AGILE Pre-conference workshop
“Bridging space, time, and semantics in GIScience”
Wageningen, The Netherlands, 9 May 2017

Semantic enrichment of user-generated educational scenarios with spatial concepts and entities

Marinos Kavouras
National Technical University of Athens
9, H. Polytechniou Str., 15780
Athens, Greece
mkav@mail.ntua.gr

Margarita Kokla
National Technical University of Athens
9, H. Polytechniou Str., 15780
Athens, Greece
mkokla@survey.ntua.gr

Eleni Tomai
National Technical University of Athens
9, H. Polytechniou Str., 15780
Athens, Greece
etomai@mail.ntua.gr

Abstract

The paper presents ongoing work on how to semantically enrich user-generated content of an educational authoring environment with spatial concepts and entities. The main purpose of such an enrichment is to improve search results for content. Analysis of existing user-generated educational scenarios has shown that although they contain a wealth of spatial concepts and entities, the search mechanisms supporting keyword-, map-, and graph-based search return positive searches only for a limited number of concepts and entities based on those attached to scenarios which are very limited as opposed to what is mentioned in the content. The methodology makes use of NLP techniques and open tools and exploits the environment’s semantic network of spatial thinking concepts to further annotate user-generated educational scenarios. Even though, the work is in its initial phases, results indicate that the approach has potential in improving semantic search and discovery.

Keywords: semantic enrichment, spatial concepts, user-generated content, semantic search, semantic metadata.

1 Introduction

XXX is an online, collaborative repository and authoring environment for the collection, creation, and sharing of educational resources for enhancing spatial thinking developed in the framework of a European project through an innovative ICT-based approach.

Resources are either simple educational objects (such as documents, images, videos, webpages, etc.) or more elaborate educational scenarios that describe a whole lesson plan or educational activity. Currently the platform has 714 registered members who have contributed in different ways to the development of 437 resources.

A semantic network of 327 concepts and 802 taxonomic relations supports the development of educational scenarios in 6 languages. The semantic network supports semantic organization of spatial knowledge but also enables the graph-based search of educational resources. Furthermore, it supports the visualization of scenarios by highlighting concepts (nodes) of the overall network present in the scenario.

Another functionality of the platform is the visualization of educational scenarios in geographic space. This functionality is supported by dynamically drawing 2,158,751 geographic entities from GeoNames geographical database which may be attached by users to scenarios as instances of certain concepts (e.g. country, lake, city etc.).

The platform provides various possibilities for searching the content, such as keyword-search, metadata search (based on metadata provided by the authors such as target audience, language, discipline, etc.), map-based search (based on geographic entities attached to a scenario), and graph-based search (based on the spatial concepts attached to a scenario).

The ability to attach spatial concepts and instances to educational scenarios is an important feature of the platform because it supports the semantic and map-based search and visualization of content. However, this feature was not widely used by users limiting thus the potential for advanced search and visualization of the scenarios in semantic and map space. Although various spatial concepts and entities are mentioned in the majority of educational scenarios, users have not actually attached these in their scenarios for allowing the subsequent semantic and map visualization. For example, a semantic search on “geological phenomena in the Mediterranean” would ideally return results about “deforestation in Cyprus” or “degradation in Greece”.

The semantic enrichment of user-generated scenarios would greatly improve the search and discovery capabilities provided by the system, for example by supporting the following:

1. Search for scenarios based on multiple concepts to provide more specific answers to the users’ search
2. Identification of similar scenarios based on the sets of concepts treated therein.
3. Linking of related/similar content
4. Enrichment of the search results by also taking into account the relations among concepts. For example, equivalence relations may provide for identifying relevant content using synonymous terms that refer to the same concepts. Hierarchical relations may contribute in expanding or limiting the exploration of relevant content.

The remainder of the paper is organized as follows. Section 2 reviews relevant literature regarding semantic enrichment and Section 3 outlines the elaboration of the existing semantic infrastructure into an ontology. Section 4 presents the method for the semantic enrichment of user-generated educational scenarios and Section 5 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. Semantic enrichment is used for information organization and retrieval, semantic search and knowledge discovery, 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 video (Nixon et al., 2013). Semantic enrichment mainly focuses on extracting entities and concepts; however, there are approaches targeting at extracting semantic metadata for more specific types of features such as events (Romero and Becker, 2016), movement data (Fileto et al., 2015), and places (Alves et al., 2009; Tardy et al., 2016).

The paper focuses on the semantic enrichment of user-generated educational scenarios describing lesson plans for enhancing students’ spatial thinking abilities. These include both unstructured and structured parts. Unstructured parts are documents describing the lesson plans whereas structured parts include tags, metadata, concepts, and geographic entities attached to the scenarios. The work aims at extracting semantic information from the unstructured parts scenarios in order to connect it with the structured parts, primarily with concepts and geographic entities. Text enrichment with geographic entities is more straightforward supported by semantic annotation tools and gazetteers. On the other hand, semantic enrichment of unstructured text with concepts is not a straightforward process, since existing semantically-enabled tools are based on their own underlying ontologies and do not support the extraction of specialized concepts.

For example, the sentence:

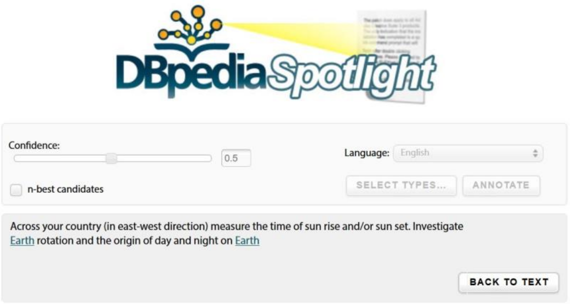
“*Across your country (in east-west direction) measure the time of sun rise and/or sun set. Investigate Earth rotation and the origin of day and night on Earth*”

includes spatial concepts such as: country, east, west direction, rotation and Earth. Figures 1 and 2 show the semantic enrichment of this sentence as performed by DBpedia Spotlight (Daiber et al., 2013) and Open Calais (Butuc, 2009) respectively. DBpedia Spotlight annotates the sentence with the concept Earth, whereas Open Calais does not identify any semantic annotations in this sentence.

3 From semantic network to ontology

A set of 327 spatial concepts with definitions derived from WordNet (2010) were organised in a three-level hierarchy that reflects clusters of basic notions and subject areas such as geometric primitives, spatial relations, space-time primitives, geography etc. Four major clusters form the upper level hierarchy, further dived into 15 clusters that represent the second level and finally the 327 concepts are organised within them. WordNet synsets were scrutinized to identify lemmas and definitions that better served the spatial thinking domain.

Figure 1. Semantic annotation with DBpedia Spotlight.



Confidence: 0.5

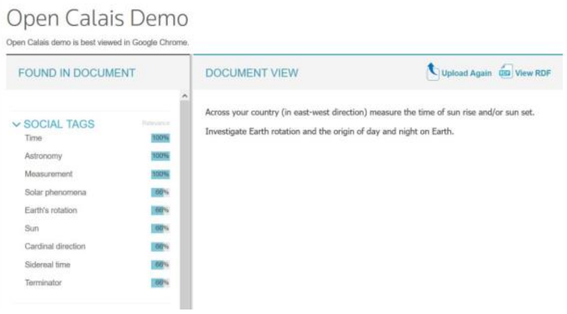
□ n-best candidates

Language: English

Across your country (in east-west direction) measure the time of sun rise and/or sun set. Investigate Earth rotation and the origin of day and night on Earth

Source: DBpedia Spotlight (http://demo.dbpedia-spotlight.org/).

Figure 2. Semantic annotation with Open Calais.



SOCIAL TAGS	Relevance
Time	100%
Astronomy	100%
Measurement	100%
Solar phenomena	66%
Earth's rotation	66%
Sun	66%
Cardinal direction	66%
Sidereal time	66%
Terminator	66%

Across your country (in east-west direction) measure the time of sun rise and/or sun set. Investigate Earth rotation and the origin of day and night on Earth.

Source: Open Calais (http://www.opencalais.com/).

In an attempt to keep the structure as simple as possible, concepts were semantically interlinked in a rather “loose” way through 802 unlabelled relations forming a semantic network.

However, a deeper WordNet analysis provided further information that was not originally used for the development of the semantic network but would greatly enhance the platform’s capabilities in terms of knowledge search and discovery. More specifically, the analysis showed excessive interlinkage of concepts in terms of semantic relations (subtype / supertype, part-of / has part, sister terms, etc.) and indicated that 26 domains that the 327 concepts belong to. A concept belongs to more than one domains as shown by the analysis. In average, concepts belong to three or four domains, while very few of them belong only to one domain. This finding reveals additional 461 relations among concepts and provides ground for developing an ontology in a more rigid and formal way.

An example shown in Figures 3 and 4 indicates how the semantic network is elaborated once the relations between concepts are explicitly denoted and domains are added to the structure. Figure 3 shows how the concept Cartesian coordinate is interrelated with other concepts in the semantic network. Straight lines denote the hierarchy of the concepts, while curves denote relations, indicating that certain concepts are somehow interrelated. On the contrary, Figure 4 constitutes an excerpt of the ontology visualization where all relations are explicitly stated and enriched with ontological domains not previously included in the semantic network.

Therefore, in order to semantically enriching user-generated content and improve semantic search and knowledge discovery, an ontology was developed that relates the 327 original concepts with explicit relations among them and domains that they belong to.

Figure 3. Relations currently in the semantic network.

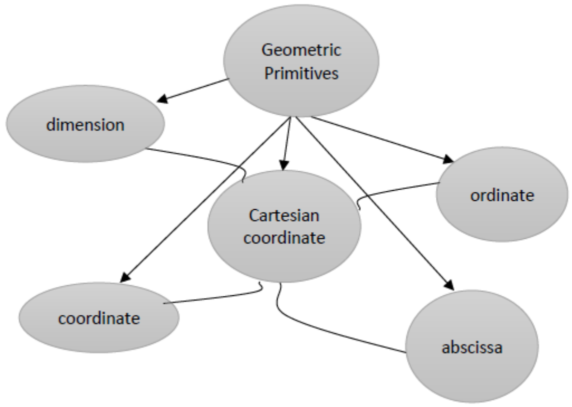
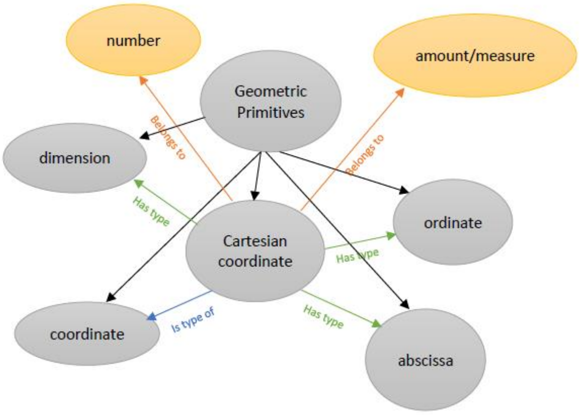


Figure 4. How the ontology changes the semantic network.



4 Semantic enrichment of user-generated scenarios with spatial concepts and entities

An examination of the implemented scenarios reveals that although users have attached concepts and instances to the scenarios (2947 and 1475 attachments of concepts and instances respectively) they have not fully exploited this potential.

There are various scenarios with less than five attached concepts that mention a multitude of concepts in their descriptions. These free text descriptions could be semantically annotated and enriched with these concepts to enforce search and discovery results. Moreover, an even larger number of scenarios do not have any geographic entity attached to them, although various geographic entities are included in the corresponding free-text descriptions. These geographic entities can be discovered through the platform’s keyword-based search mechanisms, but cannot be retrieved in the map-search.

In what follows, the procedure of such an enrichment is detailed. To perform the task, we used the GATE Information Extraction System (Cunningham et al., 2017) that integrates several NLP systems apart from having its own, ANNIE, as default.

4.1 Enrichment process

The enrichment endeavour comprises of four phases.

4.1.1 Text selection

Text collection constitutes the pre-processing phase which considers user-generated scenarios to spot those with a limited number of concepts that usually have no instances of concepts attached to them as well.

4.1.2 Preprocessing a sample

Once the texts to be enriched have been identified, a sample was used as input to GATE to be tested and analyzed. This formed a corpus of the scenarios involved.

The authoring environment allows downloading of the content in the form of pdf files which when uploaded on GATE present some problems in structure (e.g. connection or repetition of words etc.). Before running the NLP process, the files were processed manually to tackle these issues.

4.1.3 Processing the corpus

For the NLP processing, we ran two tests. First, we used ANNIE, GATE’s default system with its set of NLP tools (tokenizer, gazetteer, sentence splitter, and POS tagger). In a separate session, we used the OpenNLP (2010) package, an open source library for processing texts using machine learning tools developed by the Apache Software Foundation. OpenNLP is fully integrated in GATE and offers an alternative to ANNIE. Both toolkits may be used together in a single processing task with modules taken from each in any combination.

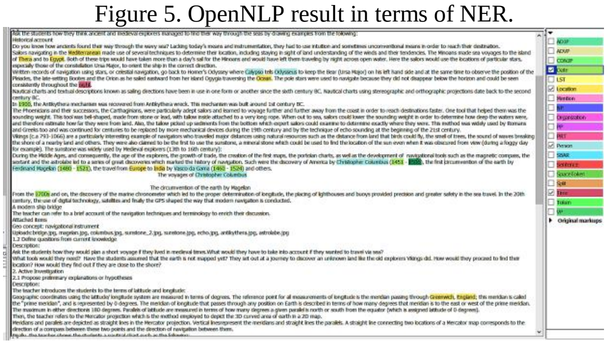
In our case, the NLP processing gave better results using OpenNLP toolkit (Figure 5). OpenNLP identified nearly all named entities in the file (e.g. London, Mediterranean, Thera, etc.) that could serve as instances of certain concepts during the enrichment phase that follows. OpenNLP has its own models for the following named entities: location, time, organization, person, money, date, percentage. In our case, location is the one that we need the most to identify instances for concepts such as city, country, continent, island, etc.

4.1.4 Enriching the scenario

Upon completion of the NLP processing phase with noun phrases being identified, we are able to annotate concepts and geographic entities in the text by uploading the developed ontology in GATE using the ontology tools plugin.

The ontology was uploaded as an xml file in order to identify concepts in the processed corpus. Furthermore, instances of certain concepts, identified in step 3 as named entities, were annotated as well, populating thus the ontology itself. Figure 6 shows an excerpt of such an enrichment in terms of annotated concepts and instances, while Figure 7 shows an excerpt of the resulting populated ontology.

Figure 5. OpenNLP result in terms of NER.



Ask the students how they think ancient and medieval explorers managed to find their way through the seas by drawing examples from the following:
Historical account
Do you know how ancients found their way through the wavy sea? Lacking today's means and instrumentation, they had to use intuition and sometimes unconventional means in order to reach their destination.
Sailors navigating in the Mediterranean made use of several techniques to determine their location, including staying in sight of land understanding of the winds and their tendencies. The Minoans made sea voyages to the island of Thera and to Egypt. Both of these trips would have taken more than a day's sail for the Minoans and would have left them traveling by night across open water. Here the sailors would use the locations of particular stars, especially those of the constellation Ursa Major, to orient the ship in the correct direction.
Written records of navigation using stars, or celestial navigation, go back to Homer's Odyssey where Calypso tells Odysseus to keep the Bear (Ursa Major) on his left hand side and at the same time to observe the position of the Pleiades, the late-setting Bootes and the Orion as he sailed eastward from her island Ogygia traversing the Ocean. The pole stars were used to navigate because they did not disappear below the horizon and could be seen consistently throughout the night.
Nautical charts and textual descriptions known as sailing directions have been in use in one form or another since the sixth century BC. Nautical charts using stereographic and orthographic projections date back to the second century BC.
In 1900, the Antikythera mechanism was recovered from Antikythera wreck. This mechanism was built around 1st century BC.
The Phoenicians and their successors, the Carthaginians, were particularly adept sailors and learned to voyage further and further away from the coast in order to reach destinations faster. One tool that helped them was the sounding weight. This tool was bell-shaped, made from stone or lead, with tallow inside attached to a very long rope. When out to sea, sailors could lower the sounding weight in order to determine how deep the waters were, and therefore estimate how far they were from land. Also, the tallow picked up sediments from the bottom which expert sailors could examine to determine exactly where they were. This method was widely used by Romans and Greeks too and was continued for centuries to be replaced by more mechanical devices during the 19th century and by the technique of echo sounding at the beginning of the 21st century.
Vikings (c.a 793-1066) are a particularly interesting example of navigators who travelled major distances using natural resources such as the distance from land that birds could fly, the smell of trees, the sound of waves breaking the shore of a nearby land and others. They were also claimed to be the first to use the sunstone, a mineral stone which could be used to find the location of the sun even when it was obscured from view (during a foggy day for example). The sunstone was widely used by Medieval explorers (13th to 16th century):
During the Middle Ages, and consequently, the age of the explorers, the growth of trade, the creation of the first maps, the portolan charts, as well as the development of navigational tools such as the magnetic compass, the sextant and the astrolabe led to a series of great discoveries which marked the history of navigation. Such were the discovery of America by Christopher Columbus (1451 - 1506), the first circumvention of the earth by Ferdinand Magellan (1480 - 1521), the travel from Europe to India by Vasco da Gama (1460 - 1524) and others.
The voyages of Christopher Columbus

The circumvention of the earth by Magellan
From the 1700s and on, the discovery of the marine chronometer which led to the proper determination of longitude, the placing of lighthouses and buoys provided precision and greater safety in the sea travel. In the 20th century, the use of digital technology, satellites and finally the GPS shaped the way that modern navigation is conducted.
A modern ship bridge
The teacher can refer to a brief account of the navigation techniques and terminology to enrich their discussion.
Attached items
Geo concept: navigational instrument
Uploads: bridge.jpg, magellan.jpg, columbus.jpg, sunstone_2.jpg, sunstone.jpg, echo.jpg, antikythera.jpg, astrolabe.jpg
1.2 Define questions from current knowledge
Description:
Ask the students how they would plan a short voyage if they lived in medieval times.What would they have to take into account if they wanted to travel via sea?
What tools would they need? Have the students assumed that the earth is not mapped yet? They set out at a journey to discover an unknown land like the old explorers Vikings did. How would they proceed to find their location? How would they find out if they are close to the shore?
2. Active Investigation
2.1 Propose preliminary explanations or hypotheses
Description:
The teacher introduces the students to the terms of latitude and longitude:
Geographic coordinates using the latitude/ longitude system are measured in terms of degrees. The reference point for all measurements of longitude is the meridian passing through Greenwich, England; this meridian is called the "prime meridian", and is represented by 0 degrees. The meridian of longitude that passes through any position on Earth is described in terms of how many degrees that meridian is to the east or west of the prime meridian. The maximum in either directionis 180 degrees. Parallels of latitude are measured in terms of how many degrees a given parallel is north or south from the equator (which is assigned latitude of 0 degrees).
Then, the teacher refers to the Mercator projection which is the method employed to depict the 3D curved area of earth in a 2D map.
Meridians and parallels are depicted as straight lines in the Mercator projection. Vertical linesrepresent the meridians and straight lines the parallels. A straight line connecting two locations of a Mercator map corresponds to the direction of a compass between these two points and the direction of navigation between them.

Source: OpenNLP (2010)

Figure 6. Annotation result.

4.2 Use of enriched scenarios for improving semantic search.

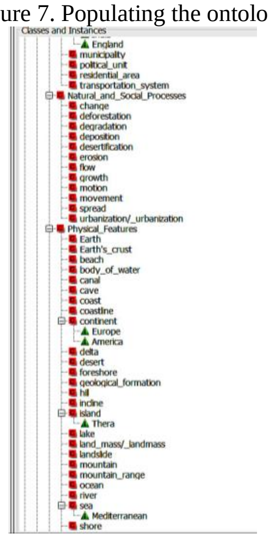
The enrichment process ends with an xml file of the annotated scenario. This is then uploaded on the platform and forms an edited (new version) of the processed scenario.

The difference is quite obvious once the semantic networks in each version are examined. Figures 8 and 9 show the semantic networks of the two scenario versions (before and after the semantic enrichment process).

A great advantage is the creation on the platform of an “image” of the existing scenario that has not altered the original content (no content was added or deleted from the original lesson plan), but was only semantically enriched in terms of concepts and instances (Figure 10 shows the instances added in the new version for map-based search).

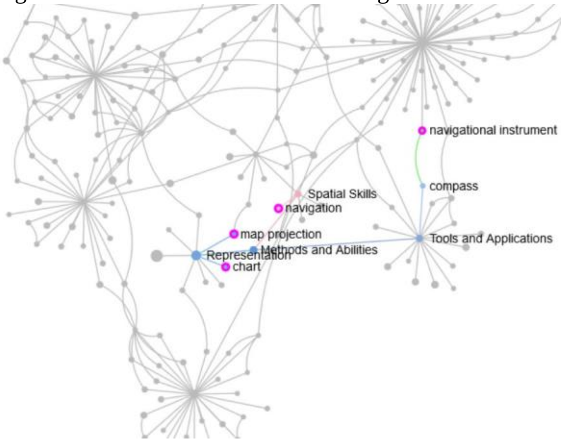
The enriched version results in a scenario that can contribute to keyword-, graph-, and map-based search capabilities of the platform and allow users to browse more efficiently its content.

Figure 7. Populating the ontology.



Classes and Instances
- England
- municipality
- political_unit
- residential_area
- transportation_system
- Natural_and_Social_Processes
  - change
  - deforestation
  - degradation
  - deposition
  - desertification
  - erosion
  - flow
  - growth
  - motion
  - movement
  - spread
  - urbanization/_urbanization
- Physical_Features
  - Earth
  - Earth's_crust
  - beach
  - body_of_water
  - canal
  - cave
  - coast
  - coastline
  - continent
    - Europe
    - America
  - delta
  - desert
  - foreshore
  - geological_formation
  - hill
  - incline
  - island
    - Thera
  - lake
  - land_mass/_landmass
  - landslide
  - mountain
  - mountain_range
  - ocean
  - river
  - sea
    - Mediterranean
  - shore

Figure 8. Semantic network of the user-generated scenario.



5 Conclusions

Semantic enrichment is currently an important feature in content retrieval, big data management, semantic web technologies etc.

Available tools for text-based semantic annotation are based on their own underlying ontologies and usually identify concepts and named entities for specific categories such as location, person, time, etc. The work presented in this paper attempted a different approach for the semantic enrichment of unstructured texts based on an elaborate ontology of spatial concepts and their between semantic relations. The first results show that there is great potential for the semantic annotation of user-generated educational resources in order to add semantic metadata to characterize and emphasize the meaning of educational scenarios, improve semantic search and discovery of content, and build connections among similar or related content.

Figure 9. Semantic network of the enriched scenario.

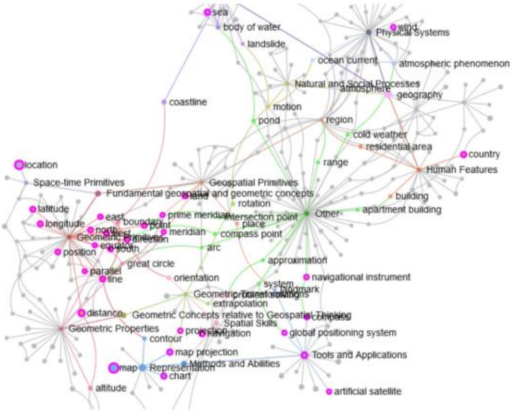
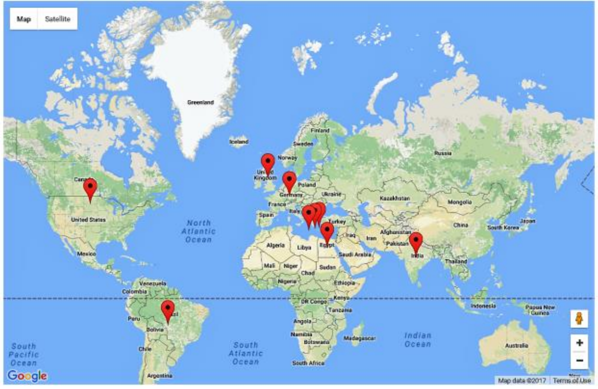


Figure 10. Instances (named entities) put the scenario on the map for map-based search and discovery.



6 References

Alves, A., Antunes, B., Pereira, F. C., and Bento, C. (2009) Semantic enrichment of places: Ontology learning from web. In: *Proceedings of the International Journal of Knowledge-based and Intelligent Engineering Systems*, 13(1), pp. 19-30, 2009

Butuc, M.G (2009). Semantically Enriching Content Using OpenCalais. *EDITIA*, 9, 77-88.

Cunningham. H et al. (2017), *Developing Language Processing Components with GATE Version 8 (a User Guide)*. The University of Sheffield, Department of Computer Science 2001-2017 [Online] Available from https://gate.ac.uk/userguide [Accessed 11th March 2017].

Daiber, J., Jakob, M., C., Hokamp, Mendes, N. P. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. In: *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics).*

Ennis, A., Nugent, C., Morrow, P., Chen, L., Ioannidis, G., Stan, A., and Rachev, P. (2015). A Geospatial Semantic Enrichment and Query Service for Geotagged Photographs. *Sensors* (Basel, Switzerland), 15(7), 17470–17482. http://doi.org/10.3390/s150717470

Fileto, R., Maya, C., Renso, C., Pelekis, N., Klein, D., Theodoridis, Y. (2015). The Baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, 98, 104–122.

Hu, Y., Janowicz, K., Prasad, S., Gao, S. (2015). Enabling Semantic Search and Knowledge Discovery for ArcGIS Online: A Linked-Data-Driven Approach. Fernando Bacao, Maribel Yasmina Santos, Marco Painho (Eds.), Lecture Notes in Geoinformation and Cartography. In: *Proceedings of AGILE 20015 “Geographic Information Science as an Enabler of Smarter Cities and Communities*”, Springer, pp 107-124.

Nixon, L., Bauer, M., and Bara, C. (2013). Connected media experiences: web based interactive video using linked data. In: *Proceedings of the 22nd International Conference on World Wide Web* (WWW '13 Companion). ACM, New York, NY, USA, 309-312. DOI: http://dx.doi.org/10.1145/2487788.2487931

Open NLP, 2010 The Apache Software Foundation, http://opennlp.apache.org/ [Accessed 12th March 2017].

Pernelle, N. (2016). Semantic enrichment of data: annotation and data linking. Artificial Intelligence [cs.AI]. Universite Paris Sud, https://tel.archives-ouvertes.fr/tel-01475250/document [Accessed 12th March 2017].

Romero, S. A. P. and Becker, K. (2016) Experiments with Semantic Enrichment for Event Classification in Tweets. In: *Proceedings of the* 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, 2016, pp. 503-506.

Tardy C., Falquet, G., Moccozet L. (2016) Semantic enrichment of places with VGI sources: a knowledge based approach. In: *Proceedings of the 10th Workshop* *on Geographic Information Retrieval*, Burlingame, California — October 31 - 31, 2016

WordNet, A Lexical Database for English. Princeton University, 2010, http://wordnet.princeton.edu/Wordnet [Accessed 12th March 2017].