on Cognitive Engineering for Spatial Information Processes (CESIP). Vol. 12. Santa Fe, NM, USA.

Kotis, Konstantinos and Andreas Papasalouros (2010). "Learning Useful Kick-off Ontologies from Query Logs: HCOME Revised". In: *2010 International Conference on Complex, Intelligent and Software Intensive Systems*. 2010 International Conference on Complex, Intelligent and Software Intensive Systems. Krakow, Poland: IEEE, pp. 345–351. DOI: 10.1109/CISIS.2010.50.

Kotis, Konstantinos and George A. Vouros (2006). "Human-centered ontology engineering: The HCOME methodology". In: *Knowledge and Information Systems* 10, pp. 109–131. DOI: 10.1007/s10115-005-0227-4.

Krogh, Anders (Feb. 2008). "What are artificial neural networks?" In: *Nature Biotechnology* 26.2. Publisher: Nature Publishing Group, pp. 195–197. DOI: 10.1038/nbt1386.

Kuhn, Werner (2005). "Geospatial Semantics: Why, of What, and How?" In: *Journal on Data Semantics III*. Lecture Notes in Computer Science 3534. Ed. by Stefano Spaccapietra and Esteban Zimányi, pp. 1–24. DOI: 10.1007/11496168_1.

Kuhn, Werner, Tomi Kauppinen, and Krzysztof Janowicz (2014). "Linked Data - A Paradigm Shift for Geographic Information Science". In: *Geographic Information Science*. Ed. by Matt Duckham, Edzer Pebesma, Kathleen Stewart, and Andrew U. Frank. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 173–186. DOI: 10.1007/978-3-319-11593-1_12.

Laddada, Wissame (2018). "Vers une émergence des systèmes d'information géographique maritime fondés sur la connaissance : application aux systèmes d'aide à la navigation". PhD thesis. Brest: Université de Bretagne Occidentale. 139 pp.

Lamotte, Léa (2019). "Extraction de relations et d'entités géographiques pour le peuplement de graphes de connaissances". Master thesis. Institut National des Langues et Civilisations Orientales. 48 pp.

Lamotte, Léa, Nathalie Abadie, Éric Saux, and Eric Kergosien (2020). "Extraction de connaissances pour la description de l'environnement maritime côtier à partir de textes d'aide à la navigation". In: *Revue des Nouvelles Technologies de l'Information*. Conférence nationale EGC 2020 - Extraction et Gestion des Connaissances. Vol. Extraction et Gestion des Connaissances, RNTI-E-36. Bruxelles, Belgium: Éditions RNTI, pp. 341–348. URL: https://hal.archives-ouvertes.fr/hal-03656340.

Leadbetter, A. M., T. Hamre, R. Lowry, Y. Lassoued, and D. Dunne (2010). "Ontologies and Ontology Extension for Marine Environmental Information Systems". In: *Proceedings of the Workshop "Environmental Information Systems and Services - Infrastructures and Platforms"*. EnviroInfo2010. Vol. 679. Bonn, Germany, pp. 12–24.

Lefrançois, Maxime, Antoine Zimmermann, and Noorani Bakerally (2017). "A SPARQL extension for generating RDF from heterogeneous formats". In: *Proceedings of the Extended semantic web conference (ESWC'17)*. Portoroz, Slovenia.

Levinson, Stephen C. (1996). "Language and Space". In: *Annual Review of Anthropology* 25, pp. 353–382.

Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li (2020). "A Survey on Deep Learning for Named Entity Recognition". In: *IEEE Transactions on Knowledge and Data Engineering* 34.1, pp. 50–70. DOI: 10.1109/TKDE.2020.2981314.

Liang, Y. and J. Zhai (2018). "Construction and Representation of Shipping Domain Ontology Based on Ontology Design Patterns". In: *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*. 2018 15th International Conference on Service Systems and Service Management (ICSSSM). Hangzhou, China. DOI: 10.1109/ICSSSM.2018.8465087.

Ling, Xiao, Sameer Singh, and Daniel S. Weld (2015). "Design Challenges for Entity Linking". In: *Transactions of the Association for Computational Linguistics* 3. Place: Cambridge, MA, USA Publisher: MIT Press, pp. 315–328. DOI: 10.1162/tacl_a_00141.

Loynes, Mayeul de and Romain Ruiz (2020). "Construction d'une base de connaissances géoréférencées sur l'environnement hydrographique côtier". Projet de fin d'études Sciences de l'information et de la modélisation. Brest: École navale. 62 pp.

Ma, Xiaogang (2022). "Knowledge graph construction and application in geosciences: A review". In: *Computers & Geosciences* 161, p. 105082. ISSN: 0098-3004. DOI: 10.1016/j.cageo.2022.105082.

Malyankar, R. (2001). "Maritime Information Markup and Use in Passage Planning". In: *Proceedings of the National Conference on Digital Government*. National Conference on Digital Government. Marina del Rey, California, USA, pp. 25–32.

Maniraj, V. and Sivakumar Ramakrishnan (2010). "Ontology Languages – A Review". In: *International Journal of Computer Theory and Engineering* 2, pp. 887–891. DOI: 10.7763/IJCTE.2010.V2.257.

Mansfield, Martin, Valentina Tamma, Phil Goddard, and Frans Coenen (2021). "Capturing Expert Knowledge for Building Enterprise SME Knowledge Graphs". In: *Proceedings of the 11th on Knowledge Capture Conference*. Online (USA): ACM, pp. 129–136. DOI: 10.1145/3460210.3493569.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot (2020). "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. The 58th annual meeting of the Association for Computational Linguistics (ACL). Online: Association for Computational Linguistics, pp. 7203–7219. DOI: 10.48550/arXiv.1911.03894.

Melo, Fernando and Bruno Martins (2017). "Automated Geocoding of Textual Documents: A Survey of Current Approaches". In: *Transactions in GIS* 21.1, pp. 3–38. ISSN: 1467-9671. DOI: 10.1111/tgis.12212.

Moncla, Ludovic (2015). "Automatic reconstruction of itineraries from descriptive texts". PhD thesis. Pau, France: Université de Pau et des Pays de l'Adour. 212 pp.

Moor, A. de, P. De Leenheer, and R. Meersman (2006). "DOGMA-MESS: A Meaning Evolution Support System for Interorganizational Ontology Engineering". In: *Proceedings of the 14th International Conference on Conceptual Structures*. Conceptual Structures: Inspiration and Application. Vol. 4068. Lecture Notes in Computer Science. Aalborg, Denmark: Springer, pp. 189–202. DOI: 10.1007/11787181_14.

Mudgal, Sidharth, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra (2018). "Deep Learning for Entity Matching: A Design Space Exploration". In: *Proceedings of the 2018 International Conference on Management of Data*. SIGMOD/PODS '18: International Conference on Management of Data. Houston, TX, USA: Association for Computing Machinery, pp. 19–34. DOI: 10.1145/3183713.3196926.

Murat, Claire (2023). "L'information nautique se numérise". In: *La Baille, revue de l'AEN et des Associations d'officiers de la Marine* 359, pp. 12–15.

Mustière, Sébastien, Nathalie Abadie, Nathalie Aussenac-Gilles, Marie-Noelle Bessagnet, Mouna Kamel, Eric Kergosien, Chantal Reynaud, Brigitte Safar, and Christian Sallaberry (2011). "Analyses linguistiques et techniques d'alignement pour créer et enrichir une ontologie topographique". In: *Revue Internationale de Géomatique* 21.2, pp. 155–179. DOI: 10.3166/RIG.21.155-179.

Nadeau, David and Satoshi Sekine (2007). "A Survey of Named Entity Recognition and Classification". In: *Lingvisticæ Investigationes* 30.1, pp. 3–26. DOI: 10.1075/li.30.1.03nad.

Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik (2021). "Named Entity Recognition and Relation Extraction: State-of-the-Art". In: *ACM Computing Surveys* 54.1, 20:1–20:39. DOI: 10.1145/3445965.

National Geospatial-Intelligence Agency (2022). *Sailing Directions (Enroute). Pub. 191 English Channel*. 20th ed.

Nguyen, Truc-Vien T. and Tru H. Cao (2007). "VN-KIM IE: Automatic Extraction of Vietnamese Named-Entities on the Web". In: *New Generation Computing* 25.3, pp. 277–292. DOI: 10.1007/s00354-007-0018-4.

Nismi Mol, E. A. and M. B. Santosh Kumar (2023). "Review on knowledge extraction from text and scope in agriculture domain". In: *Artificial Intelligence Review* 56.5, pp. 4403–4445. DOI: 10.1007/s10462-022-10239-9.

Noy, N. F. and D. L. McGuiness (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI- 2001-0880. Stanford, California, USA: Stanford University, p. 25.

Noy, Natasha, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor (2019). "Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done". In: *Queue* 17.2, pp. 48–75. DOI: 10.1145/3329781.3332266.

Paris, Pierre-Henri, Nathalie Abadie, and Carmen Brando (2017). "Linking Spatial Named Entities to the Web of Data for Geographical Analysis of Historical Texts". In: *Journal of Map & Geography Libraries* 13.1, pp. 82–110. DOI: 10.1080/15420353.2017.1307306.

Pastor-Sánchez, J.-A., F. J. Martínez Mendez, and J. V. Rodríguez-Muñoz (2009). "Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives". In: *Information Research* 14.4.

Paulheim, Heiko (2017). "Knowledge graph refinement: A survey of approaches and evaluation methods". In: *Semantic Web* 8.3, pp. 489–508. DOI: 10.3233/SW-160218.

Peroni, Silvio (2013). *Graffoo Specification*. URL: https://essepuntato.it/graffoo/specification/.

— (2016a). "A Simplified Agile Methodology for Ontology Development". In: *OWL: Experiences and Directions – Reasoner Evaluation*. 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016. Bologna, Italy.

— (2016b). "SAMOD: an agile methodology for the development of ontologies". DOI: 10.6084/M9.FIGSHARE.3189769.V2.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Asli Celikyilmaz and Tsung-Hsien Wen. Online: Association for Computational Linguistics, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14.

Qiu, Qinjun, Zhong Xie, Kai Ma, Zhanlong Chen, and Liufeng Tao (2022). "Spatially oriented convolutional neural network for spatial relation extraction from natural language texts". In: *Transactions in GIS* 26.2, pp. 839–866. DOI: 10.1111/tgis.12887.

Qiu, Qinjun, Zhong Xie, Liang Wu, and Liufeng Tao (2020). "Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques". In: *Earth Science Informatics* 13.4, pp. 1393–1410. DOI: 10.1007/s12145-020-00527-9.

Raskin, R. G. and M. J. Pan (2005). "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)". In: *Computers & Geosciences*. Application of XML in the Geosciences 31.9, pp. 1119–1125. DOI: 10.1016/j.cageo.2004.12.004.

Rawsthorne, Helen Mair, Nathalie Abadie, Adrien Guille, Pascal Cuxac, and Cédric Lopez (2024a). "Défi TextMine'24 : Reconnaissance d'entités géographiques dans un corpus des Instructions nautiques". In: *TextMine '24 : Atelier sur la Fouille de Textes*. Atelier TextMine'24, 24ème conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC'24). Dijon, France, pp. 87–92. URL: https://cnrs.hal.science/hal-04434981.

Rawsthorne, Helen Mair, Nathalie Abadie, Eric Kergosien, Cécile Duchêne, and Éric Saux (2022a). "ATLANTIS : Une ontologie pour représenter les Instructions nautiques". In: *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA 2022)*. Ingénierie des Connaissances. Saint-Étienne, France, pp. 154–163. URL: https://hal.archives-ouvertes.fr/hal-03695242.

— (2022b). "ATONTE: Towards A New Methodology for Seed Ontology Development from Texts and Experts". In: *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*. Poster and short article. 23rd International Conference on Knowledge Engineering and Knowledge Management. Vol. 3256. Bozen-Bolzano, Italy: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-3256/paper4.pdf.

— (2023). "Automatic Nested Spatial Entity and Spatial Relation Extraction From Text for Knowledge Graph Creation: A Baseline Approach and a Benchmark Dataset". In: *7th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities '23), November 13, 2023, Hamburg, Germany*. 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities. Hamburg, Germany: Association for Computing Machinery, pp. 21–30. DOI: 10.1145/3615887.3627754.

— (2024b). "Extraction automatique d'entités spatiales imbriquées et de relations spatiales à partir de texte pour la création de graphes de connaissances : Une approche et deux jeux de données". In: *TextMine '24 : Atelier sur la Fouille de Textes*. Atelier TextMine'24, 24ème conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC'24). Dijon, France, pp. 75–86. URL: https://hal.science/hal-04444358.

Reyes-Ortiz, José A. (2019). "Criminal Event Ontology Population and Enrichment using Patterns Recognition from Text". In: *International Journal of Pattern Recognition and Artificial Intelligence* 33.11. Publisher: World Scientific Publishing Co., p. 1940014. DOI: 10.1142/S0218001419400147.

Rodríguez, M. Andrea and Max J. Egenhofer (2004). "Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure". In: *International Journal of Geographical Information Science* 18.3, pp. 229–256. DOI: 10.1080/13658810310001629592.

Rust, Phillip, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych (2021). "How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Lan-

guage Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL-IJCNLP 2021. Online: Association for Computational Linguistics, pp. 3118–3135.
DOI: 10.18653/v1/2021.acl-long.243.

Sallaberry, Christian, Mustapha Baziz, Julien Lesbegueries, and Mauro Gaio (2007). "Une approche d'extraction et de recherche d'information spatiale dans les documents textuels-évaluation." In: *Coria 2007: actes de la quatrième conférence en recherche d'information et applications*. Conférence en Recherche d'Information et Applications. Saint-Étienne, France, pp. 53–64. URL: http://www.asso-aria.org/coria/2007/53.pdf.

Sauvage-Vincent, Julie (2017). "Un langage contrôlé pour les instructions nautiques du Service Hydographique et Océanographique de la Marine". PhD thesis. Université Bretagne Loire. 261 pp. URL: https://www.theses.fr/2017IMTA0001.

Schreiber, Guus, Hans Akkermans, Anjo Anjewierden, Robert de Hoog, Nigel R. Shadbolt, Walter Van de Velde, and Bob J. Wielinga (1999). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press. 471 pp. ISBN: 978-0-262-19300-9. DOI: 10.7551/mitpress/4073.001.0001.

Schröder, Markus, Christian Jilek, and Andreas Dengel (2021). "Spread2RML: Constructing Knowledge Graphs by Predicting RML Mappings on Messy Spreadsheets". In: *Proceedings of the 11th on Knowledge Capture Conference*. K-CAP '21: Knowledge Capture Conference. Online (USA): ACM, pp. 145–152.
DOI: 10.1145/3460210.3493544.

Sequeda, Juan F., Willard J. Briggs, Daniel P. Miranker, and Wayne P. Heideman (2019). "A Pay-as-you-go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases". In: *The Semantic Web – ISWC 2019*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 526–545.
DOI: 10.1007/978-3-030-30796-7_32.

Shen, Wei, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan (2023). "Entity Linking Meets Deep Learning: Techniques and Solutions". In: *IEEE Transactions on Knowledge and Data Engineering* 35.3. Publisher: IEEE Computer Society, pp. 2556–2578. DOI: 10.1109/TKDE.2021.3117715.

Shimizu, C., K. Hammar, and P. Hitzler (2022). "Modular Ontology Modeling". In: *Semantic Web*, pp. 1–31.

Shom (2016). *Signalisation maritime*. 3e. Ouvrages généraux. 62 pp. ISBN: 978-2-11-139457-5.

— (2019a). *Balisage maritime. Descriptif de contenu du produit externe*.

— (2019b). *Symboles, abréviations et termes utilisés sur les cartes marines papier*. 7e. 124 pp. ISBN: 978-2-11-139487-2.

— (2020). *Guide de rédaction des Instructions Nautiques du Shom*. Procédure spécifique. Shom, p. 75.

— (2021a). *Instructions nautiques. C22 : France (côtes Nord et Ouest). Du cap de La Hague à la pointe de Penmarc'h [Version à jour au 20 octobre 2021]*. Brest, France.

— (2021b). *Instructions nautiques. C4 : Afrique (côte Ouest). De Râs Spartel à Cape Palmas - Îles du Large [Version à jour au 27 octobre 2021]*. Brest, France.

— (2021c). *Instructions nautiques. D22 : France (côte Sud). Du cap Croisette à la frontière italienne [Version à jour au 17 novembre 2021]*. Brest, France.

— (2021d). *Instructions nautiques. D6 : Mer Méditerranée, côtes d'Afrique et du Levant [Version à jour au 13 octobre 2021]*. Brest, France.

Shom (2021e). *Instructions nautiques. G4 : Saint-Pierre-et-Miquelon [Version à jour au 22 septembre 2021]*. Brest, France.
— (2021f). *Instructions nautiques. K11 : Îles de l'Océan Pacifique (partie centrale). Île Clipperton [Version à jour au 20 octobre 2021]*. Version à jour au 20 octobre 2021. Brest, France.
— (2021g). *Instructions nautiques. L9 : Îles de l'Océan Indien (partie Sud). Terre Adélie [Version à jour au 27 octobre 2021]*. Version à jour au 20 octobre 2021. Brest, France.
— (2022). *Groupe d'Avis aux Navigateurs en Ligne*. URL: https://gan.shom.fr/diffusion/home.
— (2023). *Assemblage des cartes marines (RasterMarine)*. Brest, France. URL: https://data.shom.fr/.
Shusterman, Anna and Peggy Li (2016). "Frames of reference in spatial language acquisition". In: *Cognitive Psychology* 88, pp. 115–161. DOI: 10.1016/j.cogpsych.2016.06.001.
Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin (2019). "The Performance of LSTM and BiLSTM in Forecasting Time Series". In: *2019 IEEE International Conference on Big Data (Big Data)*. 2019 IEEE International Conference on Big Data (Big Data). Los Angeles, CA, USA, pp. 3285–3292. DOI: 10.1109/BigData47090.2019.9005997.
Simsek, Umutcan, Kevin Angele, Elias Kärle, Juliette Opdenplatz, Dennis Sommer, Jürgen Umbrich, and D. Fensel (2021). "Knowledge Graph Lifecycle: Building and Maintaining Knowledge Graphs". In: *Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference (ESWC 2021)*. 2nd International Workshop on Knowledge Graph Construction. Vol. 2873. Online: CEUR Workshop Proceedings. URL: https://ceur-ws.org/Vol-2873/paper12.pdf.
Southall, Humphrey, Ruth Mostern, and Merrick Lex Berman (2011). "On historical gazetteers". In: *International Journal of Humanities and Arts Computing* 5.2, pp. 127–145. DOI: 10.3366/ijhac.2011.0028.
Stadler, Claus, Jens Lehmann, Konrad Höffner, and Sören Auer (2012). "LinkedGeoData: A core for a web of spatial open data". In: *Semantic Web* 3.4, pp. 333–354. ISSN: 1570-0844.
Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii (2012). "brat: a Web-based Tool for NLP-Assisted Text Annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, pp. 102–107.
Suárez-Figueroa, M. C., A. Gómez-Pérez, and M. Fernández-López (2012). "The NeOn Methodology for Ontology Engineering". In: *Ontology Engineering in a Networked World*. Ed. by M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi. Berlin, Heidelberg: Springer, pp. 9–34.
Suchanek, Fabian, Gjergji Kasneci, and Gerhard Weikum (2008). "YAGO: A Large Ontology from Wikipedia and WordNet". In: DOI: 10.2139/ssrn.3199399.
Suchanek, Fabian M. (2014). "Information Extraction for Ontology Learning". In: *Perspectives on Ontology Learning*. Ed. by Johanna Völker and Jens Lehmann. Vol. 18. Studies on the Semantic Web. IOS Press, pp. 135–151.
Suchanek, Fabian M., Georgiana Ifrim, and Gerhard Weikum (2006). "LEILA: Learning to Extract Information by Linguistic Analysis". In: *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*.

Sydney, Australia: Association for Computational Linguistics, pp. 18–25. URL: https://aclanthology.org/W06-0503.

Sugathadasa, Keet Malin, Vindula Jayawardana, Dimuthu Lakmal, Nisansa de Silva, Amal Shehan Perera, Buddhi Ayesha, and Madhavi Perera (2018). "Word Vector Embeddings and Domain Specific Semantic based Semi-Supervised Ontology Instance Population". In: *The International Journal on Advances in ICT for Emerging Regions* 11.1. URL: https://journal.icter.org/index.php/ICTer/article/view/257.

Sure, Y. (2003). "A Tool-Supported Methodology for Ontology-Based Knowledge Management". In: *The Ontology and Modelling of Real Estate Trasnsactions*. Ed. by H. Stuckenschmidt, E. Stubkjær, and C. Schlieder. International Land Management Series. London: Routledge, pp. 115–126.

Sure, Y. and R. Studer (2001). *On-To-Knowledge Methodology — Employed and Evaluated Version*. D16. Karlsruhe, Germany: Institute AIFB, University of Karlsruhe, p. 56.

Tao, Liufeng, Zhong Xie, Dexin Xu, Kai Ma, Qinjun Qiu, Shengyong Pan, and Bo Huang (2022). "Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model". In: *ISPRS International Journal of Geo-Information* 11.12, p. 598. DOI: 10.3390/ijgi11120598.

To, Huy Quoc, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen (2021). "Monolingual vs multilingual BERTology for Vietnamese extractive multi-document summarization". In: *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. PACLIC 2021. Shanghai, China: Association for Computational Lingustics, pp. 692–699. URL: https://aclanthology.org/2021.paclic-1.73.

Tual, Solenn, Nathalie Abadie, Joseph Chazalon, Bertrand Duménieu, and Edwin Carlinet (2023). "A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents". In: *Proceedings of the 17th International Conference on Document Analysis and Recognition*. The 17th International Conference on Document Analysis and Recognition. San José, CA, USA: Springer. URL: https://hal.science/hal-03994759.

Tzitzikas, Y., C. Allocca, C. Bekiari, Y. Marketakis, P. Fafalios, M. Doerr, N. Minadakis, T. Patkos, and L. Candela (2013). "Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology". In: *Metadata and Semantics Research*. Research Conference on Metadata and Semantic Research. Ed. by E. Garoufallou and J. Greenberg. Communications in Computer and Information Science. Thessaloniki, Greece: Springer International Publishing, pp. 289–301. DOI: 10.1007/978-3-319-03437-9_29.

Vandecasteele, A. and A. Napoli (2012). "Spatial Ontologies for Detecting Abnormal Maritime Behaviour". In: *OCEANS 2012 MTS/IEEE Yeosu: The Living Ocean and Coast - Diversity of Resources and Sustainable Activities*. OCEANS 2012 MTS/IEEE Yeosu Conference. Yeosu, Republic of Korea: IEEE. DOI: 10.1109/OCEANS-Yeosu.2012.6263532.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Velankar, Abhishek, Hrushikesh Patil, and Raviraj Joshi (2022). "Mono vs Multilingual BERT for Hate Speech Detection and Text Classification: A Case Study in Marathi". In: *Artificial Neural Networks in Pattern Recognition*. 10th IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition. Ed. by Neamat El Gayar, Edmondo Trentin, Mirco Ravanelli, and Hazem Abbas. Vol. 13739. Lecture Notes in Computer Science. Dubai, UAE: Springer, pp. 121–128. DOI: 10.1007/978-3-031-20650-4_10.

Vrandečić, D., S. Pinto, C. Tempich, and Y. Sure (2005). "The DILIGENT knowledge processes". In: *Journal of Knowledge Management* 9.5, pp. 85–96. DOI: 10.1108/13673270510622474.

Wang, Chengbin, Xiaogang Ma, Jianguo Chen, and Jingwen Chen (2018). "Information extraction and knowledge graph construction from geoscience literature". In: *Computers & Geosciences* 112, pp. 112–120. DOI: 10.1016/j.cageo.2017.12.007.

Weikum, Gerhard, Xin Luna Dong, Simon Razniewski, and Fabian Suchanek (2021). "Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases". In: *Foundations and Trends in Databases* 10.2, pp. 108–490. DOI: 10.1561/1900000064.

Weinstein, Peter and Gene Alloway (1997). "Seed ontologies: growing digital libraries as distributed, intelligent systems". In: *Proceedings of the second ACM international conference on Digital libraries*. New York, NY, USA: Association for Computing Machinery, pp. 83–91. DOI: 10.1145/263690.263799.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3. Article number: 160018. DOI: 10.1038/sdata.2016.18.

Worboys, Michael and Kathleen Hornsby (2004). "From Objects to Events: GEM, the Geospatial Event Model". In: *Geographic Information Science*. Ed. by Max J. Egenhofer, Christian Freksa, and Harvey J. Miller. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 327–343. DOI: 10.1007/978-3-540-30231-5_22.

Wu, Kehan, Xueying Zhang, Yulong Dang, and Peng Ye (2022). "Deep learning models for spatial relation extraction in text". In: *Geo-spatial Information Science* 26.1, pp. 58–70. DOI: 10.1080/10095020.2022.2076619.

Yadav, Vikas and Steven Bethard (2018). "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In: *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2145–2158. URL: https://aclanthology.org/C18-1182.

Yang, Jiannan, Hong Jia, and Hanbing Liu (2022). "Spatial Relationship Extraction of Geographic Entities Based on BERT Model". In: *Journal of Physics: Conference Series*. 2022 4th International Conference on Artificial Intelligence and Computer Science (AICS 2022). Vol. 2363. Article number: 012031. Beijing, China: IOP Publishing. DOI: 10.1088/1742-6596/2363/1/012031.

Yangarber, Roman, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen (2000). "Automatic Acquisition of Domain Knowledge for Information Extraction". In: *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*. COLING 2000. URL: https://aclanthology.org/C00-2136.

Yao, Yuan, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun (2019). "DocRED: A Large-Scale Document-

Level Relation Extraction Dataset". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 764–777. DOI: 10.18653/v1/P19-1074.

Zhong, Zexuan and Danqi Chen (2021). "A Frustratingly Easy Approach for Entity and Relation Extraction". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2021. Online: Association for Computational Linguistics, pp. 50–61. DOI: 10.18653/v1/2021.naacl-main.5.

Zou, Jinming, Yi Han, and Sung-Sau So (2009). "Overview of Artificial Neural Networks". In: *Artificial Neural Networks: Methods and Applications*. Ed. by David J. Livingstone. Methods in Molecular Biology™. Totowa, NJ: Humana Press, pp. 14–22. DOI: 10.1007/978-1-60327-101-1_2.