Linked Open Piracy

Willem R. van Hage[1], Véronique Malaisé[2], and Marieke van Erp[1]

[1] Department of Computer Science, VU University Amsterdam
{W.R.van.Hage,Marieke.van.Erp}@vu.nl
[2] Elsevier Content Enrichment Center (CEC)
v.malaise@elsevier.com

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. In this paper we present a new linked data set and a method for automatically adding such RDF metadata to semi-structured reports to speed up the creation of geographical mashups and visual analytics applications. We showcase our method on piracy attack reports issued by the International Chamber of Commerce (ICC-CCS). We show how the semantic representation 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.

1 Introduction

In this paper we present a new data set on the Web of Data, Linked Open Piracy (LOP), how it was constructed, and how it can be used to answer complex questions about piracy. We expose descriptions of piracy attacks at sea published on the Web by the International Chamber of Commerce's International Maritime Bureau (ICC-CCS IMB)[3] and the US National Geospatial-Intelligence Agency (NGA)[4] as Linked Data RDF[5].

LOP can be seen as an Open Government Data[6] initiative for intergovernmental data. The goal of Open Government Data is to reduce the time to do analytics and mashups with open government data. The piracy reports are, like most open government data, published in a human readable format[7]. We show how we can reduce the commonly acknowledged bottleneck of data preprocessing time in the workflow from question to answer. This format and type of publication (following a given pattern for a year of publication, daily update of the webpage) makes it an ideal test case for automatic RDF event extraction; the

[3] http://www.icc-ccs.org/home/imb

[4] NGA, http://www.nga.mil/portal/site/maritime/

[5] LOP, http://semanticweb.cs.vu.nl/lop

[6] http://data-gov.tw.rpi.edu/wiki/Open_Government_Data

[7] A notable exception is data.gov.uk where the data are exposed directly as machine friendly RDF.

topic of the reports is also of contemporary socio-economic concern and are related to research questions that go beyond what classic data mining can easily answer. We therefore chose to take this example as a showcase for the feasibility and usability of event extraction coupled with novel research question answering methods.

We represent LOP data in RDF with the Simple Event Model (SEM) [7] and demonstrate that an event model is not only an intuitive way of representing (inter)governmental data, but also a powerful tool for data integration. We evaluate the usefulness of SEM as a model for Open Government Data by answering complex domain questions derived from authorities in the domain of piracy analysis, namely UNITAR UNOSAT and the ICC-CCS IMB. We use SWI-Prolog[8] to extract event descriptions from the web, represent them in SEM and store them in a ClioPatria RDF repository [10] extended with the SWI-Prolog space package [8] for spatial and temporal indexing. The entire ICC-CCS data set is hosted as Linked Data, all URIs in the data set are resolvable. A SPARQL endpoint is available at `http://semanticweb.cs.vu.nl/lop/sparql/`.

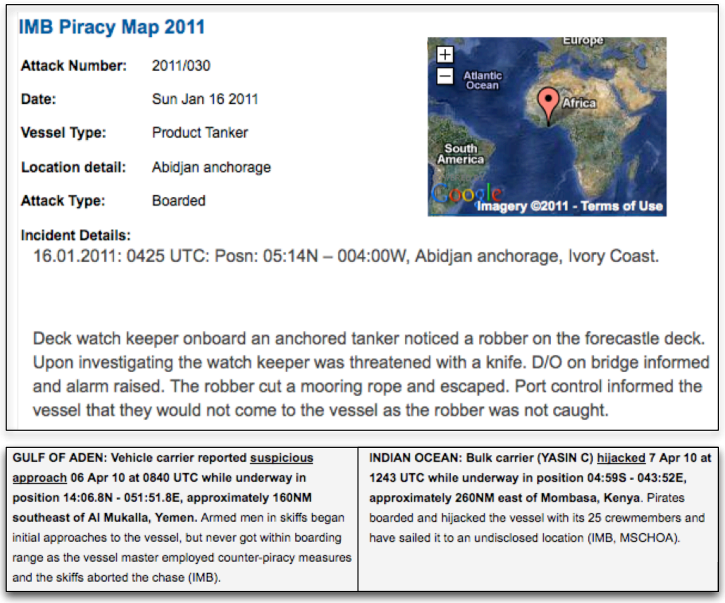
This paper is organized as follows. In Section 2, we show how we created RDF event descriptions from web pages. In Section 3, we discuss the modeling of the events in SEM. In Section 4, we show example domain questions from UNOSAT that can easily be answered using our event representation. In Section 5, we discuss related work and in Section 6, we conclude with a discussion and future plans.

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. Figure 1 (top) shows what an ICC-CCS piracy report looks like. 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 (e.g., firedupon vs fired upon) 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. The geographic and temporal coordinates are repeated in an independent field each.

For each of these pages we follow all the links in the descriptions of the placemarks on the overview map, returning us one semi-structured description pages for each event. We fetch the various fields from these pages using XPath queries and Prolog rules for value conversion and fixing irregularities. In this way we fetch: (1) The IMB's attack number, which consists of the year and a

[8] SWI-Prolog, `http://www.swi-prolog.org/`



IMB Piracy Map 2011

Attack Number: 2011/030

Date: Sun Jan 16 2011

Vessel Type: Product Tanker

Location detail: Abidjan anchorage

Attack Type: Boarded

Incident Details:

16.01.2011: 0425 UTC: Posn: 05:14N – 004:00W, Abidjan anchorage, Ivory Coast.

Deck watch keeper onboard an anchored tanker noticed a robber on the forecastle deck. Upon investigating the watch keeper was threatened with a knife. D/O on bridge informed and alarm raised. The robber cut a mooring rope and escaped. Port control informed the vessel that they would not come to the vessel as the robber was not caught.

GULF OF ADEN: Vehicle carrier reported suspicious approach 06 Apr 10 at 0840 UTC while underway in position 14:06.8N - 051:51.8E, approximately 160NM southeast of Al Mukalla, Yemen. Armed men in skiffs began initial approaches to the vessel, but never got within boarding range as the vessel master employed counter-piracy measures and the skiffs aborted the chase (IMB).

INDIAN OCEAN: Bulk carrier (YASIN C) hijacked 7 Apr 10 at 1243 UTC while underway in position 04:59S - 043:52E, approximately 260NM east of Mombasa, Kenya. Pirates boarded and hijacked the vessel with its 25 crewmembers and have sailed it to an undisclosed location (IMB, MSCHOA).

Fig. 1. Example of an IMB piracy report (top) and two NGA piracy reports (bottom)

counter. From this we generate an event identifier by prepending a namespace and by appending a suffix whenever there are duplicate attack numbers in a year; (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 detail, which we use as a label for the place of the event; (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 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 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. The time fetched from the date (3) or narrative (6) field has a number of different representations in the source pages. Some time indications are in local time, while others are in UTC. Often there is no indication of the time zone. For many events the indicated time is 00:00 (midnight) to denote the time of attack is unknown. These inconsistencies in the time notation, in combination with the fact that there are few events on the same day, led us to the decision to use the date without a time indication whenever there is ambiguity about the time.

To demonstrate that representing extracted events in SEM aids the integration of data sources, we take another set of piracy reports and integrate these with the IMB reports. For this, we use the Worldwide Threat to Shipping reports by the US National Geospatial-Intelligence Agency describing 36 piracy events between 26 March 2010 and 16 April 2010. 31 of these events overlap with the IMB reports. The remaining 5 come from other sources: Reuters (2)[9], UKMTO[10], MSCHOA[11], and ReCAAP[12]. These reports are (re)posted on many websites, some of which are plain-text representations of the reports, while others add some additional layout tags to separate the place, time, and state of the ship during the attack from the narrative. Two example NGA reports are shown in Figure 1 (bottom).

By changing the XPath and grammar rules to suit the different structure of the NGA reports we were able to recognize the same 7 attributes we got from the IMB website. The event terminology is nearly the same as on the IMB website, except there is a distinction between boardings and robberies. There is also some extra information in 34 of the 36 reports about the state of the ship during the attack, (e.g., moored or underway). 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 events we look up the coordinates of Port Victoria using GeoNames[13], which returns RDF. From this location we use trigonometry along the geoid with the haversine formula in the specified direction. For example, in the case of 150NM northwest we compute the coordinates 150 minutes of angle at a bearing of 315 degrees. We treated time in the NGA reports in the same way as in the IMB reports, reducing them to an ISO 8061 date.

We match the NGA reports to the IMB reports by picking the nearest event that occurred on the same day that has compatible actor types, i.e., when the types are not the same, one has to be sem:subTypeOf the other. This enables us to automatically map 30 of the 31 overlapping reports correctly. We store these matches with an owl:sameAs property between the two matching events. We believe the single unmatched report was mistakingly identified as a distinct IMB report, because it is extremely similar to another report (the same date, place, time, victim vessel type, and similar narrative) which has a matching IMB report. Therefore, we believe there should only have been 30 overlapping reports, which we were all able to match.

3 Event Representation in SEM

We use the set of 7 elements (see Section 2) extracted per report to generate a semantic event description using SEM. We generate a URI for the event described in each report and a URI for the victim ship, which we represent as a sem:Actor, based on the IMB attack number (nr. 1). The date (nr. 2) is attached to the

[9] Reuters, `http://www.reuters.com/`

[10] UK Maritime Trade Operations,`http://www.mschoa.org/Links/Pages/UKMTO.aspx`

[11] The Maritime Security Center – Horn of Africa, `http://www.mschoa.org/`

[12] The Regional Cooperation Agreement on Combating Piracy and Armed Robbery against Ships in Asia, `http://www.recaap.org/`

[13] GeoNames search, `http://sws.geonames.org/search`

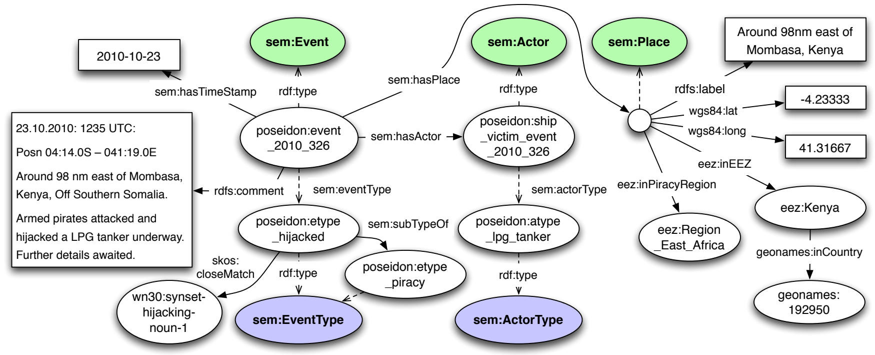


Fig. 2. The complete RDF graph of a piracy report modeled in SEM including mappings to types in WordNet 3.0, a VLIZ exclusive economic zone, its corresponding GeoNames country, and its Piracy Region.

sem:Event by means of the sem:hasTimeStamp property. The sem:hasTimeStamp datatype property was chosen over the sem:hasTime object property, because we do not need type hierarchies over time instances to answer our domain questions. The vessel type (nr. 3) is typed as a sem:ActorType attached to the victim ship sem:Actor with the sem:actorType property, a subproperty of rdf:type. The location detail (nr. 4) is made a rdfs:label of the blank node representing the location of the attack. We chose not to use the Exclusive Economic Zones (EEZs)[14] (usually defined as 200 nautical miles from the coast of the nearest state), or the GeoNames identifier of the nearest relevant place, as the URI of the location of the attack because this would have removed the distinction between the exact location of the attack and the more general region. We did use the EEZs for an initial partitioning of the world into regions (e.g. Gulf of Aden, Carribean). The remaining surface of the earth, including the international waters and inland seas is partitioned based on the nearest EEZ. The area nearest to an EEZ is assigned a new URI, e.g., the international waters off the coast of Liberia and closest to Liberia's EEZ (i.e., not closest to Ascension's, Côte d'Ivoire, Sierra Leone's, or Saint Helena's EEZs) is assigned the URI eez:Nearest_to_Liberia. Based on the distribution of the piracy events, we grouped particular sections of the world together. This grouping is only specific to the piracy event domain.

The attack type (nr. 5) is modeled analogously to the vessel type as a sem:EventType, which is attached to the event using the sem:eventType property. The event type *robbery* that we found in the NGA set was modeled as a sem:subTypeOf the IMB event type *boarding*. The *mooring* and *underway* vessel states are modeled as additional event types of the piracy event using sem:eventType properties attached to the event. All event types used in this data set are sem:subTypeOf the piracy event type, poseidon:etype_piracy. The narrative of the report (nr. 6) is attached to the event as a rdfs:comment. The WGS84

[14] http://www.vliz.be/vmdcdata/marbound/

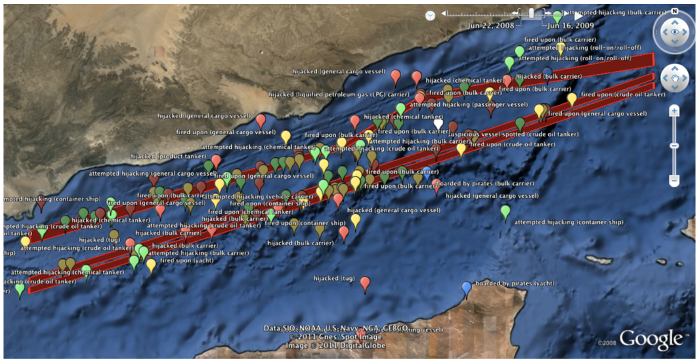

attempted hijacking (bulk carrier)
Jun 22, 2008
Jun 16, 2009
fired upon (bulk carrier)
attempted hijacking (roll-on/roll-off)
attempted hijacking (roll-on/roll-off)
hijacked (general cargo vessel)
hijacked (chemical tanker)
hijacked (bulk carrier)
hijacked (liquified petroleum gas (LPG) carrier)
fired upon (bulk carrier)
fired upon (bulk carrier)
fired upon (crude oil tanker)
attempted hijacking (passenger vessel)
hijacked (general cargo vessel)
hijacked (chemical tanker)
fired upon (general cargo vessel)
fired upon (general cargo vessel)
fired upon (bulk carrier)
fired upon (bulk carrier)
suspicious vessel spotted (crude oil tanker)
attempted hijacking (chemical tanker)
attempted hijacking (bulk carrier)
attempted hijacking (crude oil tanker)
fired upon (crude oil tanker)
hijacked (product tanker)
attempted hijacking (general cargo vessel)
fired upon (general cargo vessel)
fired upon (bulk carrier)
fired upon (bulk carrier)
boarded by pirates (bulk carrier)
attempted hijacking (container ship)
fired upon (bulk carrier)
attempted hijacking (vehicle carrier)
hijacked (general cargo vessel)
fired upon (general cargo vessel)
fired upon (container ship)
fired upon (chemical tanker)
hijacked (general cargo vessel)
attempted hijacking (container ship)
hijacked (chemical tanker)
hijacked (bulk carrier)
fired upon (container ship)
attempted hijacking (crude oil tanker)
attempted hijacking (crude oil tanker)
hijacked (bulk carrier)
hijacked (tug)
hijacked (bulk carrier)
attempted hijacking (crude oil tanker)
attempted hijacking (bulk carrier)
fired upon (yacht)
attempted hijacking (chemical tanker)
hijacked (tug)
boarded by pirates (yacht)
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2011 Cnes/Spot Image
Image © 2011 DigitalGlobe
©2008 Google

Fig. 3. Attacks plotted in Google Earth.

coordinates (nr. 7) are assigned to the blank node with the W3C WGS84 vocabulary. Additional ship names are attached to the sem:Actor using the ais:name property, a domain-specific label for ship names.

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_yacht). The SEM piracy events are aligned with WordNet 2.0[15], 3.0[16], OpenCyc[17] and Freebase[18]. WordNet gives us the advantage of relating different lexical variations to a unique URI e.g., mapping *highjacking* and *hijacking* to *hijacking*. This can also be used to automatically transform piracy descriptions to types. As WordNet has a hierarchy of hyponym relations between synsets (e.g., a *tankership* is a hyponym of *cargoship*) we can do hyponym inference.

We can not map all of our types to any one of these three vocabularies, but by mapping to all three of them we get a good coverage of our domain-specific type vocabulary. Our data set contains 73 ActorTypes and 26 EventTypes, which is too few to make it worthwhile to use an automatic mapping method, so we manually created the following mappings: 70 skos:closeMatch (24 to Freebase, 24 to OpenCyc, 25 to WordNet);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.

4 Answering Domain Questions

In this section, we show how the SEM representation simplifies answering domain questions through visualizations and analyses. We first show how the enriched

[15] WordNet 2.0, `http://www.w3.org/2006/03/wn/wn20/`
[16] WordNet 3.0, `http://semanticweb.cs.vu.nl/lod/wn30/`
[17] OpenCyc, `http://sw.opencyc.org/`
[18] Freebase, `http://{www|rdf}.freebase.com/`

data could be used to recreate UNOSAT questions. Then we show the added value of the mappings and hierarchies in an additional set of domain questions.

4.1 Rebuilding UNOSAT Reports

The analysis performed and compiled for the UNOSAT reports [5] have mostly been carried out manually and sometimes with the aid of a GIS. The analyses are thorough and insightful, but do require painstaking manual sifting through the data because only the unprocessed attack reports are used. Human researchers then plot these data on maps, and assign attack types to them. With the RDF version and the mappings to the VLIZ economic zones and geospatial reasoning the analyses that require a combination of data sources can be sped up immensely. SPARQL and Prolog rules make many complex questions as simple as a graph query.

The conclusion of map 1 in the UNOSAT 2009 Q1 report, namely that the attacks have shifted southward and extended further east-west along the axis of the International Recommended Transit Corridor (IRTC)[19] can be reproduced by combining plotting the attacks on a map along with information about the IRTC. This is illustrated in Figure 3, a time animation in KML is available online[20]. Although more coastguard and marine vessels are present in the recommended corridor, pirates also know that there are more ships there, hence more chances of finding a victim.

4.2 Additional Questions

We start with an easy visualization of number of attacks per region per year (top left Figure 4). We can see that the most active regions are the Gulf of Aden, Indonesia, India and East Africa. The graph also shows that Indonesia used to be the most active region, but sometime in 2007 activity in the Gulf of Aden and East Africa have become the regions with most piracy activity.

Although the narrative section of each report are not split up and represented in RDF yet, we can give some ideas on differences in weapon use by comparing the number of occurrences of the terms “guns” and “knives” in the different reports. For instance, there are no reports that mention knives in the Gulf of Aden region at all, while there are 109 in the Indonesia region while there are 85 that mention guns in the Gulf of Aden and only 25 in Indonesia. The pie charts in Figure 4 show an overview of five weapons types. In order to properly analyse these we will use more sophisticated NLP techniques in future work.

If we further look into the four most active areas, we can use the ship type mapping to compare differences in ships attacked in different regions. The stacked bar chart in Figure 4 immediately highlights the difference between Indonesia and the other areas, namely that in the Indonesia region far more tugs

[19] http://www.icc-ccs.org/news/163-coalition-warships-set-up-maritime-security-patrol-area-in-the-gulf-of-aden

[20] http://semanticweb.cs.vu.nl/poseidon/piracy_reports_2005-2010.kmz

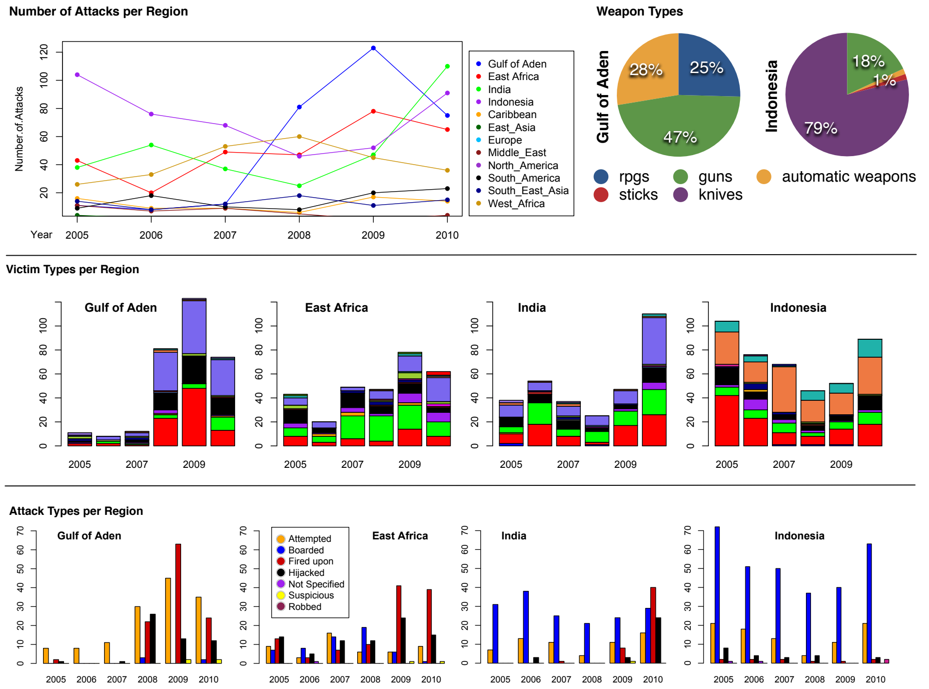


Fig. 4. Number of attacks reported per region per year, weapon types per region, victim types per region and attack types per region.

are attacked than in the other regions. In the Gulf of Aden, for a larger number of attacks the ship type of the victim is not known. Interestingly, the attacks on bulk carriers has been declining in the Asian regions until 2009, whereas it was on the rise in the African regions. In order to explain this, extra information is needed, for example on the number of ship movements in these areas. Unfortunately, such data is not openly available.

We can also split out the attacks by types of attack to see whether pirates take a different approach in different regions. Plotting these statistics in a graph, split out per region, has the advantage that one can quickly see the differences, whereas plotting these on a map still requires interpretation from the user. Here, the region clustering shows its merit. In the last series of charts in Figure 4, one can see that significant differences exist between the regions in the types of attacks. In Asia, for example, far more often ships are boarded (which often also means robbed) than in the African regions. In the Gulf of Aden attacks have become more aggressive and more often victim ships are fired upon. In the Gulf of Aden, also more attempted hijackings occur than elsewhere.

5 Related Work

This work essentially describes an Open Government Data project, like data.gov [2] and data.gov.uk [3], with the exception that data are intergovernmental. The case we present deals with scraping event description from web pages. In the past we have done similar work with different types of data sources, such as user ratings of museum pieces [9], historical events [6], and Automatic Identification System NMEA ship data for the recognition of ship behavior from trajectories and background knowledge from the Web [11]. This is accomplished with the SWI-Prolog space package [8], which is similar to Franz Inc.'s Common Lisp-based AllegroGraph system[21]. We use SEM to describe our events, because it is a simple but not spartan model. A very similar model is LODE, which has been used for the extraction of events from Wikipedia timelines [4]. Both SEM and LODE focus on the *"Who does what, where and when?"*, but LODE does not contain a typing system, whereas SEM does. An example of a much richer event model is part of the CIDOC-CRM. The purpose of CIDOC-CRM is the integration of meta data about (museum) artifacts. A description of an integration method that, like the work presented in this paper, also combines space, time and semantics, using CIDOC-CRM can be found in [1]. The SEM specification[22] contains mappings to LODE and CIDOC-CRM.

6 Conclusions and Future Work

We have shown that the ideas behind the Open Government Data initiative can also be applied to information sources from intergovernmental organizations without the need for changing their entire information workflow. Automatic conversion of online open data can bring their data to the Web and help these organizations with their business by making it easier to answer questions about their data. In this case study, the representation we use is the Simple Event Model, which helps to integrate spatio-temporal reasoning with web semantics. SEM has an appropriate level of abstraction for the integration of piracy event data: it is more general than the differences between the data sources taken into account in this paper, but still specific enough to answer domain-specific questions. This modularity of the flexible event extraction set allows us to combine data sources with relatively little change in the code base. We have shown that different data sources provide different aspects of an event, and their combination allows for interesting and serendipitous data analysis. As future work, we aim at doing further natural language processing on each report's content description in plain text in order to extract more information: the types of weapons used during the attack, the number of pirate boats and pirates, the intervention of a coalition war ship or helicopter, the outcome of the attack which would help to answer even more domain questions. Also, we would like to investigate the possibility to interlink the Linked Open Piracy data set with news items on the

[21] `http://www.franz.com/agraph/allegrograph/`

[22] SEM, `http://semanticweb.cs.vu.nl/2009/11/sem/`

World Wide Web. This would provide additional background information to the semantic event descriptions, but also a semantic description of the news articles on the Web.

7 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.

References

1. G. Hiebel, K. Hanke, and I. Hayek. Methodology for CIDOC CRM based data integration with spatial data. In *38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology*,
2. 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 *8th International Semantic Web Conference (ISWC 2009)*, 2009.
3. T. Omitola, C. Koumenides, I. Popov, Y. Yang, M. Salvadores, M. Szomszor, T. Berners-Lee, N. Gibbins, W. Hall, M. C. Schraefel, and N. Shadbolt. Put in your postcode, out comes the data: A case study. In *7th Extended Semantic Web Conference (ESWC 2010)*, 2010.
4. R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In *4th Annual Asian Semantic Web Conference (ASWC'09)*,
5. UNOSAT. Analysis of somali pirate activity in 2009. `http://unosat-maps.web.cern.ch/unosat-maps/SO/Piracy/2009/UNOSAT_Somalia_Pirates_Analysis_Q1_2009_23April09_v1.pdf`, April 2009.
6. M. van Erp, J. Oomen, R. Segers, C. van den Akker, L. Aroyo, G. Jacobs, S. Legêne, L. van der Meij, J. van Ossenbruggen, and G. Schreiber. Automatic heritage metadata enrichment with historic events. In *Museums and the Web 2011*, 2011.
7. 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*, 9(2):128–136, July 2011.
8. W. R. van Hage, J. Wielemaker, and G. Schreiber. The space package: Tight integration between space and semantics. *Transactions in GIS*, 14(2), 2010.
9. Y. Wang. *Semantically-Enhanced Recommendations in Cultural Heritage.* PhD thesis, Technische Universiteit Eindhoven, 2011.
10. J. Wielemaker, Z. Huang, and L. van der Meij. *SWI-Prolog and the web*, volume Theory Theory and Practice of Logic Programming. Cambridge, pages 363–392. Cambridge University Press, 2008.
11. N. Willems, W. R. van Hage, G. de Vries, J. Janssens, and V. Malaisé. An integrated approach for visual analysis of a multi-source moving objects knowledge base. *International Journal of Geographical Information Science*, 24(9):1–16, Sept. 2010.