Linked Open Piracy

Willem R. van Hage
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands
W.R.van.Hage@vu.nl

Véronique Malaisé
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands
vmalaise@few.vu.nl

Marieke van Erp
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands
marieke@cs.vu.nl

ABSTRACT

There is an abundance of semi-structured reports on events being written and made available on the World Wide Web on a daily basis. These reports are primarily meant for human use. A recent movement is the addition of RDF metadata to make automatic processing by computers easier. A fine example of this movement is the Open Government Data initiative which, by adding RDF metadata to spreadsheets and textual reports, strives to speed up the creation of geographical mashups and visual analytics applications. In this paper we present a new Open Linked Data RDF dataset[1] and a method for automatically adding such RDF metadata to semi-structured reports. We showcase our method on piracy attack reports issued on the web by the International Chamber of Commerce's International Maritime Bureau (ICC-CCS IMB)[2] We create a Semantic Web representation with the Simple Event Model (SEM) from screen scrapes of the ICC-CCS website. We show how the event layer makes it possible to easily analyze and visualize the aggregated reports to answer domain questions. Our pipeline includes conversion of the reports to RDF, linking their parts to external resources from the Linked Open Data cloud and exposing them to the Web through a ClioPatria web server that hosts the RDF.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms

Experimentation, Design

1. INTRODUCTION

In 2008, the increase of piracy attacks in the Gulf of Aden made the publication and analysis of events happening at sea around the world a new priority. The ICC-CCS gathers the reports related to piracy broadcasted by ships around the world, and publishes them daily on their website[2]. The reports are semi-structured, and concern seven (predefined) types of events: Hijacked, Boarded, Robbed, Attempted, Fired Upon, Suspicious (vessel spotted) and Kidnapped. The reports contains a field for the vessel type of the ship broadcasting the report; although the types of the vessels are often recurring, this field is filled manually, which gives rise to spelling variations (firedupon vs fired upon, tanker vs tankership) and a lack of certainty in terms of coverage: a new ship type could be filled in any day. The description of the event itself is done in full text, without a specific formatting except that it is preceded, in the same field, by the geographic and temporal coordinates of the event described. The geographic and temporal coordinates are repeated in an independent field each.

2. SCREEN SCRAPING

We start crawling of the ICC-CCS IMB webpage with the links to the yearly archives in the menu of the Live Piracy Map page. For each of these pages we follow all the links in the descriptions of the placemarks on the overview map. These are injected into the DOM tree with Javascript at runtime. We fetch them from the Javascript by parsing the Javascript with SWI-Prolog grammar (DCG) rules. This gives us a collection of semi-structured description pages, one for each event. We fetch the various fields from these pages using XPath queries and Prolog rules for value conversion and fixing irregularities. The code can be found online[3]. In this way we fetch: (1) The IMB's attack number; (2) The date of the attack, which we convert to ISO 8601 format; (3) The vessel type, which we map to URIs with rules that normalize a few spelling variations of the types. (4) The location label; (5) The attack type, which we map to URIs in the same way as the vessel type; (6) The incident details, which we convert to a RDFS comment describing the event itself. The first line is split into a time and place indication. These are used as backup sources to derive the date and location, should the parsing of fields nr. 2, 4 and 7 fail; (7) The longitude and latitude of the placemark on the map insert. These are used as coordinates of a generated anonymous place (i.e., without a URI) for the event.

For some of the events there are no explicit coordinates of the location of the event, but there is a textual description, for example, "approximately 150NM northwest of Port Victoria, Seychelles". For these event we look up the coordinates of Port Victoria using the GeoNames search web service[4], which returns RDF. From this location we calculate the coordinates using trigonometry. For example, in the case of 150NM northwest we compute the coordinates 150 minutes of angle at a bearing of 315 degrees.

We use the set of 7 values (numbered 1 to 7) extracted per report to generate a semantic event description using the Simple Event Model (SEM) [3] as illustrated in Figure 1.

[1]LOP, `http://semanticweb.cs.vu.nl/poseidon/ns/`, `http://ckan.net/package/linked-open-piracy`
[2]IMB, `http://www.icc-ccs.org/home/imb/`

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.
ACM 978-1-4503-0396-5/11/06.

[3]`http://www.few.vu.nl/~wrvhage/2011/lop/`
[4]GeoNames search, `http://sws.geonames.org/search`

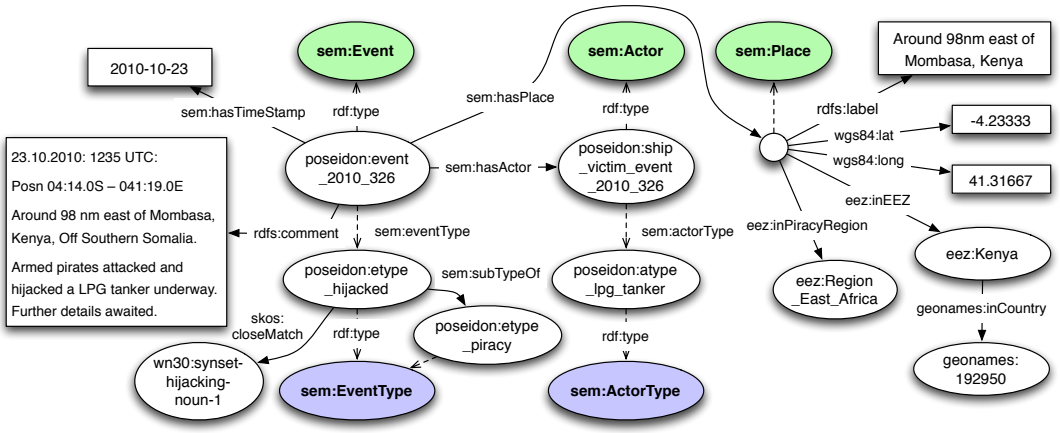

Figure 1: The complete RDF graph of a piracy report modeled in SEM including mappings to types in WordNet 3.0, a VLIZ exclusive economic zone, its corresponding GeoNames country, and its Piracy Region (see Section 3).

3. MAPPINGS

We create local URIs to represent the types of the extracted events and the types of their participants (e.g., poseidon:etype_hijacked or poseidon:atype_lpg_tanker). The shorthand for the name space of the local URIs is poseidon because the LOP data set was created during the Poseidon[5] project.

The SEM piracy events are aligned with vocabularies in the Linked Open Data cloud: WordNet 2.0[6], 3.0[7], OpenCyc[8] and Freebase[9]. Since there are only 73 sem:ActorTypes and 26 sem:EventTypes we manually created the following mappings: 70 skos:closeMatch (24 to Freebase, 24 to OpenCyc, 25 to WordNet);[10] 10 skos:broadMatch (5 to OpenCyc, 4 to WordNet, 1 to Freebase); 33 skos:relatedMatch (13 to OpenCyc, 11 to WordNet, 9 to Freebase). A "related" relation hold for example between WordNet's *to fire* and the event type *fired upon*, because *to fire* only conveys part of the meaning.

To classify each event by its place we need a classification of space. We chose to use the official geopolitical borders of the world, defined by the exclusive economic zones (EEZ, usually defined as 200 nautical miles from the coast of the nearest state). We classified all event places according to whether they are **in** or **nearest to** we need a specification of the borders of these zones. We take these from the World EEZ version 5 data set from the VLIZ Maritime Boundaries Geodatabase[11]. We make a more general partitioning of the world into regions (e.g. Gulf of Aden, Carribean) following the EEZs (using Prolog `space_intersects/3` queries on the EEZ shapes). The remaining surface of the earth, including the international waters and inland seas is partitioned based on the nearest EEZ (using Prolog `space_nearest/3` queries on the EEZ shapes). The resulting sections of the world are grouped together to form the more general domain specific partitioning of the world consisting of what we call "Piracy Regions".

[5]Poseidon `http://www.esi.nl/poseidon/`
[6]WordNet 2.0, `http://www.w3.org/2006/03/wn/wn20/`
[7]WordNet 3.0, `http://semanticweb.cs.vu.nl/lod/wn30/`
[8]OpenCyc, `http://sw.opencyc.org/`
[9]Freebase, `http://{www|rdf}.freebase.com/`
[10]We use closeMatch to represent the slight mismatch between the definitions of the concepts in SEM and the 3 target vocabularies.
[11]VLIZ, `http://www.vliz.be/vmdcdata/marbound/`

4. HOSTING THE PIRACY DATA

The entire ICC-CCS data set is hosted as Linked Data on a ClioPatria server[1]. All URIs in the data set are resolvable. A SPARQL endpoint is available at `http://semanticweb.cs.vu.nl/poseidon/ns/user/query`.

5. CONCLUSION

We present a new Open Linked Data set about worldwide maritime piracy events crawled from the web. In essence, this work is an Open Government Data project, like data.gov [1] and data.gov.uk [2], with the exception that data are intergovernmental. Our goal is the same, to reduce the cost to answer, possibly complex, domain questions on integrated government data. LOP allows easier processing of piracy reports for visualization and the computation of statistics.

6. ACKNOWLEDGEMENTS

This work has been carried out as a part of the Poseidon project and the Agora project. Work in the Poseidon project was done in cooperation with Thales Nederland, under the responsibilities of the Embedded Systems Institute (ESI). The Poseidon project is partially supported by the Dutch Ministry of Economic Affairs under the BSIK03021 program. The Agora project is funded by NWO in the CATCH programme, grant 640.004.801. We would like to thank Davide Ceolin, Juan Manuel Coleto, and Vincent Osinga for their significant contributions. We thank the ICC-CCS IMB and the NGA for providing the open piracy reports.

7. ADDITIONAL AUTHORS

Additional authors: Guus Schreiber (VU University Amsterdam, email: `guus.schreiber@vu.nl`).

8. REFERENCES

[1] D. D. Li Ding, D. L. McGuinness, J. Hendler, and S. Magidson. The data-gov wiki: A semantic web portal for linked government data. In *Proceedings of the 6th International Conference on Knowledge Capture*, 2009.

[2] T. Omitola et al. Put in your postcode, out comes the data: A case study. In *The Semantic Web: Research and Applications*, volume 6088 of *Lecture Notes in Computer Science*, pages 318–332. Springer Berlin / Heidelberg, 2010.

[3] W. R. van Hage, V. Malaisé, R. Segers, L. Hollink, and G. Schreiber. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics*, 2011.