J Data Semant (2012) 1:187–201
DOI 10.1007/s13740-012-0009-6

ORIGINAL ARTICLE

Linked Open Piracy: A Story about e-Science, Linked Data, and Statistics

Willem Robert van Hage · Marieke van Erp · Véronique Malaisé

Received: 2 February 2012 / Revised: 15 May 2012 / Accepted: 20 June 2012 / Published online: 19 July 2012

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 representing data from spreadsheets and textual reports in RDF, strives to speed up the creation of geographical mashups and visual analytic applications. In this paper, we present a newly linked dataset and the method we used to automatically translate semi-structured reports on the Web to an RDF event model. We demonstrate how the semantic representation layer makes it possible to easily analyze and visualize the aggregated reports to answer domain questions through a SPARQL client for the R statistical programming language. We showcase our method on piracy attack reports issued by the International Chamber of Commerce (ICC-CCS). 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.

Keywords Information extraction · Metadata enrichment · Linked data

W. R. van Hage · M. van Erp (✉)
Department of Computer Science, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
e-mail: marieke@cs.vu.nl

W. R. van Hage
e-mail: W.R.van.Hage@vu.nl

V. Malaisé
Elsevier Content Enrichment Center (CEC), Radarweg 29, 1043 NX Amsterdam, The Netherlands
e-mail: v.malaise@elsevier.com

1 Introduction

Governmental and commercial organisations collect a wealth of information; from census to trade data and from pollution to crime. Too often, making sense of these data is a time consuming undertaking as most data are stored in many spreadsheets or textual reports. Recent initiatives, such as the Open Government Data initiative,[1] have shown the added benefit of using Semantic Web technologies to unlock the potential of such data.

In this article, we first present a new dataset on the Web of Data, linked open piracy (LOP) describing maritime piracy events and detail its construction. Then we present an approach and tool for analyzing these types of data and show how these can be used to answer complex questions about the domain. We expose descriptions of piracy attacks at sea published to the Web by the International Chamber of Commerce's International Maritime Bureau (ICC-CCS IMB) and the US National Geospatial-Intelligence Agency (NGA)[2] as Linked Data RDF.[3]

Linked open piracy can be seen as an open government data 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, similar to most open government data that is for example processed into http://www.data.gov, published in a human readable format.[4] We show how, by converting the IMB piracy reports to RDF and linking them to LOD cloud resources,

[1] Open Government Data, http://www.data-gov.tw.rpi.edu/wiki/The_Data-gov_Wiki.

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

[3] LOP, http://www.semanticweb.cs.vu.nl/poseidon/ns/.

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

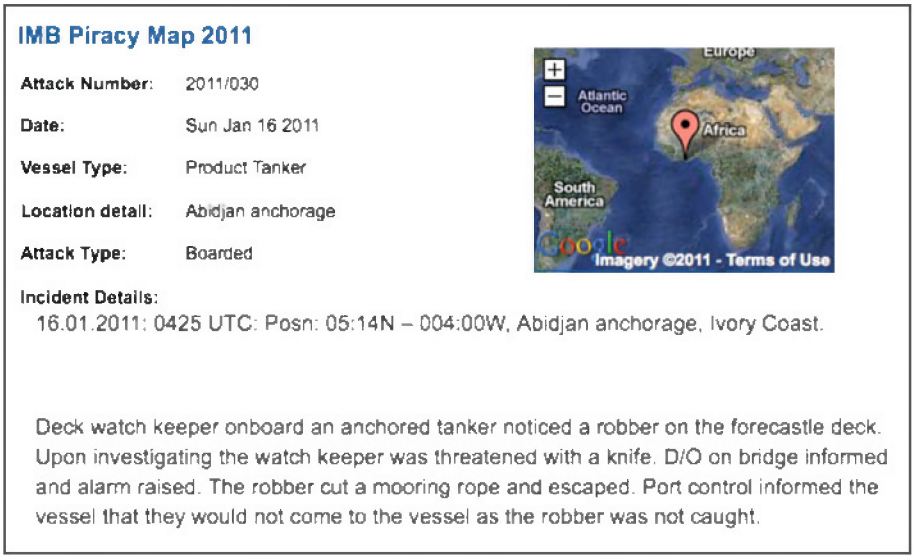

Fig. 1 Example of an IMB piracy report

we reduce the commonly acknowledged bottleneck of data preprocessing time in the workflow from question to answer.

The format and type of publication of the IMB piracy reports (following a given pattern for year of publication, daily updated to the web page) make it an ideal test case for automatic RDF event extraction; the topic of the reports is also of contemporary socio-economic concern [3] and is 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.

As the main structure for our representation of LOP in RDF, we chose the simple event model (SEM) [24] 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 use SWI-Prolog[5] to extract event descriptions from the web, represent them in SEM and store them in a ClioPatria RDF repository [27]. The SWI-Prolog space package [25] is used for spatial and temporal indexing. The added benefit of using SEM as a model for Open Government Data is evaluated by answering complex domain questions derived from authorities in the domain of piracy analysis, UNITAR UNOSAT and the ICC-CCS IMB. To perform the analysis and evaluation, we utilize the SPARQL package[6] for R[7], which bridges the gap between RDF and statistical data processing.

The remainder of the paper is organised as follows. In Sect. 2, we describe the IMB and NGA reports on the Web. In Sect. 3, we show the event extraction method we used to create RDF event descriptions from web pages. In Sect. 4, we discuss the modelling of the events in SEM. In Sect. 5, we extend the event models with extra properties about weapon use extracted automatically from the textual narratives included in the event reports. In Sect. 6, we show how the LOD dataset can be accessed online. In Sect. 7, we show how we process the event descriptions in the R statistical programming language and we evaluate which domain questions from the IMB and UNOSAT can be answered using our event representation in SEM and which additional results we achieved as corollaries. In Sect. 8, we discuss related projects, methods, and event models. In Sect. 9, we conclude with a discussion of our findings and a summary of our future work.

2 Maritime Piracy Reports on the Web

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.[8] The reports are semi-structured, and concern seven (predefined) types of events: hijacked, boarded, robbed, attempted, fired upon, suspicious (vessel spotted) and kidnapped. An example report is shown in Fig. 1. The reports contain 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., 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 written up full text, without a specific format-

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

[6] SPARQL R, http://www.cran.r-project.org/package=SPARQL.

[7] R, http://www.r-project.org/.

[8] IMB, http://www.icc-ccs.org/home/imb.

Table 1 Number of reports from 2005 to 2011.

2005	2006	2007	2008	2009	2010	2011
276	237	260	293	395	458	434

ting except that it is often 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.

The number of reported incident has risen steadily since the ICC-CCS started collecting incident reports in 2005. The number of reports for each year is shown in Table 1.

3 Collecting Piracy Reports

In this section, we first detail how the piracy reports were collected from the ICC-CCS IMB Website, followed by an example of how this approach can easily be adopted to collect piracy reports from another source. A copy of the code discussed in this section can be found online at http://www.few.vu.nl/~wrvhage/LOP/LOP_code_JoDS.zip.

3.1 ICC-CCS IMB Website

We start crawling of the ICC-CCS IMB web page 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 place marks 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 Prolog grammar 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. In this way, we fetch: (1) The IMB's report number, which consists of the year and a 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 nr. 2, 4 and 7 fail; (7) The longitude and latitude of the place mark on the map insert. These are used as coordinates of a generated anonymous place (i.e., without a URI) for the event.

Over the years the layout of the IMB reports has changed, so to get the same field we use a number of different XPath expressions. For example, to get the narrative field we can use:

```
//div[contains(@id,"narrations")]/p/
text().
```

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. We have seen examples where the indicated time without time zone has to be local time and cases where it has to be UTC. 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.

3.2 NGA WTS Reports

To demonstrate that the representation of extracted events in SEM aids the integration of data sources we take another set of piracy reports and try to integrate these with the IMB reports.

Our example set comes from the Worldwide Threat to Shipping reports by the US National Geospatial-Intelligence Agency (NGA).[9] Two example reports are shown in Fig. 2. We take a set of reports describing 36 piracy events between the 26th of march 2010 to the 16th of april 2010. 31 of these events overlap with the IMB reports. The remaining five come from other sources: Reuters (2),[10] UK Maritime Trade Operations (UKMTO),[11] The Maritime Security Center—Horn of Africa (MSCHOA),[12] and The Regional Cooperation Agreement on Combating Piracy and Armed Robbery against Ships in Asia (ReCAAP).[13]

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. By changing the XPath and grammar rules to suit the different structure of the WTS reports, we were able to recognize the same seven attributes we got from the IMB website. The event

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

[10] Reuters, http://www.reuters.com/.

[11] UKMTO, http://www.mschoa.org/Links/Pages/UKMTO.aspx.

[12] The Maritime Security Center—Horn of Africa, http://www.mschoa.org/.

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


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).

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).

Fig. 2 Example of two NGA piracy reports

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, whether it was moored or underway. Sometimes the NGA reports also mention the name of the ship. 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 the GeoNames search web service[14]. From this location, we perform trigonometry along the geoid with the haversine formula in the specified direction. For example, in the case of 150 NM northwest we compute the coordinates 150 min of angle at a bearing of 315° degrees. The same problems with time indications apply to the NGA set as to the IMB set so we treated time in the same way, reducing it to an ISO 8601 date.

We match the NGA reports to the reports extracted from the IMB site by picking the nearest event that occurred on the same day that has compatible actor types. By compatible we mean exact equivalence of types or a sem:subTypeOf relation. This way, we were able 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.

4 Event Representation

We use the set of seven report elements (numbered 1–7 in Sect. 3) extracted per report to generate a semantic event description using the simple event model (SEM) [24]. A graphical example of a SEM event description is given in Fig. 3. We first generate a URI for the event described in the report and a URI for the victim ship that is based on the IMB attack number (nr. 1). The victim ship is represented as a sem:Actor. The date (nr. 2) is attached to the sem:Event by means of the sem:hasTimeStamp property. The sem:hasTimeStamp datatype property is 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 an rdfs:label of the blank node representing the location of the attack. In our representation, we chose to represent the exact location of the attack and to not use the exclusive economic zones (EEZs)[15] (usually defined as 200 nautical miles from the coast of the nearest state), or the GeoNames identifier of the nearest relevant place, to represent the location of the attack. The reason is that this would have removed the distinction between the exact location of the attack and the more general region, resulting in the assignment of the same place to 18 events when using EEZs or over 600 events when using GeoNames identifiers. For certain types of analyses it is handy to have EEZs or GeoNames identifiers for the events, but we chose to arrange this through mappings (see Sect. 4.1). The attack type (nr. 5) is modeled analogously to the vessel type as a sem:EventType, that is attached to the event using the sem:eventType property.

The event type *robbery* that we found in the NGA set was modelled as a sem:subTypeOf the IMB event type *boarding*. The *mooring* and *underway* vessel states are modelled as additional event types of the piracy event using extra sem:eventType properties attached to the event. All event types in this dataset are sem:subTypeOf the piracy event type, poseidon:etype_piracy.

sem:subTypeOf is a subproperty of rdfs:subClassOf, which enables us to use RDFS to select any set of attacks we are interested in. The narrative of the report (nr. 6) is attached to the event as a rdfs:comment. The WGS84 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.

[14] GeoNames search, http://www.sws.geonames.org/search.

[15] http://www.vliz.be/vmcdata/marbound/.

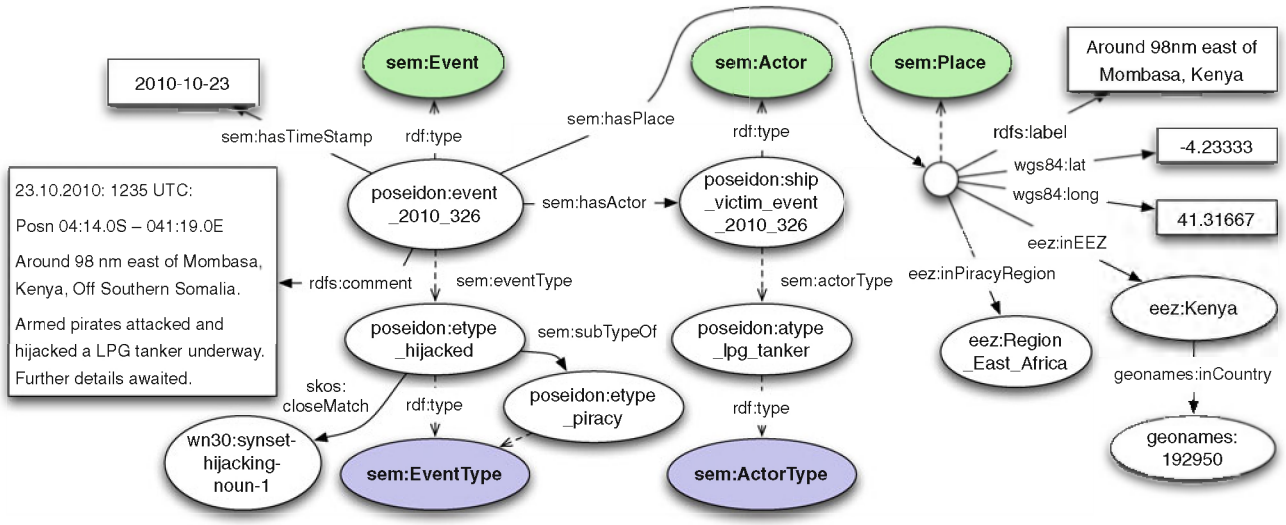

Fig. 3 The complete RDF graph of a piracy report modeled in SEM including mappings to types in WordNet 3.0, a VLIZ exclusive economic zone, its corresponding GeoNames country, and its piracy region (see Sect. 4.1)

4.1 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).[16]

The SEM piracy events are aligned with the following vocabularies in the Linked Open Data cloud: WordNet 2.0,[17] 3.0,[18] OpenCyc[19] and Freebase[20]. Even with the ICC-CCS's semi-structured format, there is still some variation in the values, because the fields are filled in manually (e.g., the term *hijacking* can be spelled *highjacking* or *hijacking*). WordNet can help us here to relate different lexical variations to a unique URI. We use this to automatically transform piracy descriptions to types. WordNet also has a hierarchy of hyponym relations between synsets (e.g., a *tankership* is a hyponym of *cargoship*), which enables us to do hyponym inference.

We cannot map all of our types to any one of these three vocabularies, but by mapping to all three of them we obtain a good coverage of our domain-specific type vocabulary. As our dataset only contains 73 ActorTypes and 26 EventTypes, it is not worthwhile to set up an automatic mapping method, so we manually created the following mappings: 70 skos:closeMatch (24 to Freebase, 24 to OpenCyc, 25 to WordNet);[21] 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 holds for example between WordNet's *to fire* and the event type *fired upon*, because *to fire* only conveys part of the meaning.

As mentioned earlier in this section, it may be useful to classify each event by its place. For this, we need a classification of space. We chose to use the official geopolitical borders of the world, defined by the exclusive economic zones (EEZs). We classified all event places according to whether they are *in* or *nearest to* an EEZ. We take the specification of the borders of these zones from the World EEZ version 5 datasets from the VLIZ Maritime Boundaries Geodatabase[22]. This dataset contains all EEZs of the world in KML format. We use the SWI-Prolog space package [25] to extract the shapes and their descriptions from the KML file and to perform containment and nearest-neighbor queries for all sem:Places of the events and all the EEZs. 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 area nearest to an EEZ is assigned a new URI. For instance, the area of 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.

For the piracy domain, we make an additional, more general, partitioning of the world into regions. This partitioning is based on the distribution of the piracy events (e.g.,

[16] The shorthand for the name space of the local URIs is poseidon because the LOP dataset was created during the Poseidon project (http://www.esi.nl/poseidon/).

[17] WordNet 2.0, http://www.w3.org/2006/03/wn/wn20/.

[18] WordNet 3.0, http://www.semanticweb.cs.vu.nl/lod/wn30/.

[19] OpenCyc, http://www.sw.opencyc.org/.

[20] Freebase, http://wwwlrdf.freebase.com/.

[21] We use closeMatch to represent the slight mismatch between the definitions of the concepts in SEM and the 3 target vocabularies.

[22] VLIZ, http://www.vliz.be/vmdcdata/marbound/.

Gulf of Aden, Carribean) and follows the EEZs (using Prolog `space_intersects/3` queries on the EEZ shapes). This grouping is domain specific and specific to the task of showing developments in piracy events.

5 Narrative Analysis

Although the SEM piracy event descriptions already provide a rich source for report analysis, there is still a treasure of information contained in the unstructured event narratives. These snippets of text that are included in the piracy reports from 2007 on contain for example information about the weapons that were used, the number of attackers, possible outcomes of the attack and whether the victim received any assistance. There is a great variety in the length and types of information that is given in the narratives, as can be seen from examples below:

poseidon:event_2010_008:

"tank stripping operations. Robbers escaped with ships stores. Pilot and port control informed."

poseidon:event_2011_140:

"22.03.2011: 2200 LT: Posn: 02:45.22N 104:24.29E, Off Tioman island, Malaysia.A group of more than 10 pirates armed with long knives in a speed boat boarded a tug towing a barge enroute from Singapore to Koh Kong, Cambodia. They took hostage the 10 crewmembers, locked them in a cabin, cut of the tracking system on the tug and hijacked the vessel. On 24.03.2011, they released the crew in a life raft and gave them some food, water, their passports and some money. By then, the tug boat had been repainted to a green colour. On 26.03.2011, a passing-by fishing boat rescued the crewmembers and landed them at Natuna Island and the crew managed to contact the owners. All relevant authorities in the region informed to lookout for the hijacked tug and barge."

The narrative sections contain a large variety of information types such as weapon types, actions of the victims, actions of the attackers, number of attackers and what type of vessel the attackers used, most of them expressed in running text. This makes a field segmenting task much harder than for example the task of segmenting addresses [2]. Closest to our dataset is a field segmentation task carried out for collection reports for specimens in the natural history domain [5,13,22]. However, those reports contain fewer free text information types (only 2 vs. 9 for the pirates), as well as many shorter fields that are easier to recognise. As deep linguistic analysis of the narratives is out of the scope of this contribution, we only detail the information extraction experiments we carried out in order to retrieve the weapon type used by the attackers.

5.1 Data Preparation

Following the distribution of the events from 2007 to 2011, we annotated 200 event instance narratives with weapon information. Due to the increase in the number of attacks, the breakdown of the events is as follows: 14 % from 2007, 16 % from 2008, 22 % from 2009, 24 % from 2010 and 24 % from 2011. We chose to take time into account, as we saw from the event types that the nature of the attacks has changed, which we suspect may also influence the weapon type. Within the years, the events were chosen randomly.

In the selected events, the following weapon classes are encountered: knives, guns, it automatic weapons, knives and guns, catapults, knives and hacksaws, automatic weapons and RPGs, rockets and guns. Furthermore, we also encounter instances where no weapon type is mentioned *armed*, or no weapons are mentioned at all *no mention*.

As the weapon types are often expressed using similar words, we chose to use a vector space approach using a modified bag of words to represent the comment sections. Our modified bag of words consists of a combination of noun phrases (unigrams, bigrams and the occasional trigram) as well as adjectives. The data are first tokenised using a simple symbol-driven tokenizer implemented in Perl, after which the noun phrases are selected. The noun phrases that are found are for example *armed pirates* and *armed security team*. This helps to discern between weapons used by attackers and by the victims better than single word features. The final data preparation step consists of stemming using the Porter Stemming Algorithm [17].

5.2 Results

We used WEKA version 3.6.2 [9] to perform our initial feature selection, bringing down the number of features from 1,142 (4 features derived from the structured data, namely, year, type of attack, eez, and ship type, 1,026 noun phrase features, and 12 adjective features) to 48 (4 structured features, 60 noun phrase features, and 4 adjective features). We then use the WEKA implementation of the RIPPER algorithm [6] to construct an initial set of rules to classify the weapon type. This set of rules gives us an F measure of 76.8 % on the 200 annotated examples in a tenfold cross-validation experiment. In Table 2, the results per weapon class are presented.

Even though the classification is not perfect, the results are actually more useful than the precision, recall and F measure would indicate. This is because the classes form a hierarchical structure where in some cases it is not so bad if the classifier makes a mistake, for example when the classifier mistakes a firearm for a gun. We chose not to merge the firearms and gun classes, as guns are only a subset of firearms, but in many cases they will be guns. We see similar examples with steel rod (of which there is only one example in our training set, and our classifier will classify that instance as 'armed'). As the RIPPER algorithm does not assign multiple classes to an instance, it also has proportionally more trouble with the 'mixed' weapons instances than with the single

Table 2 Results of weapon classification using RIPPER on 68 features

Class	#	Precision	Recall	F_1
Knives	29	0.893	0.862	0.877
Steel rod	1	0	0	0
Catapults, knives and hacksaws	1	0	0	0
Guns	11	0.6	0.818	0.692
Guns and knives	5	0.6	0.6	0.6
Firearms	23	0.818	0.783	0.8
RPGs	1	0	0	0
Automatic weapons	11	0.778	0.636	0.7
Automatic weapons + RPGs	13	0.733	0.846	0.786
Guns and RPGs	5	0	0	0
Unspecified	22	0.632	0.545	0.585
No mention of weapons	78	0.845	0.91	0.877
Overall	200	0.761	0.78	0.768

weapon instances. It, for example, classifies two instances of class knives and guns as just knives, and does this also for the instance of class catapults, knives and hacksaws.

In the LOP data, the weapon type used in the attack is represented by a separate lop:attackerWeapon property that is attached to the event. The representations of the weapon type for events 2010_326 and 2011_261 are for example given as:

poseidon:event_2010_326 lop:attackerWeapon poseidon:wtype_armed.
poseidon:event_2011_261 lop:attackerWeapon poseidon:wtype_gun, poseidon:wtype_knife.

In future work, we will look at deeper natural language processing techniques to also detect other information types from the narratives, as well to improve on the current weapons classification results.

5.3 Weapons Analysis

Although the weapons classification is not perfect, it can already give an indication of different weapon use in different regions. For this analysis, we have aggregated all non-fire arms (knives, steel rods, catapults and hacksaws) into 'Melee Weapons', all light firearms (guns and firearms) into 'Firearms' and all heavy firearms (automatic weapons, RPGs) into 'Automatic Firearms'. We have plotted the results for four piracy hotspots, namely, the Gulf of Aden, East Africa, the India Bengal zone and Indonesia and show the results in Fig. 4. These charts show that in the attacks in the Gulf of Aden and East Africa firearms are much more popular than melee weapons, whereas the opposite is true for Indonesia. In the India Bengal zone, the weapons distribution is fairly

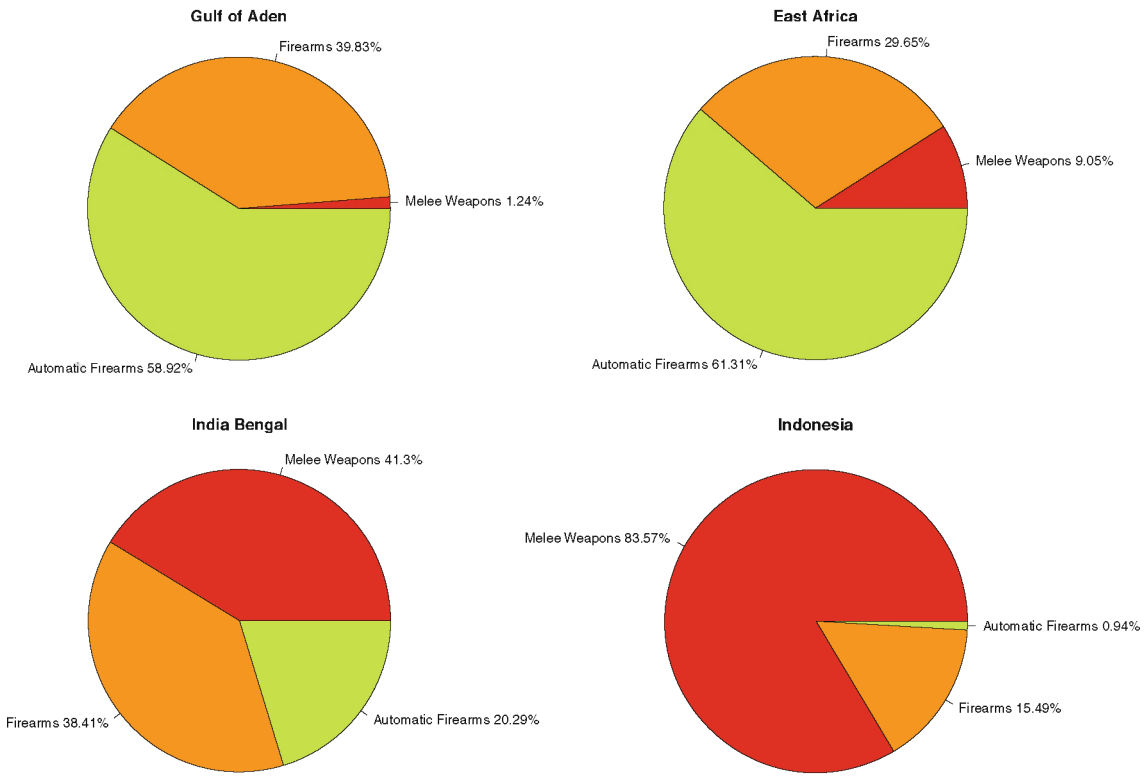

Fig. 4 Breakdown of aggregated weapon types for Gulf of Aden, East Africa, India Bengal and Indonesia. The Gulf of Aden is a war zone compared to Indonesia

equal. This type of information can be useful to estimate what type of counter-piracy measures to apply in what region.

6 Hosting the Piracy Data

The entire ICC-CCS dataset, as described in the previous sections, is hosted as Linked Data on a ClioPatria server. All URIs in the dataset are resolvable. For example, the event with the URI poseidon:event_2010_326 (shown in Fig. 3) is found at: http://www.semanticweb.cs.vu.nl/poseidon/ns/instances/event_2010_326. The SPARQL end-point is available at: http://www.semanticweb.cs.vu.nl/lop/query. A KML rendering of the dataset can be found at http://www.few.vu.nl/~wrvhage/LOP/LOP.kmz. All event descriptions in the KML version have links to the original ICC-CCS webpages and the RDF version of the event.

7 Statistical Analysis

In this section, we show how the event representation makes it easy to answer domain questions through visualizations and analyses. We first demonstrate how we access the data from R, a language and environment for statistical computing and graphics (footnote 7), using the SPARQL package for R (footnote 6). Then we show how we apply these techniques to recreate UNOSAT and IMB reports (Sects. 7.2, 7.3). Then we show the added value of the mappings and hierarchies in an additional set of domain questions (Sect. 7.4).

7.1 The R SPARQL Package

The R language allows us to easily select, aggregate and visualize the event descriptions, and to perform statistical tests. These are exactly the tools that are needed to answer commonly asked questions about piracy events, such as "Has the intensity of attacks really increased in the Gulf of Aden in the past years?" or "Is there a difference in the types of attacks that occur in the Gulf of Aden and in the rest of the world?".

To make it possible to use R to process the Linked Open Piracy event descriptions, we use the SPARQL package for R developed together with Tomi Kauppinen. This package allows us to access SPARQL end points and pose SELECT or UPDATE queries. In this case, we use SELECT queries to gather tables that describe the various properties of the events. For example, if we want to show the attack intensity in the Gulf of Aden over time we will need the time and the region of the LOP events. Figure 5 shows the R code that accomplishes this. We first define where the SPARQL end point can be found by declaring the URL of the end point, http://www.semanticweb.cs.vu.nl/lop/sparql/. Then we specify the RDF

```
library(SPARQL)
library(zoo) # provides as.yearqtr

endpoint <- 'http://semanticweb.cs.vu.nl/lop/sparql/'

# find timestamps and regions of piracy events
query <-
  'SELECT ?region ?time
   WHERE { ?event sem:hasTimeStamp ?time .
           ?event sem:hasPlace ?place .
           ?place eez:inPiracyRegion ?region . }'

# define eez namespace to shorten URIs to QNames
ns <- c('eez',
        'http://semanticweb.cs.vu.nl/poseidon/ns/eez/')

# fire query at SPARQL end point
data <- SPARQL(url=endpoint,query=query,ns=ns)$results

# select the region Gulf of Aden
slice <- data[ data[['region']] ==
               'eez:Region_Gulf_of_Aden', ]

# count the events per quarter
counts <- table(as.yearqtr(as.Date(slice[['time']])))

# plot the results
plot(counts,type='b')
```

Fig. 5 R code that uses the SPARQL package to access the Linked Open Piracy dataset. This code produces the plot shown in Fig. 7

graph pattern that connects an event's time and region in a SELECT query and we fire that query at the end point. To shorten the URIs that we get back we can declare abbreviations for namespaces. The result of the SPARQL call is a data frame with a column for each variable in the SELECT query and a row for each instantiation of these variables. To count the events in the Gulf of Aden, we make a slice of the data frame that we retrieve from the SPARQL end point. This slice selects the rows of the data frame that have eez:Region_Gulf_of_Aden as a value in the region column. Then we determine the quarter of the year the event happened in by converting the time column to quarters, and aggregate the list of events to a table of counts. This table can be used for statistical analysis or visualization. A visualization of the counts is shown in Fig. 7.

In the rest of this section, we will apply this method of using SPARQL projections of RDF graphs into R tables for visual and statistical analysis to answer questions from piracy reports of the United Nations Institute for Training and Research (UNITAR) Operational Satellite Applications Programme (UNOSAT). Prior to the conversion of the IMB reports to Linked Open Data, the statistics in these reports were time-consumingly compiled manually. Having the data in a structured and queryable format can make this considerably simpler and more efficient, so the human researcher can spend more time on interpreting the results of the analyses.

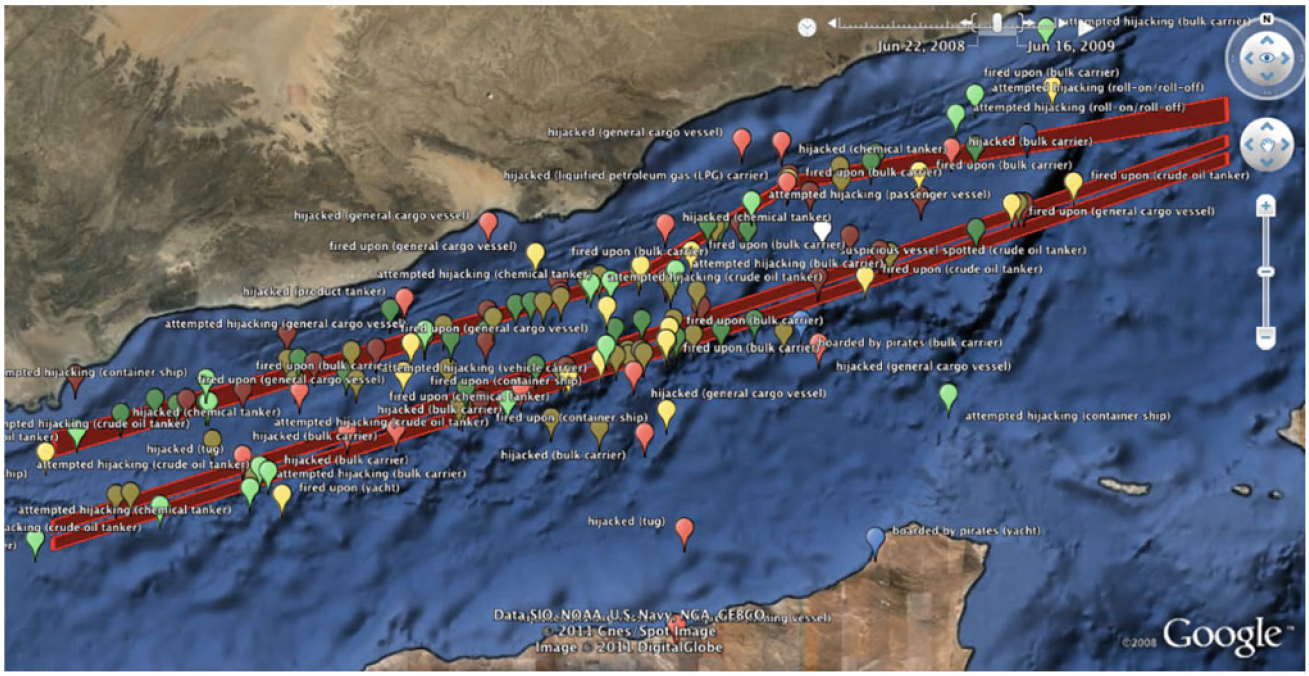

Fig. 6 Attacks plotted in Google Earth, along with shapes indicating the areas of the International Recommended Transit Corridor (IRTC) shown in *red*. The attacks follow the patrolled corridors. Pirates go where there are ships to attack. An animated version of this figure can be found online (footnote25) (colour figure online)

The code used to generate the plots shown in the rest of this article can be found online[23].

7.2 Rebuilding UNOSAT Reports

The analysis performed and compiled for the UNOSAT reports [21] are usually mostly 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 can make many complex questions as simple as a graph query. Having an RDF event model to work with makes selecting, extending, and correlating the data much easier than just having GIS map layers.

The conclusion of Section 2 in the UNOSAT 2009 report, namely, that the attacks have shifted southward and extended further east–west along the axis of the International Recommended Transit Corridor (IRTC) [24] can be reproduced by combining geographical information about the attacks with information about the (IRTC). This is illustrated in Fig. 6. A time animation in KML is available online.[25] Although more coastguard and marine vessels are present in the recommended corridor, pirates also know that there are more

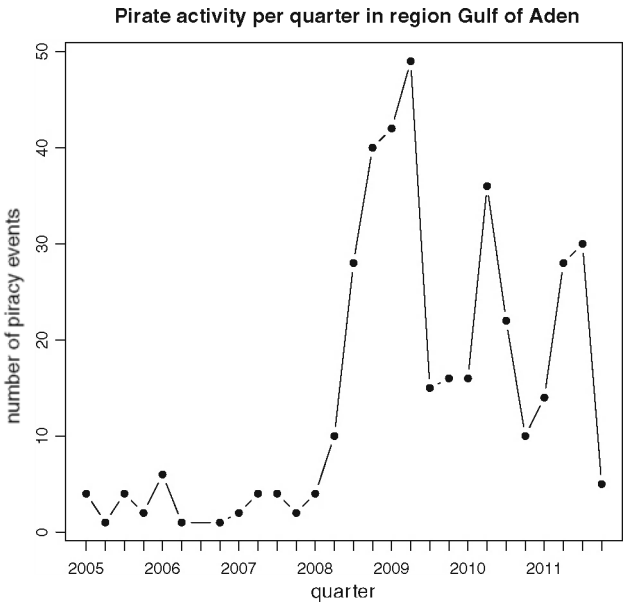

Fig. 7 Number of attacks reported in the Gulf of Aden 2005–2011, aggregated quarterly

[23] LOP R, http://www.few.vu.nl/~wrvhage/LOP/PlotLOP.R.

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

[25] LOP KML, http://www.few.vu.nl/~wrvhage/LOP/LOP.kmz.

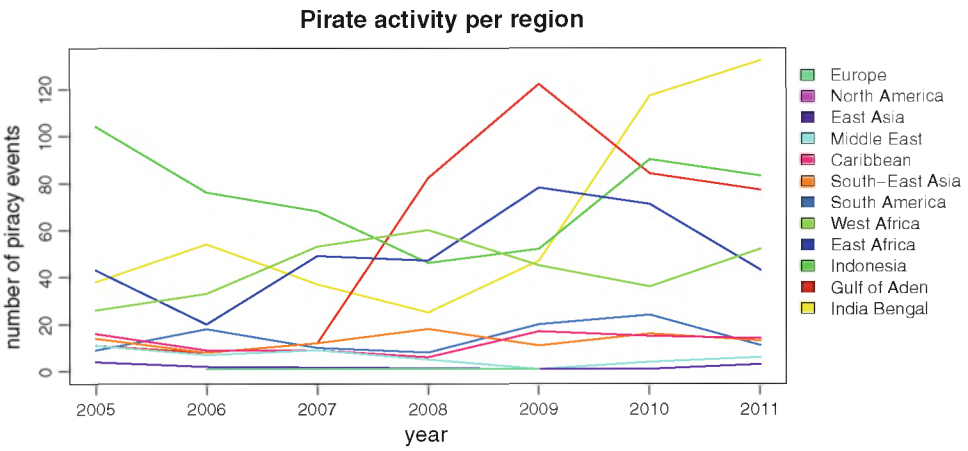

Fig. 8 Number of attacks reported per region per year. Notice the simultaneous decrease in the Gulf of Aden and increase in India and Indonesia. Figure 11 shows that changes in Aden and India are more likely to be related than Aden and Indonesia, considering the sudden change in the types of attacks in India. This spill over into the Arabian Sea, which falls in the LOP piracy region "India Bengal", after September 2009, is shown in Fig. 9

ships there, hence more chances of finding a victim. For a discussion about how to visualise and count the actual numbers of attacks in the vicinity of the patrolled corridors using the SWI-Prolog space package see [24].

Tables and graphs summarizing the number of successful hijackings, arrests and attacks, such as those on page 3 (summarizing the total number of reported attacks per region) and 5 (type of attacks per quarter) of the report, are simply pulled out of the data using a few queries and then some adding up in one's favorite statistics program such as R (footnote 7). For example, the hijackings can be found by querying for sem:eventType poseidon:etype_hijacked and the attempts that failed by querying for poseidon:etype_attempted instead. In Fig. 11, an overview of the counts of every event type is shown for the four most notorious piracy regions of the world, the Gulf of Aden, East Africa, India and the Bay of Bengal, and Indonesia. Comparing the red to the pink bars gives an indication of the success rate of pirates in these regions.

7.3 Visualizing IMB Highlights

The IMB piracy reporting centre regularly posts trends they detect in the piracy data on their website. In this section, we take a report from the IMB website and show how the information discussed in the report can be extracted from the processed piracy data.

On Wednesday 21st October, 2009 the IMB reports that there is an unprecedented increase in Somali pirate activity[26]. If we plot the number of attacks in the Gulf of Aden region (see Fig. 7) we see that the significant jump in the number of attacks already appeared in the third quarter of 2008 (significance computed using Welch's two sample t test, $p = 0.05$). We also see that after 2009, the number of attacks decreases in the Gulf of Aden (from Fig. 8 we can see that is shifts to the region around India). In Sect. 7.1, we will go into detail how to automatically produce the counts shown in Fig. 7.

7.4 Additional Questions

We start with an easy visualization of number of attacks per region per year (Fig. 8). From this figure, we can see that the most active regions are the Gulf of Aden, Indonesia, the India Bengal zone and East Africa. The Figure 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.

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. Figure 10 immediately highlights the difference between Indonesia and the other areas, namely, that in the Indonesia region far more tugs and other small vessels are attacked than in the other regions. Another difference that stands out is that there is such a big difference in the types of attacked ships between the Gulf of Aden and East Africa. In both regions big vessels make up the largest portion of the victims, but in East Africa these are mostly container ships, while in the Gulf of Aden many more tankers get attacked. In order to explain this, extra information is needed, for example on the number of ship movements of these types in these areas. There might simply be relatively more tankers in the Gulf of Aden region than before the coast of East Africa. Unfortunately, such data are 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 Fig. 11, 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)

[26] News about piracy boom, http://www.icc-ccs.org/news/376-unprecedented-increase-in-somali-pirate-activity.

Fig. 9 Attacks in the Arabian Sea after September 2009 plotted in Google Earth. They types of attacks are very similar to those in East Africa and the Gulf of Aden (violent attacks, shown in *red, yellow and green*), but not similar to the types of attacks previously known in the India and Bay of Bengal region (boardings with the purpose of robbery, shown in *blue*). An animated version of this figure can be found online (footnote 25) (colour figure online)

than in the African regions. In the Gulf of Aden attacks have become more aggressive and more often victim ships are fired upon. This also shows from the weapons analysis in Sect. 5.3. In the Gulf of Aden, also more attempted hijackings occur than elsewhere.

8 Related Work

Maritime piracy is a problem that is as old as maritime trade. However, in the past decade, the problem has exploded again starting with a growing number of attacks off the coast of Somalia since 2005 [1]. With the increase in attacks, and the incurred costs on trade, also the interest from the (research) community has grown to analyse the attacks and devise counter measures such as from a naval perspective [20], from socio-economics [18], and agent-based systems [11]. The work at hand provides the prerequisites for facilitating analyses from a variety of perspectives on the piracy attack data.

The past few years, a considerable body of work has been published on converting governmental data to Linked Open Data [8]. Essentially, the work at hand is also an Open Government Data project, similar to http://www.data.gov [14] and http://www.data.gov.uk [16], with the exception that these data are intergovernmental. Furthermore, our dataset was not previously published for access in an open government portal. The case we present deals with scraping event descriptions from Web pages.

All the event descriptions are represented as SEM events. We chose this model because it is a simple but expressive and flexible model. We have, for example, used it to represent user ratings of museum pieces [26], historical events [23], and automatic identification system (AIS) of NMEA ship data for the recognition of ship behavior from trajectories and background knowledge from the Web [28]. A very similar model is LODE, which has been used for the extraction of events from Wikipedia timelines [19]. 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 [7]. The purpose of CIDOC-CRM is the integration of metadata about (museum) artifacts. A description of an integration method that, as the work presented in this paper, also combines space, time and semantics, using CIDOC-CRM can be found in [10]. The SEM specification[27] contains mappings to LODE and CIDOC-CRM.

[27] SEM, http://www.semanticweb.cs.vu.nl/2009/11/sem/.

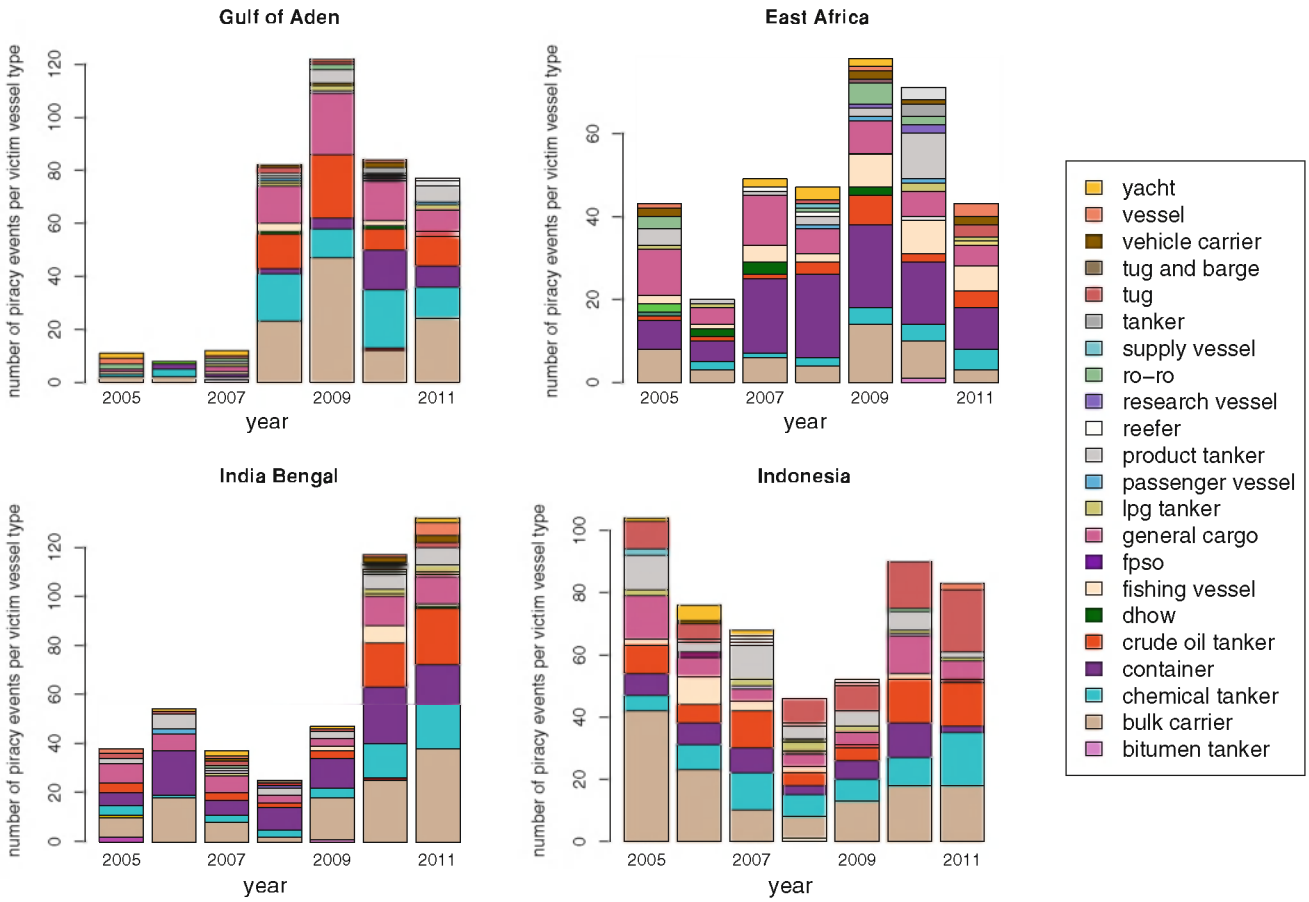

Fig. 10 Victim types per region for the Gulf of Aden, East Africa, India Bengal and Indonesia regions with the number of attacks on bulk carriers in *beige*, chemical tankers in *cyan*, container ships in *purple*, crude oil tankers in *red*, general cargo ships in *pink*, and tugs in *brown*. Attacks in Indonesia aim to rob any ship, also smaller ships in harbors, like tugs, while attacks near Somalia aim to hijack big ships, like tankers and container ships. For the sake of brevity, the legend only shows ship types with five or more occurrences in the dataset (colour figure online)

In Sect. 7.1, we discuss the SPARQL package for R. An example of the SPARQL package being adopted for a semantic, statistical, and visual analysis of Linked Data can be found at the Linked Science website [12].

9 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 to change their entire information workflow. Automatic conversion of online open data can bring them 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. The simple event model 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 (as well as those used in other cases, cf. [23,26,28]), but it is specific enough to answer the domain questions presented in Sect. 7. In Sect. 3, we show that the event extraction process is flexible and can be applied to new datasets by adapting the XPath query and the exception rules used for parsing the crawled webpages. The conversion process can also be applied to other types of datasets such as low-level GPS information (cf., [28]).

This modularity allows us to combine data sources with relatively little change in the code base. When the initial development of the IMB screen scraper was done, the adaptation to the set of NGA WTS reports could be done in an afternoon. We have shown that different data sources provide different aspects of an event, and their combination allows for interesting and serendipitous data analysis. When the dataset was ready to answer the UNOSAT and IMB domain questions discussed in Sects. 7.2 and 7.3, we got answers to the additional example domain questions in Sect. 7.4 for free. Statistical tests to compare the distribution of ship types, attack types, per time interval or region are easily done by

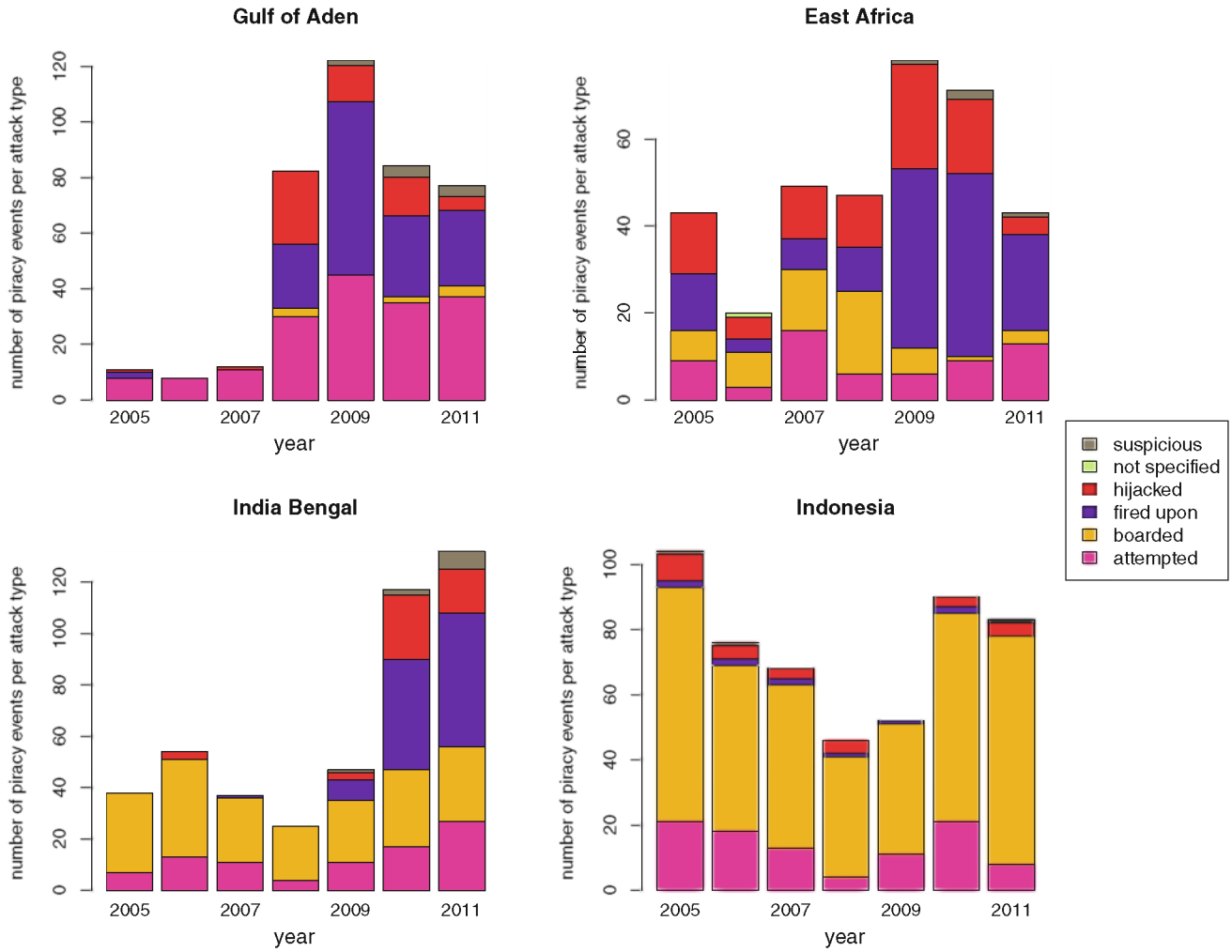

Fig. 11 Number of attacks per year sorted by attack type for Gulf of Aden, East Africa, India and Indonesia Regions. The attack types in the Gulf of Aden and East Africa differ significantly from the rest of the world (χ^2 test). In 2010, the India Region (the Arabian Sea) has become more similar to East Africa and Aden

importing the RDF data into the R statistical language with the SPARQL R package. We contributed to the Linked Open Data by new RDF datasets and their connection to existing parts of the LOD cloud. These links participate in weaving the LOD cloud, to enhance new research dimensions for future research questions to find answers.

Whereas most data resources seem unstructured, there are many report websites that would be very suited to scraping. For most programmers, putting together a scraper for a website such as ICC-CCS is a matter of days. One of the drawbacks of live scraping is that websites tend to change their formatting, which happened with the ICC-CCS database three times in the time frame covered by the scraped event descriptions (2005–2012). This can cause code that at first succeeded at scraping event descriptions to fail at scraping the same data, due to a change in presentation. Scraper code will have to be continuously maintained to keep working. Caching snapshots of the source website can help with the transition from one version of the source website to the next. A more robust solution to the entire migration to RDF would be to directly connect to a database using tools like D2RQ [4]. This way the translation code will not have to be changed when the HTML rendering of the data changes. However, database schema changes can still require a change in the translation code.

Real-world data are dirty. It is estimated that about 5 % of data entered manually are incorrect [15]. There were indeed data entry mistakes in the event descriptions. In total, there were three distinct incorrect spelling variants of event type identifiers (e.g., 'boraded' instead of 'boarded') and two naming inconsistencies (e.g., 'firedupon' instead of 'fired upon') amongst 6 unique corrected event types over 2,357 events. The main source of mistakes was the victim ship type identifiers, where there were 12 incorrect spellings (e.g.,

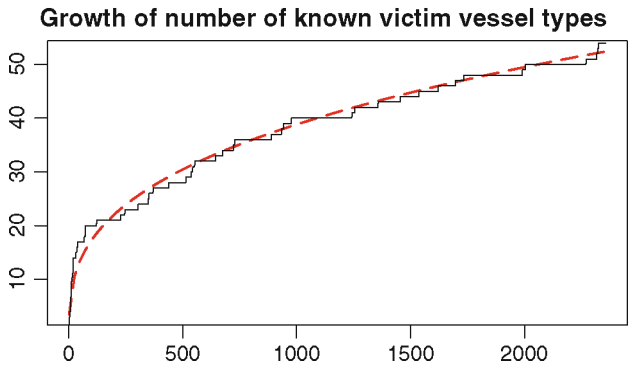

Fig. 12 Growth of the number of unique victim ship types as more event descriptions are processed. The *solid line* shows the number of victim types and the *red dashed line* the function $f(x) = 3.46x^{0.35}$ (colour figure online)

'vehicule carrier' instead of 'vehicle carrier') and 20 naming inconsistencies (e.g., 'refrigerated cargo' versus 'reefer') amongst 54 unique corrected actor types. 5 plus 32 corrections over 2,357 entries is about 1.5 %, which is a relatively low rate of mistakes. In all, the mistakes could be solved with 37 correction rules. Over time the number of correction rules that had to be added dropped very rapidly, because the growth of the number of new types slowed down. An illustration of the growth of the vocabulary of ship types is shown in Fig. 12. The growth is significantly slower than a square root of the number of event records. Many of these errors could have been easily prevented if a stricter form had been used for recording the piracy reports. For instance, auto completion in the data entry user interface could have avoided many typo's like writing 'fried upon' instead of 'fired upon'.

Event models have recently gained more interest in the research community, because they can be used to represent 'who', 'what', 'where', and 'when', which are core concepts in many different domains. A simple representation of events, actors, places, and their respective types makes it easy to analyze events. For instances, when events, actors, or places are identified by a single URI they can be counted easily. Adopting an RDF graph representation makes it easy to add in extra descriptive properties of event facets. For example, in the case described in this article it was simple to add a hierarchical classification of the places into EEZs and regions without any other changes to the data and with minimal changes (only additions) to the SPARQL queries used to count the events.

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 number of pirate boats and pirates, the intervention of a coalition war ship or helicopter, the outcome of the attack, etc. All these aspects, when present in the report, are informally stated and their formalization would help to answer further research questions such as: is there a difference in the level of aggression in the attacks in the Gulf of Aden? What is the status of most of the attacked vessels? How many cases of attacks where a warship intervened had a successful outcome? Also, we would like to investigate the possibility to interlink the Linked Open Piracy dataset with news items on the 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.

Acknowledgments 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. Bellamy C (2011) Maritime piracy—return of the world's second-oldest security problem. RUSI J 156(6):78–83
2. Bellare K, McCallum A (2007) Learning extractors from unlabeled text using relevant databases. In: Proceedings of sixth international workshop on information integration on the web (IIWeb-07), in conjunction with AAAI-07, July 23. AAAI Press, Vancouver, pp 10–16
3. Bensassi S, Martínez-Zarzoso I (2012) How costly is modern maritime piracy to the international community? Rev Int Econ (preprint)
4. Bizer C (2004) D2RQ—treating non-RDF databases as virtual RDF graphs. In: Proceedings of the 3rd international semantic web conference (ISWC2004)
5. Canisius S, Sporleder C (2007) Bootstrapping information extraction from field books. In: Proceedings of the 2007 joint meeting of the conference on empirical methods on natural language processing (EMNLP) and the conference on natural language learning (CoNLL), June 28–30. ACL, Prague, pp 827–836
6. Cohen WW (1995) Fast effective rule induction. In: Twelfth international conference on machine learning (ICML'95), pp 115–123
7. Crofts N, Doerr M, Gill T, Stead S, Stiff M (2008) Definition of the CIDOC conceptual reference model. Technical report, ICOM/CIDOC CRM Special Interest Group. version 4.2.5
8. Ding L, Lebo T, Erickson JS, DiFranzo D, Williams GT, Li X, Michaelis J, Graves A, Zheng J, Shangguan Z, Flores J, McGuinness DL, Hendler JA (2011) Twc logd: A portal for linked open government data ecosystems. J Web Semant 9(3):325–333
9. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1)
10. Hiebel G, Hanke K, Hayek I (2010) Methodology for CIDOC CRM based data integration with spatial data. In: 38th annual conference on computer applications and quantitative methods in archaeology. Granada, Spain
11. Jakob M, Vaněk O, Pěchouček M (2011) Using agents to improve international maritime transport security. IEEE Intell Syst:90–95

12. Kauppinen T, Gräler B (2012) Using the SPARQL package in R to handle Spatial Linked Data. http://www.linkedscience.org/tools/sparql-package-for-r/tutorial-on-sparql-package-for-r/
13. Lendvai P, Hunt S (2008) From field notes towards a knowledge base. In: Proceedings of the sixth international language resources and evaluation (LREC'08), 28–30 May 2008. European Language Resources Association (ELRA), Marrakech, pp 644–649
14. Li Ding DD, McGuinness DL, Hendler J, Magidson S (2009) The data-gov wiki: a semantic web portal for linked government data. In: 8th international semantic web conference (ISWC 2009)
15. Maletic J, Marcus A (2000) Data cleansing: beyond integrity analysis. In: Proceedings of the conference on information quality (IQ 2000), 20–22 Oct. Cambridge, pp 200–209
16. Omitola T, Koumenides C, Popov I, Yang Y, Salvadores M, Szomszor M, Berners-Lee T, Gibbins N, Hall W, Schraefel MC, Shadbolt N (2010) Put in your postcode, out comes the data: a case study. In: 7th extended semantic web conference (ESWC 2010)
17. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137
18. Ramsey A (2011) Alternative approaches: land-based strategies to countering piracy off the coast of somalia. Technical report, Civil Military Fusion Centre
19. Shaw R, Troncy R, Hardman L (2009) Lode: linking open descriptions of events. In: 4th annual Asian semantic web conference (ASWC'09). Shanghai, China
20. Tsilis T (2011) Counter piracy escort operations in the gulf of aden. Master's thesis, Naval Postgraduate School, Monterey
21. UNOSAT / UNITAR. Spatial analysis of somali pirate attacks in 2009. http://www.unosat-maps.web.cern.ch/unosat-maps/SO/CE20100714SOM/UNOSAT_SOM_CE2010-PiracyAnalysis_Report_HR_v1.pdf, June 2010
22. Van Erp M (2010) Accessing natural history: discoveries in data cleaning, structuring, and retrieval. PhD thesis, Tilburg University
23. van Erp M, Oomen J, Segers R, van den Akker C, Aroyo L, Jacobs G, Legêne, van der Meij L, van Ossenbruggen J, Schreiber G (2011) Automatic heritage metadata enrichment with historic events. In Museums and the Web 2011
24. van Hage WR, Malaisé V, Segers R, Hollink L, Schreiber G (2011) Design and use of the simple event model (SEM). J Web Semant 9(2):128–136
25. van Hage WR, Wielemaker J, Schreiber G (2010) The space package: tight integration between space and semantics. Trans in GIS 14(2)
26. Wang Y (2011) Semantically-enhanced recommendations in cultural heritage. PhD thesis, Technische Universiteit Eindhoven
27. Wielemaker J, Huang Z, van der Meij L (2008) SWI-prolog and the web, volume theory and practice of logic programming. Cambridge University Press, Cambridge, pp 363–392
28. Willems N, van Hage WR, de Vries G, Janssens J, Malaisé V (2010) An integrated approach for visual analysis of a multi-source moving objects knowledge base. Int J Geogr Inf Sci 24(9):1–16