AGILE: GIScience Series, 2, 38, 2021. https://doi.org/10.5194/agile-giss-2-38-2021
Proceedings of the 24th AGILE Conference on Geographic Information Science, 2021.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras.
This contribution underwent peer review based on a full paper submission.

Educing knowledge from text: semantic information extraction of spatial concepts and places

Evangelos Papadias[a], Margarita Kokla[b] (corresponding author) and Eleni Tomai[b]

papadias@hua.gr, mkokla@survey.ntua.gr, etomai@mail.ntua.gr

[a]Geography Department, Harokopio University of Athens, Athens, Greece
[b]School of Rural and Surveying Engineering, National Technical University of Athens, Athens, Greece

Abstract. A growing body of geospatial research has shifted the focus from fully structured to semi-structured and unstructured content written in natural language. Natural language texts provide a wealth of knowledge about geospatial concepts, places, events, and activities that needs to be extracted and formalized to support semantic annotation, knowledge-based exploration, and semantic search. The paper presents a web-based prototype for the extraction of geospatial entities and concepts, and the subsequent semantic visualization and interactive exploration of the extraction results. A lightweight ontology anchored in natural language guides the interpretation of natural language texts and the extraction of relevant domain knowledge. The approach is applied on three heterogeneous sources which provide a wealth of spatial concepts and place names.

Keywords: semantic annotation, ontology-based information extraction, natural language, spatial concepts, spatial entities

1 Introduction

A growing body of geospatial research has shifted the focus from fully structured content (such as spatial databases and GISs) to semi-structured (such as webpages) and unstructured content (natural language texts). Although ambiguous and imprecise, unstructured content provides a wealth of information on places, geographic features, events, and activities capturing authoritative as well as common-sense conceptualizations. Place descriptions from user-generated content have proved to be valuable tools to elicit linguistic, semantic, and cognitive aspects of places and landscapes (Ballatore and Adams, 2015; Derungs and Purves, 2014).

Semantic information extraction aims at eliciting salient, specific types of information from natural language texts. Although there exists semantically-enabled tools for information extraction, the process is not straightforward, since existing tools are based on their own general underlying knowledge and do not support the extraction of domain concepts.

Ontology based information extraction (OBIE) aims at making domain knowledge explicit by employing domain ontologies in order to formally describe domain knowledge and assist the extraction of pre-defined domain concepts, properties, relations, and instances. Ontologies provide the basis for relating natural language terms to their meaning through the concepts they express and other conceptual knowledge. Educing concepts, relations, and place names from natural language texts and linking these to other relevant resources supports semantic annotation, knowledge-based exploration, and semantic search.

The paper presents a web-based prototype for the extraction of geospatial entities and concepts, and the subsequent semantic visualization and interactive exploration of the extraction results. A lightweight generic geospatial ontology with a natural language anchorage, is used to interpret the input texts and guide the semantic information extraction process. The aim of the prototype is:

- to enrich natural language texts with spatial concepts and entities
- to unveil immanent spatial knowledge from texts that can be formally described and further processed for semantic analysis of textual resources

The remainder of the paper is organized as follows. Section 2 reviews relevant work regarding geospatial information extraction. Section 3 presents the workflow of the spatial and semantic information extraction and the development of the web-based prototype. Finally, Section 4 draws conclusions and discusses future directions.

2 Related Work

Semantic enrichment aims at enhancing content interlinkage, search, and discovery by adding well-defined semantic metadata that help machines make sense of the content and reveal latent relations. It is used for information organization, semantic search, and ontology development and population. Semantic enrichment has been used to add semantic metadata to different types of content, such as unstructured documents (Pernelle, 2016), maps (Hu et al., 2015), images (Ennis et al., 2015; Tardy et al., 2016) and videos (Nixon et al., 2013).

Information extraction plays a central role in this process since it supports the automatic processing of unstructured or semi-structured natural language texts and the retrieval of certain types of information that are relevant for the task at hand while ignoring other types of information. Ontology based information extraction (OBIE) (Wimalasuriya and Dou, 2010) is a subfield of IE, in which an ontology that formally describes domain knowledge guides the extraction of concepts, properties, and instances inherent of the domain.

Geospatial-oriented approaches to semantic information extraction are used for tasks such as the spatialization of text corpora, the exploration of linguistic descriptions of space and places, and geographic information search and retrieval. These approaches explore natural language texts with the aim of eliciting various types of information, such as places (O'Hare and Murdock, 2013; Purves et al., 2011; Vasardani et al., 2013), events (Wang and Stewart, 2015), locative expressions (Liu et al., 2014), activities (Hobel and Fogliaroni, 2016), and emotions (Ballatore and Adams, 2015).

Most of these approaches to semantic information extraction from natural language texts commonly use gazetteers to extract place names and relations among them. Vocabularies and taxonomies are also used to extract types of places, events or activities. In terms of acquiring conceptual geospatial knowledge, topic modeling techniques are mostly employed to identify abstract topics that describe a text collection. However, semantic information extraction can also be guided by an ontology and its rich representation of domain knowledge.

Hu et al. (2015) designed a specific ontology based on ArcGIS Online schema to extract entities and classes from map titles and descriptions, to support knowledge discovery for ArcGIS Online.

Ballatore and Adams (2015) developed a vocabulary of place nouns of natural and built places and extracted place emotions from a corpus of travel blog posts based on the emotion vocabulary WordNet-Affect (Strapparava and Valitutti, 2004).

Wang and Stewart (2015) extracted spatiotemporal and semantic information for natural hazards from web news reports. The process was based on a hazard ontology developed from authoritative sources, integrated with spatial, temporal, and semantic gazetteers to account for three aspects of hazards.

Stock and Yousaf (2018) propose an instance-based learning approach for the interpretation of natural language descriptions of location. The approach is based on two ontologies that model spatial relations between point, line, and area features and characteristics used to represent context for geographic features.

The present paper describes the implementation of a web-based prototype for semantic information extraction and visualization of textual resources. The prototype employs ontology-based information extraction of spatial concepts and places and also supports the semantic visualization and exploration of the results.

3 Description of the Web-based Prototype

A web-based prototype is developed using the Shiny R package created by RStudio for building interactive web applications. The prototype implements the extraction of locations and concepts and the subsequent semantic visualization of the extraction results. Significant effort has been put to create a pipeline of processes, to achieve easy future extendibility for similar projects.

The demonstration of the prototype involves three sources of documents: (a) a corpus of 159 geospatial educational resources derived from a crowdsourced educational platform, (b) 26 Chapters from the book 'World Regional Geography: People, Places and Globalization' (2016) and (c) 11 articles with various themes from heterogeneous sources (BBC, NY Times, Nature, etc.). All three sources include wealth of geospatial information in terms of place names and spatial concepts referring to natural and manmade spatial features, but also to geospatial primitives, spatial relations, natural and social processes, etc.

To support the extraction of such a wealth of geospatial knowledge, the semantic information extraction process is guided by a lightweight generic geospatial ontology (Kavouras et al., 2016; Kokla et al., 2018) with a natural language anchorage based on WordNet (Fellbaum, 1998). The ontology is used to formally describe domain knowledge and assist the extraction of pre-defined ontology concepts. It includes 342 concepts referring to spatial features (e.g., mountains, cities, countries, etc.) but also to geospatial primitives, spatial relations, natural and social processes, human and physical systems, etc.

3.1 Semantic Information Extraction

Fig. 1 provides an overview of the semantic information extraction process, which consists of four steps: (1) pre-processing, (2) core natural language processing, (3) location extraction, and (4) concept extraction and linking.

3.1.1 Pre-processing

Initially, a pre-processing step is performed to prepare the texts for the subsequent processing. The Poppler utility library (Poppler, 2021) was used for rendering Portable Document Format (pdf) documents. The texts were then cleaned in order to remove html characters, email addresses, hyphenation, and other special characters and symbols such as bullets, etc. that do not add value to the annotation process.

3.1.2. Natural Language Processing

This step performs the core natural language analysis using the Stanford CoreNLP Natural Language Processing Toolkit. The toolkit provides an annotation-based NLP processing pipeline that takes as input the text corpus and carries out linguistic analysis to derive annotations for the texts: (a) tokenization to split text into words, phrases, symbols, or other meaningful elements called tokens, (b) sentence splitting to divide the texts into sentences, (c) part-of-speech (POS) tagging to mark up each phrase as corresponding to a particular part of speech, i.e., noun phrases, verb phrases, adjective phrases, adverb phrases, etc., and (d) lemmatization to identify the base or dictionary form of a word (lemma).

3.1.3 Location Extraction

The process of entity extraction involves the identification of mentions of entities in a text, such as persons, locations, organizations, time and their association to a reference Knowledge Base (Martinez-Rodriguez et al., 2018). The work focused specifically on the extraction of locations mentioned in the input texts. The Named Entity Recognizer from Stanford CoreNLP software was applied, which uses machine-learning sequence models to label entities. For example, for the Chapter entitled “Regions in Geography”, the process retrieved place names such as France, Canada, Rocky Mountains, New England, United States, Mexico, Rio Grande, Europe, Switzerland, Italy, etc. (Fig. 2).

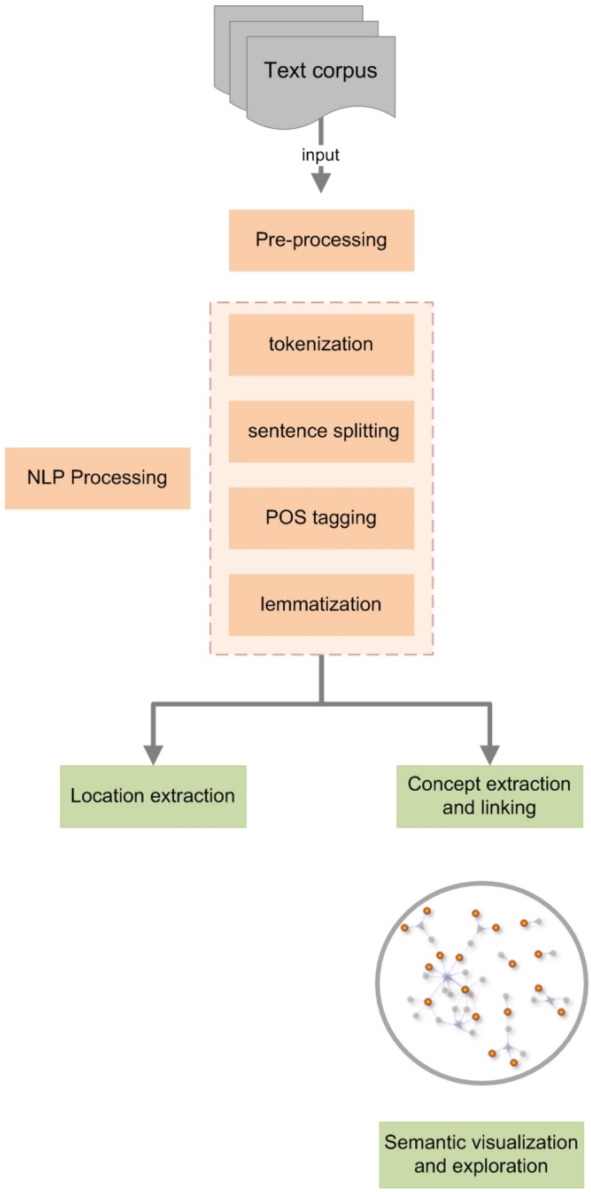

Figure 1. Overview of the semantic information extraction process.

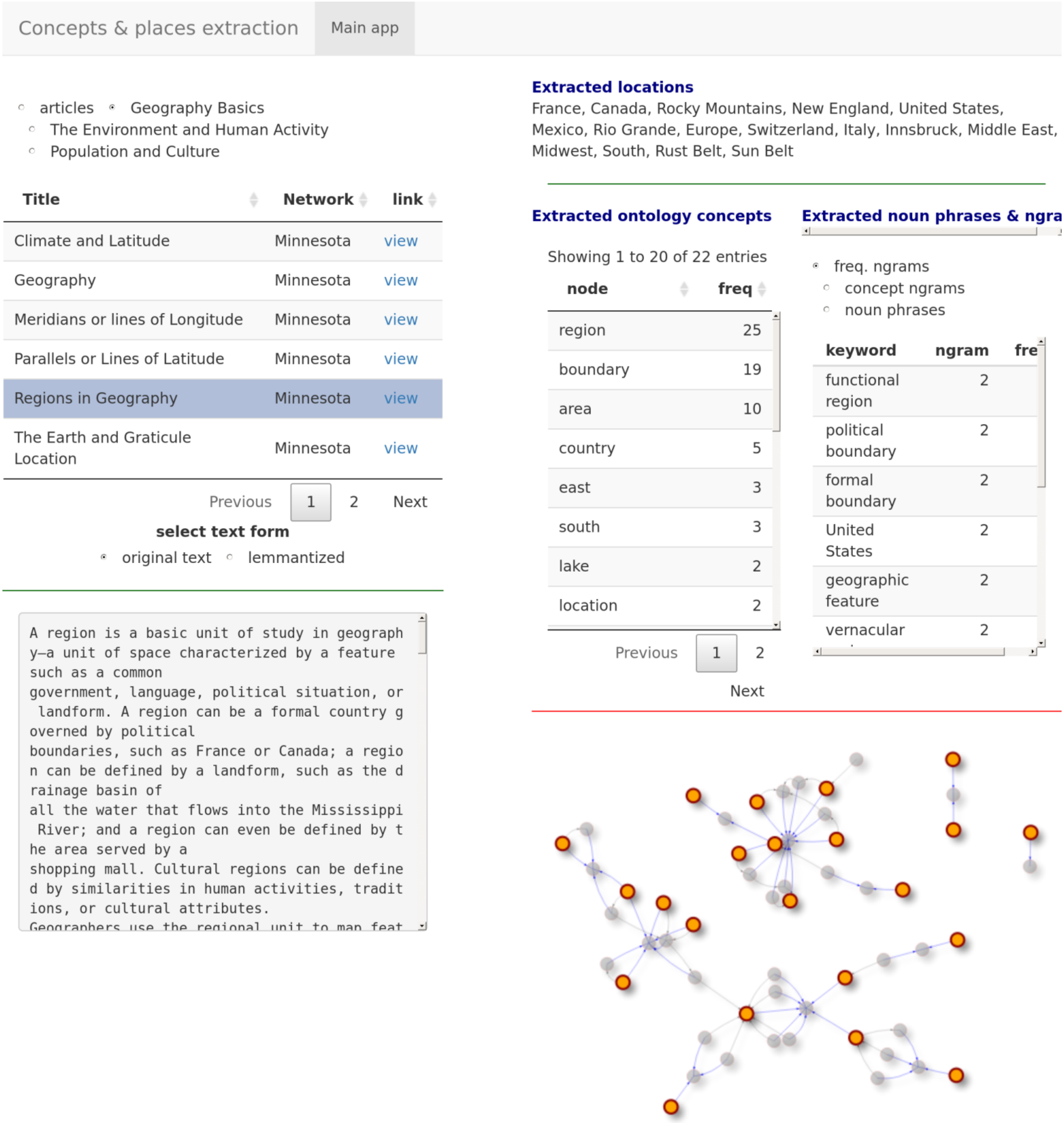

Figure 2. Location and concept extraction for the Chapter "Regions in Geography".

3.1.4. Concept Extraction and Linking

Concept extraction refers to the identification of keywords and key-phrases that represent spatial concepts (Martinez-Rodriguez et al., 2018). Typically, these are nouns and noun phrases correspondingly that appear frequently in a given document and are considered to correspond to the main topic/ theme of the document. A combination of extraction methods is implemented in the workflow:

(a) an ontology-based concept extraction method is used to identify nouns and noun phrases corresponding to ontology concepts. The concept extraction process used a string-matching technique between ontology concept labels and the input texts.

(b) POS-tagging and shallow syntactic patterns are used to identify noun phrases ('population growth', 'winter season', 'map scale', time zone'), as well as noun phrases prefixed by adjectives ('climatic seasons', 'periodic motion') and calculate their frequencies.

(c) a window-based method is used to extract the most frequently used n-grams, as well as n-grams that contain an ontology concept.

The extraction methods b and c are implemented in order to assist the identification of candidate ontology concepts for subsequent ontology enrichment.

The identified keywords and key-phrases that correspond to spatial concepts are further linked to the reference ontology to support the subsequent semantic visualization and exploration.

Fig. 2 shows the extraction of concepts (e.g., region, boundary, area, country, east, south, etc.) and frequent n-grams (e.g., functional region, political boundary, formal boundary, geographic feature, etc.) from the chapter 'Regions in Geography'.

3.2 Semantic Visualization and Exploration

The web-based prototype supports the visualization and exploration of the extraction results (Fig. 2). The user is able to browse the list of texts, select and view a text (both the original and lemmatized version), and also the extracted locations and concepts. The concepts extracted are also semantically visualized using a semantic network. Fig. 3 shows the semantic visualization of the Chapter "The Earth and Graticule Location": the extracted concepts are highlighted in orange. The user may also explore the immediate neighbours of the extracted concepts and also view the definitions of concepts based on the reference ontology.

3.3 Data and Software Availability

Research data and code chunks supporting the findings of this publication are available in https://github.com/veegpap/Concepts-places-extraction under a MIT license and are accessible via the DOI https://doi.org/10.5281/zenodo.4717699. The DOI also includes a minimal reproducible example of the computational workflow supporting this publication as an R script with instructions included in the file README.md in the repository. All used packages are available on CRAN.

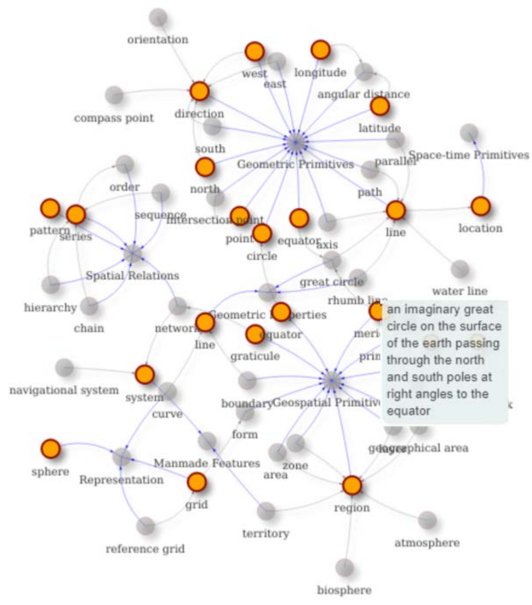

Figure 3. Semantic visualization of extracted concepts.

4 Discussion and Conclusions

The paper presents the implementation of a web-based prototype for the extraction and visualization of geospatial semantic information from natural language texts with the use of a light-weight ontology. The demonstration of the prototype on a heterogeneous corpus revealed the validity of the approach for extracting geospatial information and gave insights on the transferability and extensibility potential to other types of corpora.

The workflow may be used for tagging large collections of texts with geospatial concepts and place names to explore how this knowledge is described in natural language. The result may also be used to connect the extracted locations with the spatial concepts to which they refer to further explore the semantic relationships between places and spatial concepts latent in the input documents.

The proposed OBIE workflow may also be applied for a corpus-based ontology validation. Such validation endeavour compares the learned ontology with the content of a text corpus that covers significantly a given domain and seeks to explore how much a given domain is covered by the ontology.

Extracting the semantic and spatial information from unstructured and semi-structured texts poses significant challenges due to the use of natural language and the resulting issues of lexical polysemy and contextdependence. Place names and concept terms are not always monosemous requiring place name and word sense detection and disambiguation. For the specific corpora, place name disambiguation did not present significant challenges. On the other hand, word sense disambiguation presents more challenges due to the polysemy and ambiguity related to natural language and has not been substantially dealt with herein. Machine learning techniques such as deep learning may be used to augment the existing techniques and support more challenging processes such as entity and concept disambiguation, attribute and relation extraction, and axiom learning.

Acknowledgements

The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the First Call for H.F.R.I. Research Projects to support "Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: HFRI-FM17-2661).

References

Ballatore, A. and Adams, B.: Extracting Place Emotions from Travel Blogs, in: AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities, Lecture Notes in Geoinformation and Cartography, 1–5, 2015.

Derungs, C. and Purves, R. S.: From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus, International Journal of Geographical Information Science, 28, 1272–1293, https://doi.org/10.1080/13658816.2013.772184, 2014.

Ennis, A., Nugent, C., Morrow, P., Chen, L., Ioannidis, G., Stan, A., and Rachev, P.: A Geospatial Semantic Enrichment and Query Service for Geotagged Photographs, Sensors, 15, 17470–17482, https://doi.org/10.3390/s150717470, 2015.

Fellbaum, C.: WordNet: an electronic lexical database, MIT Press, Cambridge, Mass, 1998.

Hobel, H. and Fogliaroni, P.: Extracting Semantics of Places from User Generated Content, in: The 19th AGILE International Conference on Geographic Information Science, 2016.

Hu, Y., Janowicz, K., Prasad, S., and Gao, S.: Enabling Semantic Search and Knowledge Discovery for ArcGIS Online: A Linked-Data-Driven Approach, in: AGILE 2015: Geographic Information Science as an Enabler of Smarter Cities and Communities, edited by: Bacao, F., Santos, M. Y., and Painho, M., Springer International Publishing, Cham, 107–124, https://doi.org/10.1007/978-3-319-16787-9_7, 2015.

Kavouras, M., Kokla, M., Tomai, E., Darra, A., and Pastra, K.: GEOTHNK: A Semantic Approach to Spatial Thinking, in: Progress in Cartography: EuroCarto 2015, edited by: Gartner, G., Jobst, M., and Huang, H., Springer International Publishing, Cham, 319–338, https://doi.org/10.1007/978-3-319-19602-2_20, 2016.

Kokla, M., Papadias, V., and Tomai, E.: Enrichment and Population of a Geospatial Ontology for Semantic Information Extraction, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLII–4, 309–314, https://doi.org/10.5194/isprs-archives-XLII-4-309-2018, 2018.

Liu, F., Vasardani, M., and Baldwin, T.: Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis, in: Proceedings of the 4th International Workshop on Location and the Web, New York, NY, USA, event-place: Shanghai, China, 9–16, https://doi.org/10.1145/2663713.2664426, 2014.

Martinez-Rodriguez, J. L., Hogan, A., and Lopez-Arevalo, I.: Information extraction meets the Semantic Web: A survey, SW, 1–81, https://doi.org/10.3233/SW-180333, 2018.

Nixon, L., Bauer, M., and Bara, C.: Connected Media Experiences: Web Based Interactive Video Using Linked Data, in: Proceedings of the 22Nd International Conference on World Wide Web, New York, NY, USA, event-place: Rio de Janeiro, Brazil, 309–312, https://doi.org/10.1145/2487788.2487931, 2013.

O'Hare, N. and Murdock, V.: Modeling locations with social media, Inf Retrieval, 16, 30–62, https://doi.org/10.1007/s10791-012-9195-y, 2013.

Pernelle, N.: Semantic enrichment of data: annotation and data linking, thesis, Université Paris Sud, 2016.

Poppler: https://poppler.freedesktop.org/, last access: 23 April 2021.

Purves, R., Edwardes, A., and Wood, J.: Describing place through user generated content, FM, https://doi.org/10.5210/fm.v16i9.3710, 2011.

Stock, K. and Yousaf, J.: Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data, International Journal of Geographical Information Science, 32, 1087–1116, https://doi.org/10.1080/13658816.2018.1432861, 2018.

Strapparava, C. and Valitutti, A.: Wordnet-affect: an affective extension of wordnet, in: 4th International Conference on Language Resources and Evaluation, 1083–1086, 2004.

Tardy, C., Falquet, G., and Moccozet, L.: Semantic enrichment of places with VGI sources: a knowledge based approach, in: Proceedings of the 10th Workshop on Geographic Information Retrieval - GIR '16, the 10th Workshop, Burlingame, California, 1–2, https://doi.org/10.1145/3003464.3003470, 2016.

University of Minnesota: World Regional Geography, University of Minnesota Libraries Publishing edition, https://doi.org/10.24926/8668.2701, 2016.

Vasardani, M., Winter, S., and Richter, K.-F.: Locating place names from place descriptions, 27, 2509–2532, https://doi.org/10.1080/13658816.2013.785550, 2013.

Wang, W. and Stewart, K.: Spatiotemporal and semantic information extraction from Web news reports about natural hazards, Computers, Environment and Urban Systems, 50, 30–40, https://doi.org/10.1016/j.compenvurbsys.2014.11.001, 2015.

Wimalasuriya, D. C. and Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches, Journal of Information Science, 36, 306–323, https://doi.org/10.1177/0165551509360123, 2010.