# RDFising Events in the World: Raising News to the LOD Cloud

Marieke van Erp[1], Marco Rospocher[2], Piek Vossen[1], Itziar Aldabe[3], and Aitor Soroa[3]

[1] VU University Amsterdam
marieke.van.erp,piek.vossen@vu.nl
[2] Fondazione Bruno Kessler
rospocher@fbk.eu
[3] The University of the Basque Country (UPV/EHU)
itziar.aldabe,a.soroa@ehu.es

**Abstract.** News articles report on events happening in the world. Up to now, information about these events has been difficult to capture in a structured form such as RDF due to the complexity of language. However, with natural language processing technology coming of age, it is starting to become feasible to tap into this wealth of information. In this paper, we present our pipeline for extracting event information from large sets of news articles in English and Spanish in order to (1) create structured data of who did what when and where as well as opinions and speculations, and (2) link the extracted content to entities of the LOD cloud. Whilst our information extraction pipeline may not be perfect, the redundancy of the data smoothes out recall, and the linking to ontologies and LOD sources enable filtering the data based on cleaner background knowledge. By tracking the provenance of the extracted information, so that user can always refer back to the original article, our pipeline produces rich datasets that contains both unstructured (raw text) and structured content (RDF) in an interlinked manner. We demonstrate this by presenting two datasets that we have produced, highlighting their structure, volume and complexity. This illustrates how our platform can process daily streams of news and publish the reported events as a structured resource for researchers to use without having to process data or set up state-of-the-art NLP technology themselves. The pipelines and produced datasets are open and available to the public.

**Keywords:** natural language processing, events, unstructured text

## 1 Introduction

Using natural language processing technology, it is starting to become feasible to tap into the wealth of information contained in news articles. Large-scale news analysis is used by companies to gather competitive intelligence [8]. Government information specialists utilise news banks to inform parliament and collect information for enquiries [7]. Oftentimes, their tasks involve querying a news archive,

filtering the articles and performing an analysis on the textual data of the news, which can be a time-consuming manual process. Natural language processing (NLP) technology is being utilised to ease some of the manual labour, for example by clustering the information around certain topics (cf. Google News[4] or extracting entities and/or showing how 'trending' a topic is (cf. European Media Monitor[5]). However, thus far information from automatically processed news articles have not been published in a format that is easily queryable using common vocabularies. Moreover, most solutions rely on the mentions in the text to display quantitative information on news. They do not represent the core changes reported in the news as instances of events across all these mentions.

Within the NewsReader project,[6] we have set up a state-of-the-art NLP pipeline that performs syntactic and semantic analyses of documents to extract events, participants, locations, times, opinions and speculations which are then aggregated across documents to obtain a structured dataset of the aggregated information contained in the documents. The extracted information is enriched with links to external sources such as DBpedia and WordNet to abstract from the mentions of entities in the articles to a semantic representation at the instance level. That is, our pipeline enables us to expose the content of any news article as proper linked data, thus favouring the publication of news dataset in the LOD cloud. We applied our pipeline to two news corpora, one specific to the global automotive industry and one general corpus, thus producing two large-scale LOD-compliant structured datasets. Our contributions are the following:

- end-to-end pipeline (openly available through a virtual machine) for extracting events from text and converting them to RDF
- dataset about the Global Automotive Industry 2003-2013 containing $46,359,301$ triples
- English and Spanish Wikinews dataset 2005 - 2014 containing $9,819,141$ triples and $1,869,387$ respectively.

The remainder of this paper is organised as follows. In Section 2, we present background and related work. In Section 3, we explain our processing pipeline. In Section 4, we present our Global Automotive Industry dataset. In Section 5, we present our Wikinews dataset. We discuss the limitations and open issues of our approach in Section 6 and finish with conclusions and future work in 8.

## 2 Background

Since the start of the Linking Open Data project in 2007, the Linked Open Data cloud has seen an explosive growth in a wide variety of datasets such as data about persons, locations, census results and drugs [1]. Most of these datasets are RDF versions of data that was already structured in databases or spreadsheets. In NewsReader, we focus on textual data, from which we extract event

[4] `http://news.google.com`
[5] `http://emm.newsbrief.eu/overview.html`
[6] `http://www.newsreader-project.eu`

structures, as events form the core of explaining news stories. There are already some datasets of events present in the LOD cloud, most notably Last.FM[7] and EventMedia.[8] Last.FM is an RDF version of the Last.FM website,[9] containing information about events, artists, and users. EventMedia is an aggregate of three public event directories (last.fm, eventful, and upcoming) and two media directories (flickr, YouTube). Events are represented using the LODE ontology,[10] while media are represented with the W3C Ontology for Media Resources.[11] It is interlinked with DBpedia, Freebase, Geonames[12] and it also contains links to numerous related web pages from MusicBrainz,[13] Last.fm, Eventful[14] Upcoming,[15] and Foursquare.[16]

As for the coverage of news data, we have found a few datasets in this domain such as the New York Times dataset,[17] which mainly consists of subject headings (10K) that are tagged with People, Organizations, Locations, and Subject Descriptors information. El Viajero's tourism dataset[18] integrates about $20,000$ resources about travel from newspapers and digital platforms belonging to the Prisa Digital Group. However, the integration mostly takes place at the metadata level, rather than deep processing of the content of the resources.

Probably the closest to our datasets is the Ontos News Portal dataset.[19] In this dataset, persons, organisations, and locations, as well as some facts about these entities are extracted from news articles. The articles themselves are not part of the dataset, instead links are provided. The Ontos News Portal focuses on topics in five different categories: politics, business, technology, sports and stock intelligence. The main difference with the NewsReader datasuite is that in the Ontos News Portal shallow natural language processing techniques are applied, making this dataset less interesting for language researchers or users interested in deeper layers beyond topics and entities.

The NewsReader datasets differ from these datasets in a few key aspects. Firstly, we provide links to the exact position in the news article where an event or entity is mentioned, enabling analyses at a much more fine-grained level than for example the Ontos News Portal offers. We can thus create time-ordered sequences of events from large data sets that show characteristic development, trends and storylines over time. While we currently slice our datasets around particular topics, to keep them manageable in size, we can treat any topic as our

[7] http://datahub.io/dataset/rdfize-lastfm
[8] http://eventmedia.eurecom.fr/
[9] http://www.last.fm
[10] http://linkedevents.org/ontology/
[11] http://www.w3.org/TR/mediaont-10/
[12] http://www.geonames.org
[13] http://musicbrainz.org/
[14] http://www.eventful.com
[15] http://www.upcoming.org
[16] http://www.foursquare.com
[17] http://data.nytimes.com/
[18] http://datahub.io/nl/dataset/elviajero
[19] http://datahub.io/nl/dataset/ontos-news-portal, http://news.ontos.com/

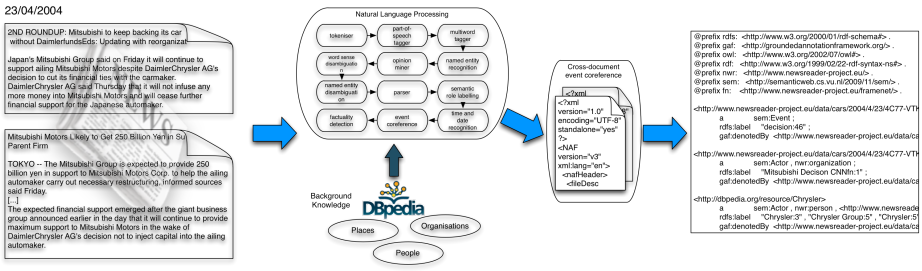


Fig. 1: Overview of the processing pipeline

tools are not focused on particular topics or domains. In this paper, we present a dataset around the global car industry which provides users with a rich data source around one particular domain, as well as a general news dataset which enables users to explore links between different topics as particular entities or types of events occur across different domains.

There are already tools to convert output from standard NLP modules to RDF such as those developed by the NLP2RDF project.[20] However, after executing a fairly standard NLP processing pipeline, we perform an additional cross-document aggregation step to move from text mentions to instances, which is an additional step beyond what such tools currently offer. Our instance representation furthermore allows us to aggregate information across many different sources, showing complementarity and differences across these sources and the provenance of the information provided.

## 3 Data Processing Methodology

Figure 1 shows a schematic overview of our processing pipeline. There are two main steps to our data processing methodology: information extraction from the document and cross-document event coreference. The document information extraction pipeline is a typical natural language processing pipeline, where we extract mentions of events, actors and locations from the text. During the cross-document event coreference step, the mentions are crystallized into instances, effectively forming a bridge between the document (NLP) and instance (SW) levels.

To link the two levels, we defined the Grounded Annotation Framework or GAF [4][21] for representing events and actors in text and as structured data. GAF combines a representation of the interpretation of mentions in text with a representation of instances of events and entities in a semantic web resource. For the textual representation, we used the Natural Language Processing Annotation Format (NAF [5][22]) and for the representation of instances of events and

[20] `http://nlp2rdf.org/`
[21] `http://groundedannotationframework.org`
[22] `http://wordpress.let.vupr.nl/naf/`

entities, we use the Simple Event Model (SEM [6][23]). Mentions of events and entities in texts are linked to instances of events and entities in SEM through GAF:DENOTEDBY and GAF:DENOTES links.

In the remainder of this section, we detail the steps in our pipeline and conclude with details on our implementation.

### 3.1 Information Extraction from Documents

The information extraction pipeline extracts entities and events from raw text of newspaper articles. To facilitate easy processing and keeping track of the provenance of the system, we mark up the raw text with metadata such as title and publication date using the NAF format. Every module in the pipeline adds an element to the header indicating the version of the module that was used and a timestamp as well as a layer encoding the information that was extracted from the text.

The information extraction processing starts with the TOKENISER which splits the text into sentences and words. The PART-OF-SPEECH TAGGER adds information on the word type to each word, for example indicating whether the word is a noun or a verb. The MULTIWORDS TAGGER detects multiword expressions in WordNet, partly resolving ambiguity. Then, the NAMED ENTITY RECOGNISER detects named entities and tries to categorise them as person, location, organisation or miscellaneous names. This module is followed by the OPINION MINER, which is aimed at detecting opinions (whether there is a positive or negative sentiment towards something), opinion holders (who has the opinion) and opinion targets (what is the opinion on). The WORD SENSE DISAMBIGUATION MODULE discerns between different meanings of words and encodes the most plausible meaning depending on its context. The NAMED ENTITY DISAMBIGUATION MODULE attempts to resolve the named entities against a knowledge base (in this case DBpedia), in order to link them to an entity instance. The PARSER aims to detect the syntactic structure of the sentence, e.g. the subject and object of the clause. The SEMANTIC ROLE LABELLER tries to detect the semantic arguments to predicates, e.g. who is the agent in *How far can you go with a Land Rover?*. We then employ a TIME AND DATE RECOGNISER to tag temporal expressions so events can later be organised on a timeline. Which entities and events are the same within the document is computed via COREFERENCE MODULES. Finally, the FACTUALITY CLASSIFIER is employed to determine which events have taken place or will probably take place, and which are denied or are mentioned in a speculative manner.

### 3.2 Cross-document Event Coreference

The information extraction from documents step results in the interpretation of a single textual source (i.e. document) in NAF. The text is treated as a sequence of tokens that are interpreted by the various modules. The same event and the

[23] `http://wordpress.let.vupr.nl/sem/`

same entities can be mentioned several times within such a sequence. Information on each of these mentions can be incomplete: one sentence may make reference to the time and place, while another sentence may specify the actors involved in an event. If we consider a large set of textual sources, we will also find many references across these sources that overlap and complement each other. To go from these mention-based representations in NAF to an instance representation in SEM, we go through a number of steps resolving co-reference across mentions (see [3] for a detailed description of our approach).

- Within-document co-reference (mention-based)
  - entity coreference
  - event coreference through mention similarity
- Cross-document co-reference (instance-based)
  - clustering events within the same time and place constraints
  - event coreference through instance similarity

The NLP modules already identify the entities in the text and where possible assign a URI to each of them. The Entity Coreference Module uses the available information to decide which entities refer to the same instance but also resolves anaphoric expressions. Each entity coreference set is used to represent a unique entity instance. If we have a unique URI (e.g. to DBPedia), it is used to identify the entity, otherwise, we create a unique URI based on the identifiers in NAF. Entity instances across documents can share the same URI if it is based on an external LOD resource. They get a single representation with GAF:denotedBy links to all the places in the NAF files where they were mentioned. Entities without an external LOD URI are therefore local to a single NAF file and have only mentions within that source. Entity instances can be persons, organisations, places and time points and durations. For each instance, we also provide the types and the labels given by the NLP modules.

```
(In this example, to shorten some long URIs just for the sake of readibility, we ‘‘improperly’’
abbreviated them with QNAMEs (e.g., nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#coe56 should
be read as <http://www.newsreader-project.eu/data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#coe56>))

PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX time:<http://www.w3.org/TR/owl-time#>
PREFIX gaf:<http://groundedannotationframework.org/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX nwr:<http://www.newsreader-project.eu/>
PREFIX sem:<http://semanticweb.cs.vu.nl/2009/11/sem/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX prov: <http://www.w3.org/ns/prov>

nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#coe56
    a              sem:Event ;
    rdfs:label     "provide" ;
    gaf:denotedBy
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#char=2742,2749&word=w506&term=t506,
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#char=2812,2820&word=w517&term=t517,
    nwr:data/cars/2013/1/1/57DF-TK31-DXF1-N0P1.xml#char=438,446&word=w79&term=t79,
    nwr:data/cars/2013/1/1/57DG-05S1-DXF1-N18B.xml#char=1452,1460&word=w257&term=t257,
    nwr:data/cars/2013/1/1/57DG-05S1-DXF1-N18Y.xml#char=1576,1583&word=w283&term=t283,
    nwr:data/cars/2013/1/1/57DG-05S1-DXF1-N197.xml#char=508,516&word=w93&term=t93,
```

```
    nwr:data/cars/2013/1/1/57K5-FKK1-DYBW-2534.xml#char=718,726&word=w114&term=t114.

dbpedia:Democratic_Republic_of_the_Congo
  a              sem:Place , nwr:location ;
  rdfs:label     "democratic republic of the congo" , "DRC" ;
  gaf:denotedBy
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#char=2662,2694&word=w489,w490,w491,w492,w493
    &term=t.mw489,
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#char=2696,2699&word=w495&term=t495 ,
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#char=2842,2845&word=w522&term=t522 .

nwr:data/cars/2013/1/1/57DF-TK31-DXF1-N0P1.xml#pr12,rl23 {
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#coe56
    sem:hasActor  dbpedia:Ford_Motor_Company .
}

nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#docTime_1 {
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#coe56
    sem:hasTime  nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#dct .
}

nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#pr76,rl195 {
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#coe56
    sem:hasPlace  dbpedia:Democratic_Republic_of_the_Congo .
}

nwr:time/20130101
    a time:DateTimeDescription ;
    time:unitType time:unitDay ;
    time:day "1"^^<http://www.w3.org/2001/XMLSchema#gDay> ;
    time:month "1"^^<http://www.w3.org/2001/XMLSchema#gMonth> ;
    time:year "2013"^^<http://www.w3.org/2001/XMLSchema#gYear> .

nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#dct
    a sem:Time, time:Instant ;
    rdfs:label "2013-01-01" ;
    gaf:denotedBy
   nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#nafHeader_fileDesc_creationtime ;
    time:inDateTime nwr:time/20130101 .

#factuality details
nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#facValue_2 {
    nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#coe56
    nwr:value/hasFactBankValue "CT+" .
}

#provenance details
nwr:data/cars/2013/1/1/57DF-TK31-DXF1-N0P1.xml#pr12,rl23
    gaf:denotedBy
     nwr:data/cars/2013/1/1/57DF-TK31-DXF1-N0P1.xml#rl23 ,
        nwr:data/cars/2013/1/1/57DF-TK31-DXF1-N0P1.xml#rl25 ;
    prov:wasAttributedTo nwr:sourceowner/India_Automobile_News .
```

We also create instances for events. In the case of events, we usually do not have an external URI. Within a single NAF file, we compare mentions of events coming from the Semantic Role Labeling. Those mentions with a similarity score above a threshold[24] constitute a single instance and we combine all the SEMANTIC ROLE LABELLER as RDF triples for that instance. For example, one sentence mentions an acquisition of a company in relation to the date, whereas another sentence mentions the amount of money paid for the acquisition and the com-

[24] Currently this is on the basis of the lemma (baseline approach) but will be extended to more complex classifiers combining more features (e.g. words, syntax, semantics).

pany that is the new owner. The instance representation for the acquisition event thus will be the subject of three triples aggregated from different mentions. The relation triples are stored as named-graphs to be able to express properties of the relations as statements, such as the provenance of the relation (what source made what statement, see the "provenance details" block in the example) or the factuality (did it really happen or is it speculated, see "factuality details" block in the example).

Since we deal with large corpora of documents, we first separate potential instances of events on the basis of disjunctive features. We assume that events are bound by time an place. We therefore store all events with the same time and place relations in a temporarily bucket. Events without time and place information are stored in a timeless and placeless bucket. We then only need to compare all instances within the same bucket. Since the actors and places are, where possible, normalised to their unique URI (preferably to external LOD repositories), we can now compare the event instances within the same bucket for matches of their event descriptions (currently the lemma), the actors and the places involved. We require that at least one actor and one place needs to match across instances to establish cross-document event-coreference. If this is the case, we merge the event-instances into a single instance and create the proper representation of all the relations for the merged instance, i.e. the subject of each relation is the identifier for the merged instance, while the predicate and object URIs can remain as they are. We can establish different degrees of overlap in the semantic components to experiment with higher and lower degrees of matching. For example, timeless events may require more overlap of other event components than events with very detailed time information. When merging instances, all pointers to mentions and provenance relations are merged as well.

### 3.3 Implementation

Modules that perform deep NLP processing such as for example the Semantic Role Labeller are resource intensive and time consuming processes (see Table 1). As we aim to process large amounts of textual data in a timely manner, we implemented an NLP pipeline that can scale up with the number of documents through parallelization.

We use Virtual Machines (VMs) as building blocks for the linguistic pipeline. We have defined one VM per language with all the required modules for event extraction. This approach allows the project to scale horizontally (or scale out) as a solution to the problem of dealing with massive quantities of data, just by deploying all NLP modules into VMs and making as many copies of the VMs as necessary to process the required documents.

Inside each VM, the modules are managed using the Storm framework for streaming computing.[25] Storm is an open source, general-purpose, distributed, scalable and partially fault-tolerant platform for developing and running distributed programs that process continuous streams of data. Storm allows setting

[25] http://storm.incubator.apache.org/

Table 1: Minimum (Min), Maximum (Max), Average number of seconds taken for a document per module and the standard deviation in times ($\sigma$). The time taken increases with the length of the documents.

| Module | Min | Max | Average | $\sigma$ |
|---|---|---|---|---|
| Tokenizer | 0.3 | 16.1 | 0.5 | 0.4 |
| Part-of-speech tagger | 1.0 | 164.4 | 2.1 | 2.4 |
| Multiword tagger | 2.6 | 895.1 | 4.2 | 12.0 |
| Named Entity Recognition | 0.6 | 135.4 | 2.3 | 2.2 |
| Opinion Miner | 0.2 | 646.9 | 1.3 | 10.6 |
| Word Sense Disambiguation | 0.1 | 36.0 | 0.8 | 1.4 |
| Named Entity Disambiguation | 0.5 | 1826.0 | 12.9 | 70.6 |
| Semantic Role Labeller | 18.1 | 942.1 | 39.8 | 22.9 |
| Time and date recognizer | 1.0 | 338.7 | 5.5 | 7.2 |
| Event Coreference | 0.2 | 229.3 | 1.7 | 2.9 |
| Factuality | 0.5 | 32.1 | 0.8 | 0.8 |

scalable clusters with high availability using commodity hardware and minimizes latency by supporting local memory reads and avoiding disk I/O bottlenecks.

The main abstraction structure of Storm is the *topology*, which describes the processing node that each message passes through. The topology is represented as a graph where nodes are processing components, while edges represent the messages sent between them. In our implementation, each topology node is a proxy to an external NLP module as described in Section 3.1. Each module receives a NAF document, creates annotations on top of it, and passes the enriched NAF to the next module. The current version of the VM containing our pipeline is available from `http://bit.ly/1hHVwVc`. All modules are also freely available through `https://github.com/newsreader`.

With this setting we have deployed eight copies of the VM machine and process the documents in parallel. We were able to process the Global Automotive Industry dataset (described in the next section), which consists of more than 64,000 documents, in less than 5 days. The cross-document coreference and the conversion to SEM takes a couple of hours to run.

The source news, the NAF files, and the RDF content resulting from the conversion to SEM, are all uploaded into the KnowledgeStore [2], a scalable, fault-tolerant, and Semantic Web grounded storage system that, thanks to a tight interlinking between the news, the structured content of the NAF documents, and the corresponding RDF entities and facts, enables to jointly store, manage, retrieve, and semantically query, both structured and unstructured content.

## 4 Global Automotive Industry

The NewsReader project targets large and complex global developments that span longer time periods of times. With its focus on financial and economic news,

Table 2: Statistics on the produced datasets

(a) Global Automotive Industry

| | Global Automotive Industry |
|---|---|
| Event mentions | 5,247,872 |
| Events | 1,784,532 |
| Location mentions | 1,049,711 |
| Locations | 62,255 |
| Actor mentions | 3,127,146 |
| Actors | 445,286 |
| # provenance triples | 12,851,504 |
| Total # triples | 46,359,301 |

(b) Wikinews

| | Wikinews | |
|---|---|---|
| | English | Spanish |
| Event mentions | 1,033,293 | 158,757 |
| Events | 811,885 | 158757 |
| Location mentions | 296,296 | 70,010 |
| Locations | 11,610 | 9047 |
| Actor mentions | 570,803 | 246,140 |
| Actors | 103,496 | 144,793 |
| # provenance triples | 3,470,932 | 725,733 |
| Total # triples | 9,819,141 | 1,869,387 |

the financial crisis is a good example of such a development. We selected the car industry as one of the specific cases as it is a complex global market involving not only industry but also governments and well-organized labor unions. This makes the dataset an interesting playground for researchers from different fields, as well as information specialists within companies or other organisations. We also expect to find interesting events in this dataset as the automotive industry has changed dramatically due to the financial crisis and upcoming new economies.

The documents were selected by first performing a query on the LexisNexis[26] News database by selecting specific keywords related to car companies ("Alfa Romeo", "Aston Martin", "BMW", etc). This initial query retrieved around 6 million News documents, which are further filtered by considering only documents spanning between 2001 and 2013 years and containing more than 2 car company names, resulting in 64,540 documents from 3,111 different publications. The goal is to obtain relevant documents describing events which involve two or more car companies. All document were then converted from the NITF format[27] that is used at LexisNexis to NAF and processed in the pipeline. The dataset can be downloaded from `http://datahub.io/dataset/global-automotive-industry-news`. Note that the dataset does not include the content of the source news documents as we are not authorized to redistribute them due to licensing issues although the articles can be looked up in the LexisNexis database by users with a subscription.

Despite the fact that this structured dataset has not yet been derived from all relevant articles about the global automotive industry, it already enables us to answer some complex questions about the domain. We can for example retrieve the different events and the locations and times they took place at to obtain insights into the localities of the global automotive industry. If we visualise these results on a map ordered by year as shown in Figure 2, we can see that there is little mention of India in our dataset in 2003, but in 2008 it is

[26] LexisNexis is one of the content partners in the NewsReader project

[27] `http://www.iptc.org/site/News_Exchange_Formats/NITF/`

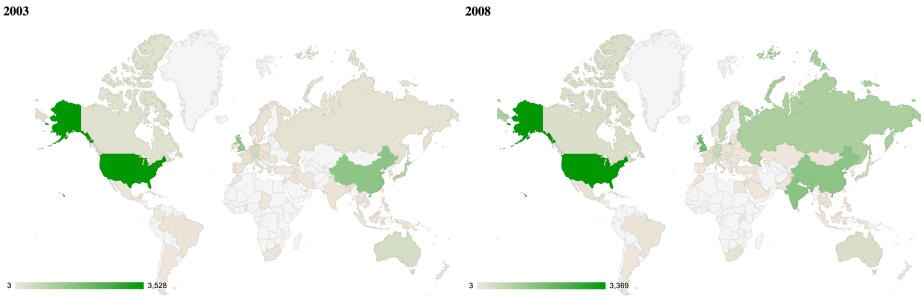


Fig. 2: Overview of number of events taking place in particular locations in the car dataset

an important player in the Global Automotive Industry. When further looking into these results, we find that in 2008 Tata Motors, an Indian car manufacturer, launched the Nano, an affordable car and bought Jaguar Land Rover from Ford.

As the data is centered around events and includes the actors that were involved in the events, users can also for example easily find out who interacted with whom in the domain. In Figure 3, we have plotted the network of the key actors in the car domain (based on the frequency of their occurrence in our dataset). The persons in the turquoise bubbles are all (former) CEOs or board members of car companies (e.g. Wendelin Wiedeking was the former president of Porsche, Alan Mullaly is the president of the Ford Motor Company). Furthermore, we also find politicians (orange) such as the president of the United States, Barack Obama. He is linked to the CEO of Chrysler, Sergio Marchionne, as the US government bailed out Chrysler after the 2008 - 2010 automotive industry crisis. Prince Bernhard of Lippe-Biesterfeld (orchid) is involved here, as both he and Jürgen Schrempp have participated in the Bilderberg conferences. Similarly, Kirk Kerkorian is involved in the car industry as an investor, and Ron Gettelfinger was the president of the Auto Workers Union.

Furthermore, the graph also uncovers a few errors from the entity disambiguation module (red), as Abraham Lincoln shows up here, where most likely the mentions of "Lincoln" in our corpus refer to the Lincoln Motor Company. Another suspicious entity in this graph is "Jim Hall (musician)" whereas mentions of this entity should have been linked to the Jim Hall who is managing director of the automotive consulting firm 2953 Analytics'. However, this Jim Hall is not present in DBpedia, hence the module chose a suboptimal entity to link to. The same cause underlies the suboptimal linking of David Healy (footballer), David Smith (sculptor), David Cole (producer), Matthew Taylor (footballer) and Liu Qi (Liu Biao's son). This shows an interesting area of future work to explore the limitations of current LOD sources and options to improve such problems by including information other sources corporate databases to provide alternative knowledge bases to link to.

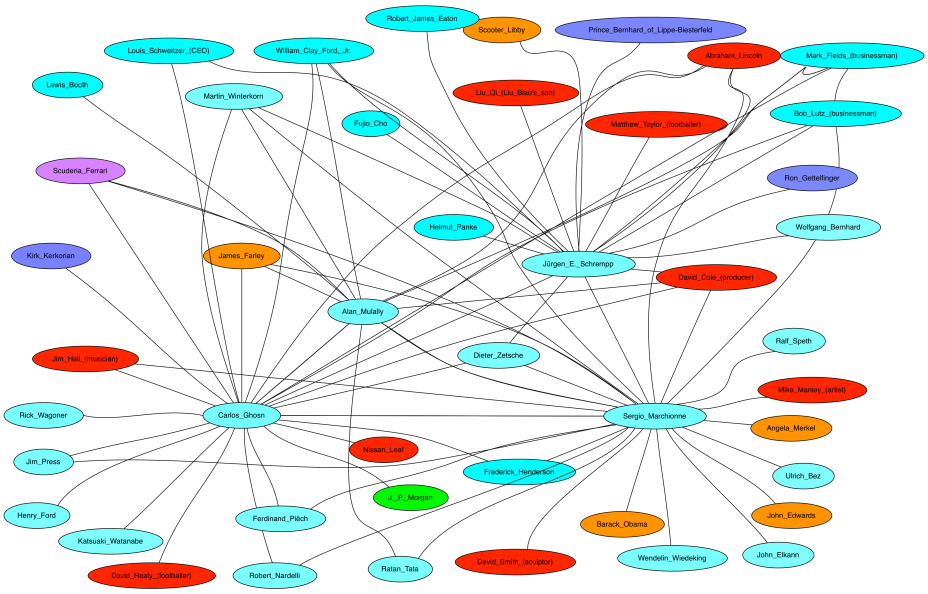


Fig. 3: The key actors' network in the global automotive industry domain

## 5 WikiNews

Wikinews[28] is a free multilingual open news source operated and supported by the Wikimedia foundation. We chose to use this source as it enables us to link entities and events across different language as well as its broad coverage. For the Wikinews dataset, we processed 18,886 English and 7,603 Spanish news articles. Contrary to the Global Automotive Industry dataset, this dataset is freely available, thus the original articles, as well as the full grammatical processing results are available. The Wikinews dataset can be found at `http://datahub.io/dataset/structured-wikinews`. For this dataset, we have made available the structured dataset, as well as the NAF files containing the grammatical analyses after the document processing pipeline, providing a multilayered playground for experimenting with NLP and SW technologies.

One thing that we can query the data for is which entities are most often mentioned in conjunction with President Barack Obama? Here we find that out of the mentions in total of Barack Obama in our corpus together with another entity, Mitt Romney co-occurs 112 times, United States co-occurs 80 times, The US Senate 56 times, The Republican Party 52 times and John McCain 51 times. The Democratic Party and Hillary Rodham Clinton do not occur in the top 5 (see Figure 4).

We also find that Barack Obama is mostly involved in statement events e.g. giving speeches (113 times), text creation e.g. signing bills (56 times) and

[28] `http://en.wikinews.org`

| Entity | Occurrences |
|---|---|
| <http://dbpedia.org/resource/Mitt_Romney> | 112 |
| <http://dbpedia.org/resource/United_States> | 80 |
| <http://dbpedia.org/resource/United_States_Senate> | 56 |
| <http://dbpedia.org/resource/Republican_Party_(United_States)> | 52 |
| <http://dbpedia.org/resource/John_McCain> | 51 |
| <http://dbpedia.org/resource/Democratic_Party_(United_States)> | 46 |
| <http://en.wikinews.org/wiki/Obama,_Romney_battle_over_foreign_policy_in_final_U.S._presidential_debate#c010> | 41 |
| <http://dbpedia.org/resource/Hillary_Rodham_Clinton> | 33 |
| <http://en.wikinews.org/wiki/Obama's_choice_for_Treasury_issues_warning_on_China#c01> | 31 |
| <http://dbpedia.org/resource/CNN> | 29 |

Fig. 4: Top 10 entities co-occurring with Barack Obama

arriving e.g. travelling (23 times). However, our processing pipeline also classifies Barack Obama as a location in 755 events. This is most likely due to an incorrect location assignment by the semantic role labeller in sentences such as *Gordon Brown stood next to Obama*. Future versions of our pipeline will be less sequential (such that the named entity disambiguation module can influence the semantic role labeller and prevent an agent from being linked to a location) and contain more fine-grained domain knowledge. However, as one can filter queries through DBpedia ontology categories, it is possible to detect these inconsistencies and leave them out of analyses when one is not concerned with perfect recall.

Another interesting example we found in the dataset is the mention of the suing of Apple, the computer company, by Apple Corps, the multimedia corporation founded by The Beatles[29]. The two mentions of Apple companies are correctly disambiguated in our dataset, and linked to the appropriate DBpedia resources.

## 6 Discussion

The datasets we have described are the first versions of the NewsReader datasuite. As such there are still a few issues to resolve. First of all, the data is not quite yet exposed as Linked Open Data. The main reason for the global automotive industry data for this is that we are working out a publication scheme with our content partners to create links from the derived information from the sources (the events and entities) to the original articles in a closed database. For the Wikinews articles, there are no such restrictions, but there is the issue of there being the original articles, the articles enriched with grammatical analyses and the final cross-document aggregated information. We are currently devising a model to expose these three different information levels in an intuitive manner.

Furthermore, our datasets are also "alive" in the sense that they are currently periodically updated and in the future continually. We are also improving the modules in our pipeline, which will refine the results as we progress. The main

[29] `http://en.wikinews.org/wiki/Beatles'_Apple_Corps_sues_Apple_Computer`

area of concern here is to extend the use of background knowledge (beyond DBpedia) to include more focused domain knowledge. This should improve the performance of our named entity disambiguation module.

## 7 Value to the Research Community

All the tools used are available as open source software, including the Virtual Machines that contain the English and Spanish pipelines[30] This enables anyone to process their own dataset. Although we condense the information in the textual documents into events and entities, one can always relate the information to the source as the *evidence* of every statement is available through the GAF:DENOTEDBY links.

The Global Automotive Industry and WikiNews datasets already provide a rich resource for the research community as a starting point to explore information that is obtained automatically from text. There are many interesting avenues of research here, such as investigating at which stages in the NLP pipeline it is beneficial to add background knowledge. Another avenue could be the exploitation of structured information already present in the resource to improve the quality of the resource (such as for example wikilinks and categories in the WikiNews dataset). Within the NewsReader project, we are currently in the first stages of investigating the use of *knowledge crystallisation* for these datasets. This includes "cleaning" the data by exploiting reasoning techniques and background knowledge to keep the most compact description possible of an event or entity extracted from text. We aim to remove redundant content and noise. More in general, the information automatically extracted from the NLP pipeline could be considered as "evidence" of facts mentioned in text, and only when the amount of evidence exceeds an appropriate threshold or certain conditions are met, these facts should be considered as *first class citizens* to be included in the final dataset.

Beyond the Semantic Web community, we envision our datasets to be useful to NLP researchers for benchmarking as well as creating timelines and storylines and as a resource for resolving entity ambiguity. Journalism researchers could utilise it for comparison of coverage between different sources as we keep careful track of what fact was extracted from what data source. Furthermore, we are already working with information professionals in the financial-economic domain and the Dutch House of Representatives to investigate how such structured datasets can aid them in performing their daily tasks.

[30] Currently available through for evaluation purposes at: `http://bit.ly/1hHVwVc` (English) and `http://bit.ly/1jTukXL` (Spanish). A public release via the NewsReader website is planned for this summer. Demos of both pipelines are also available at `http://bit.ly/1uXZjXu` (English) and `http://bit.ly/1onO8FE` (Spanish). A demo of the cross-document event coreference module is available at `http://bit.ly/1mTThBs` and the code at our Github page

## 8 Conclusions and Future Work

We have presented a pipeline to convert textual news articles to structured RDF, as well as two datasets showcasing the advantages of such a conversion. Our pipeline uses state-of-the-art technologies from NLP and SW. The Global Automotive Industry dataset presents a large-scale dataset from which observations about the car industry between 2003 and 2013 can be made. The Wikinews dataset shows how articles in multiple languages differ in their coverage of events.

We are currently working on optimising our pipeline to be able to process articles faster and scale up to process daily news streams and expand to Dutch and Italian. Furthermore, we are investigating other datasets to link to both withing and beyond the LOD cloud to provide better coverage for the entity disambiguation module.

Currently the datasets are available as regularly updated, but static data dumps, but future versions of the data will be published as Linked Open Data, enabling anyone to link their resources to the changes reported on in the news.

**Acknowledgements.** This research was funded by European Union's 7th Framework Programme via the NewsReader project (ICT-316404).

## References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
2. Corcoglioniti, F., Rospocher, M., Cattoni, R., Magnini, B., Serafini, L.: Interlinking unstructured and structured knowledge in an integrated framework. In: 7th IEEE International Conference on Semantic Computing (ICSC), Irvine, CA, USA (2013)
3. Cybulska, A., Vossen, P.: Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014). Reykjavik, Iceland (May 26-31 2014)
4. Fokkens, A., van Erp, M., Vossen, P., Tonelli, S., van Hage, W.R., Serafini, L., Sprugnoli, R., Hoeksema, J.: Gaf: A grounded annotation framework for events. In: Proceedings of the 1st workshop on Events: Definition, Detection, Coreference, and Representation at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013). No. ISBN 978-1-937284-47-3, Association for Computational Linguistics, Atlanta, GA, USA (Jun 9-15 2013)
5. Fokkens, A., Soroa, A., Beloki, Z., Rigau, G., van Hage, W.R., Vossen, P.: Naf: the nlp annotation format. Tech. Rep. NWR-2014-3, VU University (2014)
6. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the Simple Event Model (SEM). J. Web Sem. 9(2), 128–136 (2011)
7. Stouten, P., Kortleven, R., Hopkinson, I.: Test data and scenarios. Deliverable 8.1, NewsReader Project (2013)
8. Ziegler, C.N.: Competitive intelligence capturing systems. In: Mining for Strategic Competitive Intelligence, pp. 51–62. Springer Berlin Heidelberg (2012)