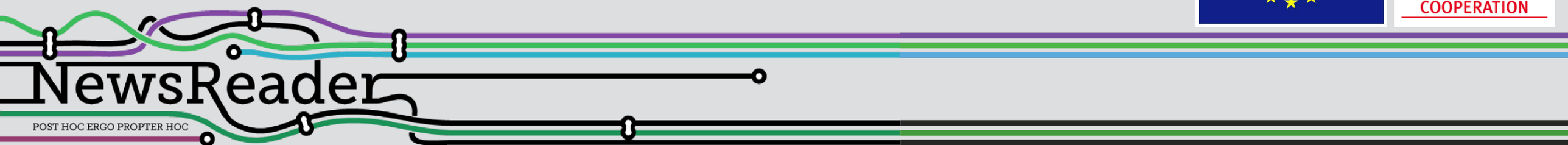
NEWSREADER

RECORDING HISTORY
BY PROCESSING MASSIVE STREAMS OF DAILY NEWS

PIEK VOSSEN, VU UNIVERSITY AMSTERDAM,
GERMAN RIGAU, BASQUE UNIVERSITY, SAN SEBASTIAN
LUCIANO SERAFINI, FEDERATION BRUNO KESSLER, TRENTO
PIM STOUTEN, LEXISNEXIS, AMSTERDAM
FRANCIS IRVING, SCRAPERWIKI, LIVERPOOL
WILLEM VAN HAGE, SYNERSCOPE, EINDHOVEN

LREC-2014, REYKJAVIK



HOW DID THE WORLD CHANGE YESTERDAY?

CAN WE HANDLE THE NEWS?

- Information broker LexisNexis:
- 1.5 millions news articles on a single working day
- 30,000 different sources

HOW DID THE CAR INDUSTRY CHANGE DURING THE FINANCIAL CRISIS

- 6 million English articles on the car industry in the LexisNexis archive for the last 10 years

- 2 million Google hits for "Volkswagen takeover" not sorted by publication date

Trends

Web Search Interest: **volkswagen**. Worldwide, 2004 - present.

Explore trends

Hot searches

Interest over time

The number 100 represents the peak search interest

News headlines Forecast

Search terms

volkswagen

+ Add term

Other comparisons

Limit to

Explore trends

Hot searches

Search terms

volkswagen

+ Add term

Other comparisons

Limit to

THE PROBLEM

10k entities per year

1995 96 97 98 99 2000 01 02 03 04 2005 06 07 08 09 2010 11 12 13 14 2015

HOW MANY CHANGES?

200k mentions per year

1995 96 97 98 99 2000 01 02 03 04 2005 06 07 08 09 2010 11 12 13 14 2015

6 MILLION ARTICLES

Past

News

News

Speculation

On 16 September 2008, Porsche *increased its shares* by another 4.89%, in effect *taking control of* the company, with more than 35% of the voting rights.

6 Jan 2009 – Porsche has been on *a quest to takeover* VW for more than two years.

DAILY NEWS TSUNAMI

- **VOLUME IS TOO BIG**: *1,5 MILLION ITEMS EACH WORKING DAY*
- **REPEATED AND DUPLICATED**: *WE CANNOT DISTINGUISH THE NEW FROM THE OLD*
- **INCOMPLETE AND PIECEMEAL**: *WE NEED TO READ ALL TO GET THE COMPLETE PICTURE*
- **ACTUAL AND SPECULATED EVENTS**: *WE CANNOT DISTINGUISH THE REALIS FROM IRREALIS, SPECULATIONS, FEARS AND HOPES*
- **INCONSISTENT AND CONTRADICTORY**: *WE CANNOT TELL TRUE FROM FALSE AND WHO TO BELIEVE*
- **OPINIONATED AND SELECTIVE**: *WE DO NOT REALIZE THE BIAS OF OUR SOURCES*

WHAT IF COMPUTERS COULD READ THE NEWS?

NEWSREADER (ICT316404)

- Technology to process massive streams of news from many different sources in 4 languages (English, Dutch, Spanish and Italian):
  - Recording the changes in the world as they are told in the media over long periods of time → history-recorder.
    - What happened, where and when, who was involved.
    - What temporal and causal relations hold, what intentions are involved.
    - Who made what statement, where do sources agree and disagree: provenance!
    - KnowledgeStore that handles dynamic growth of information, reflecting long-term developments.
    - Organise and visualise massive amount of changes as stories, scripts, plots to provide efficient access

GROUNDED ANNOTATION FRAMEWORK

- **GAF**: groundedannotationformat.org.
- Distinguishes between mentions of entities and events in sources (text, images, movies, databases, sensors) and the representation of instances in the assumed world.
- Mentions are semantically represented in the **NAF** (NLP Annotation Format) representation of the text (same instances mentioned in at different places in the text and in different texts).
- Instances are semantically represented in **SEM** (RDF-based Simple Event Model, Van Hage et al 2011) using URIs.
- gaf:denotes and gaf:denotedBy links to connect the two.

MENTIONS

Forbes
4/23/2004 @ 5:01PM
http://www.forbes.com/2004/04/23/cz_jf_0423flint.html

DaimlerChrysler just refused to make a $7 billion to $8 billion cash infusion to the floundering company (Mitsubishi). ... His tactics led to massive investments in American Chrysler (a takeover), in Mitsubishi (37% ownership and control) and Korean Hyundai (10% and no control).

New Zealand Herald,
Monday Apr 26, 2004
http://www.nzherald.co.nz

Schrempp may have suffered his own personal Waterloo on Friday when Daimler's board voted to pull the plug on troubled Japanese carmaker Mitsubishi Motors rather than pump in billions of euros to keep the company on financial life support.

INSTANCES

WHAT: decision
WHO: DaimlerChrysler
WHEN: Friday, April, 23, 2004

NOT

WHAT: invest
WHO: DaimlerChrysler
WHO: Mitsubishi
WHO: $7-8 billion euros

MENTIONS

New York Times, By MARK LANDLER
Published: April 24, 2004
http://www.nytimes.com/2004/04/24/

Even Mr. Schrempp's hold on the chief executive's job at DaimlerChrysler seems shaky in the wake of his company's unexpected refusal to aid a bailout of the financially troubled Mitsubishi.

Automotive News, Monday
Apr 26, 2004:3
http://www.autonews.com

The decision not to bail out Mitsubishi Motors Corp. raises fresh doubts about the future of DaimlerChrysler CEO Juergen Schrempp.

NAF EXAMPLE

Toyota brought Lexus to Japan in 2005.

```
<predicate id="pr36">
  <!--brought-->
  <externalReferences>
    <externalRef reference="bring.01" resource="PropBank"/>
    <externalRef reference="bring-11.3-1" resource="VerbNet"/>
    <externalRef reference="Bringing" resource="FrameNet"/>
  </externalReferences>
  <span><target id="t199"/></span>
  <role id="rl84" semRole="A0">
    <!--Toyota-->
    <externalReferences>
      <externalRef reference="bring-11.3#Agent" resource="VerbNet"/>
    </externalReferences>
    <span><target head="yes" id="t198"/></span>
  </role>
  <role id="rl85" semRole="A1">
    <!--Lexus-->
    <externalReferences>
      <externalRef reference="bring-11.3#Theme" resource="VerbNet"/>
    </externalReferences>
    <span><target head="yes" id="t200"/></span>
  </role>
  <role id="rl86" semRole="A3">
    <!--to Japan-->
    <span><target head="yes" id="t201"/><target id="t202"/>
    </span>
  </role>
  <role id="rl87" semRole="AM-TMP">
    <!--in 2005-->
    <span><target head="yes" id="t203"/><target id="t204"/>
    </span>
  </role>
</predicate>
```

```
<entities>
  <entity id="e1" type="person">
    <references>
      <span>
        <!--Toyota Motor-->
        <target id="t6"/>
        <target id="t7"/>
      </span>
    </references>
    <externalReferences>
      <externalRef reference="http://dbpedia.org
      /resource/Toyota" resource="spotlight_v1"/>
    </externalReferences>
  </entity>
```

```
<coref id="coentity1" type="person">
  <span>
    <!--Toyota motor-->
    <target id="t6"/>
    <target id="t7"/>
  </span>
  <span>
    <!--Toyota-->
    <target id="t198"/>
  </span>
</coref>
```

SEM IN TRIG FORMAT

EVENT INSTANCE

<nwr:data/cars/2013/1/1/5758-BPN1-F0J6-D2T2.xml#sellEvent>

a sem:Event , nwr:contextual , fn:Commerce_sell ;

rdfs:label "sell" ;

gaf:denotedBy

<nwr:data/cars/2013/1/1/5758-BPN1-F0J6-D2T2.xml#char=1352,1356&word=w251&term=t251> ,

<nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#char=1536,1540&word=w275&term=t275>.

SEM IN TRIG FORMAT

ENTITY INSTANCE

<http://dbpedia.org/resource/Toyota>

a sem:Actor , nwr:person , nwr:organization , nwr:framenet/Commerce_sell#Seller> ;

rdfs:label "Toyota" , "Toyota motor" ;

gaf:denotedBy

<nwr:data/cars/2013/1/1/5760-PM51-JD34-P4RM.xml#char=98,104&word=w18&term=t18> ,

<nwr:data/cars/2013/1/1/57K5-FKK1-DYBW-2534.xml#char=44934,44940&word=w8114&term=t8114> .

SEM RELATIONS AS NAMED GRAPHS

<nwr:/data/cars/2013/1/1/5758-BPN1-F0J6-D2T2.xml#pr25,rl55> {
  <nwr:data/cars/2013/1/1/5722-S821-F0J6-D48N.xml#sellEvent>
    sem:hasActor <http://dbpedia.org/resource/Magyar_Suzuki> .
}

<nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#pr46,rl114> {
  <nwr:data/cars/2013/1/1/5758-BPN1-F0J6-D2T2.xml#sellEvent>
    sem:hasPlace <http://dbpedia.org/resource/South_Africa> .
}
<nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#docTime_26> {
  <nwr:data/cars/2013/1/1/5760-PM51-JD34-P4H7.xml#sellEvent>
    sem:hasTime <nwr:time/2013-01-01> .
}

PROPERTIES OF RELATIONS

PROVENANCE

<nwr:data/cars/2013/1/1/57R8-5451-F0J6-D2GH.xml#pr25,rl55>

gaf:denotedBy

<nwr:data/cars/2013/1/1/57R8-5451-F0J6-D2GH.xml#rl55> ;

prov-o:wasAttributedTo

<nwr:sourceowner/Peru_Autos_Report> .

FACTUALITY

<nwr:data/cars/2013/1/1/57K5-FKK1-DYBW-2534.xml#facValue_1125> {

<nwr:data/cars/2013/1/1/57K5-FKK1-DYBW-2534.xml#sellEvent>

nwr:hasFactBankValue

"CT+" .}

CROSS-DOCUMENT EVENT COREFERENCE

- Instance based event-coreference:
  - All event mentions with same lemma and same time anchor
    - Share at least one actor (possibly DBPedia URI)
    - Share at least one place (possibly DBPedia URI)
- Aggregation of SEM instances from NAF mentions and the extraction of provenance layers through named graphs
- http://ic.vupr.nl/~ruben/vua-eventcoreference.ttl/

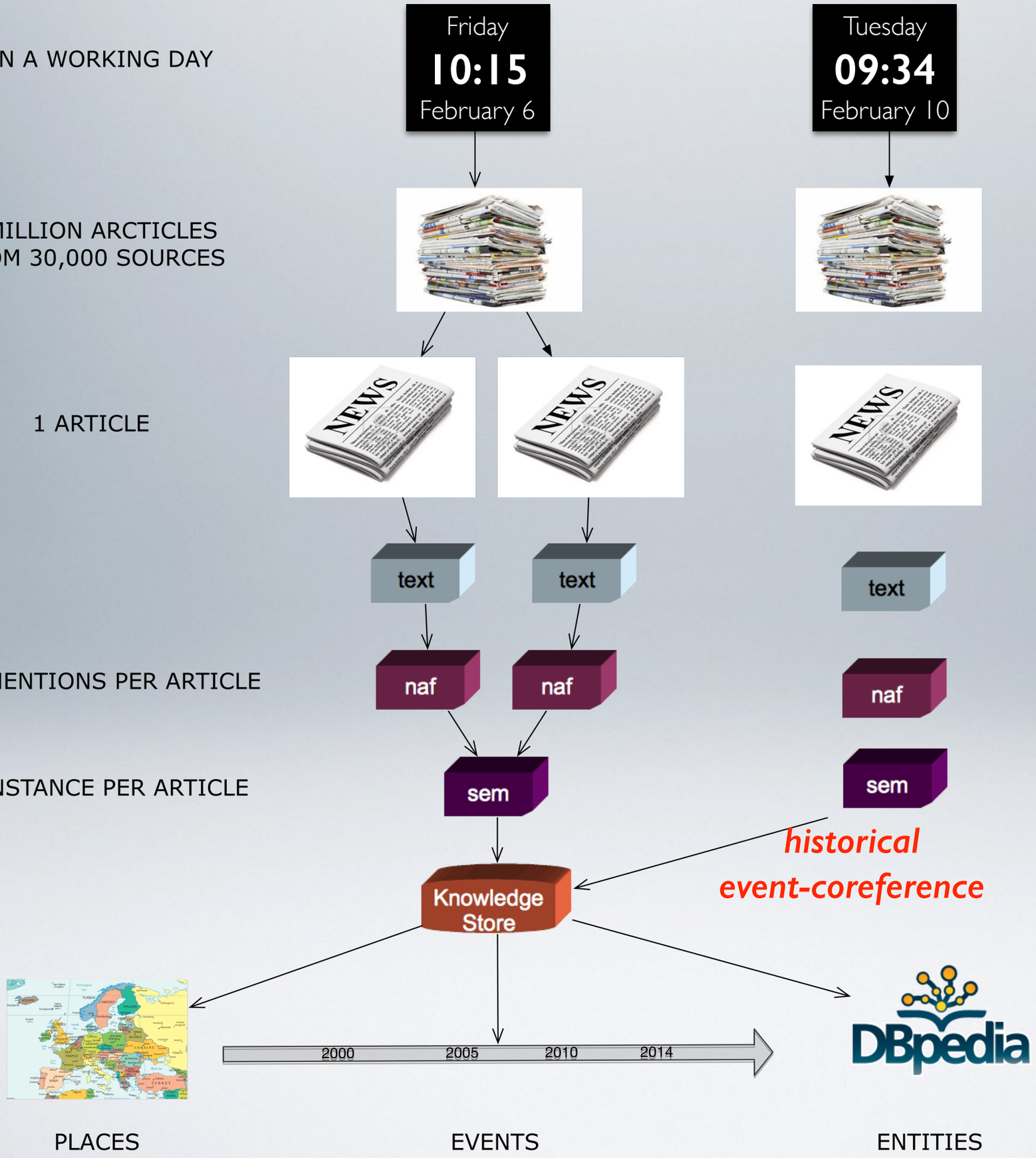
ON A WORKING DAY

1 MILLION ARCTICLES
FROM 30,000 SOURCES

1 ARTICLE

100 EVENT MENTIONS PER ARTICLE

5 EVENT INSTANCE PER ARTICLE



* *Virtual machines with 15 modules for English and Spanish*
* *Similar modules for Dutch and Italian*

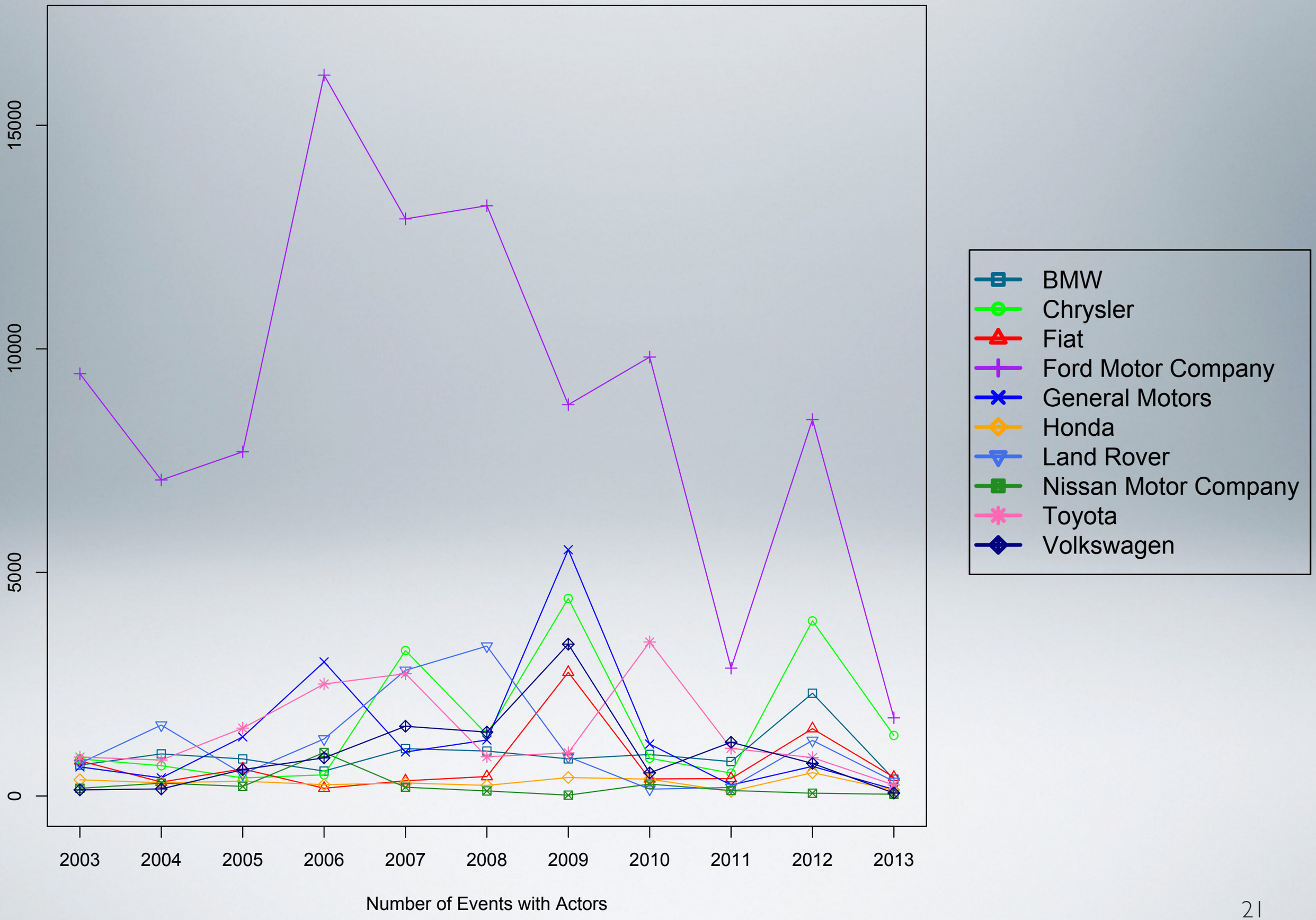
* *Combine all information across different mentions for a single instance*
* *Merge future information with previous information*

DATA FIRST YEAR

- **Car Industry news** (2003-2013): 63K articles, 1,7M event instances, 445K actors, 63K places, 41K DBpedia entities and 46M triples.
- **TechCrunch** (2005-2013): 43K articles, 1,6M event instances, 300K actors, 28K DBpedia entities and 24M triples.
- **WikiNews**: 19K English, 8K Italian, 7K Spanish and 1K Dutch. 69 Apple news documents for annotation.
- **ECB**+: 43 topics and 482 articles from GoogleNews, extended with 502 GoogleNews articles for 43+ topics (similar but different event).

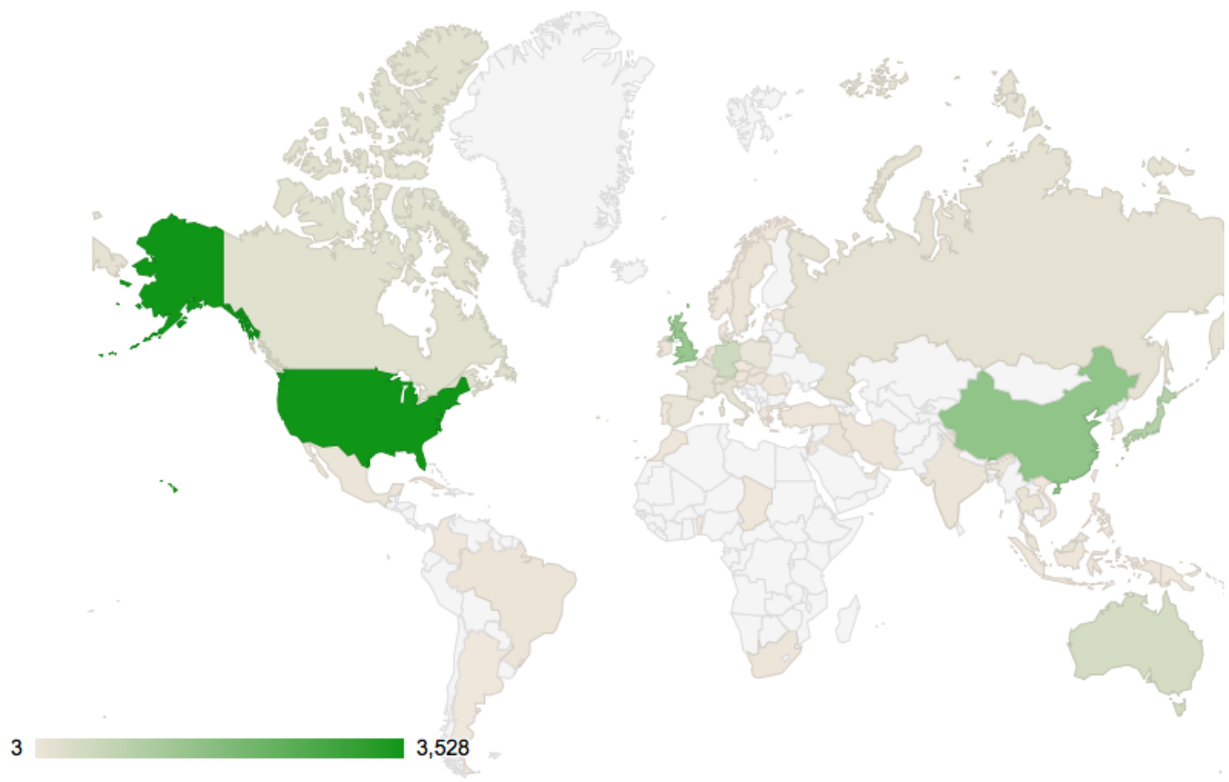
Table 2. Statistics on global automotive industry dataset

	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

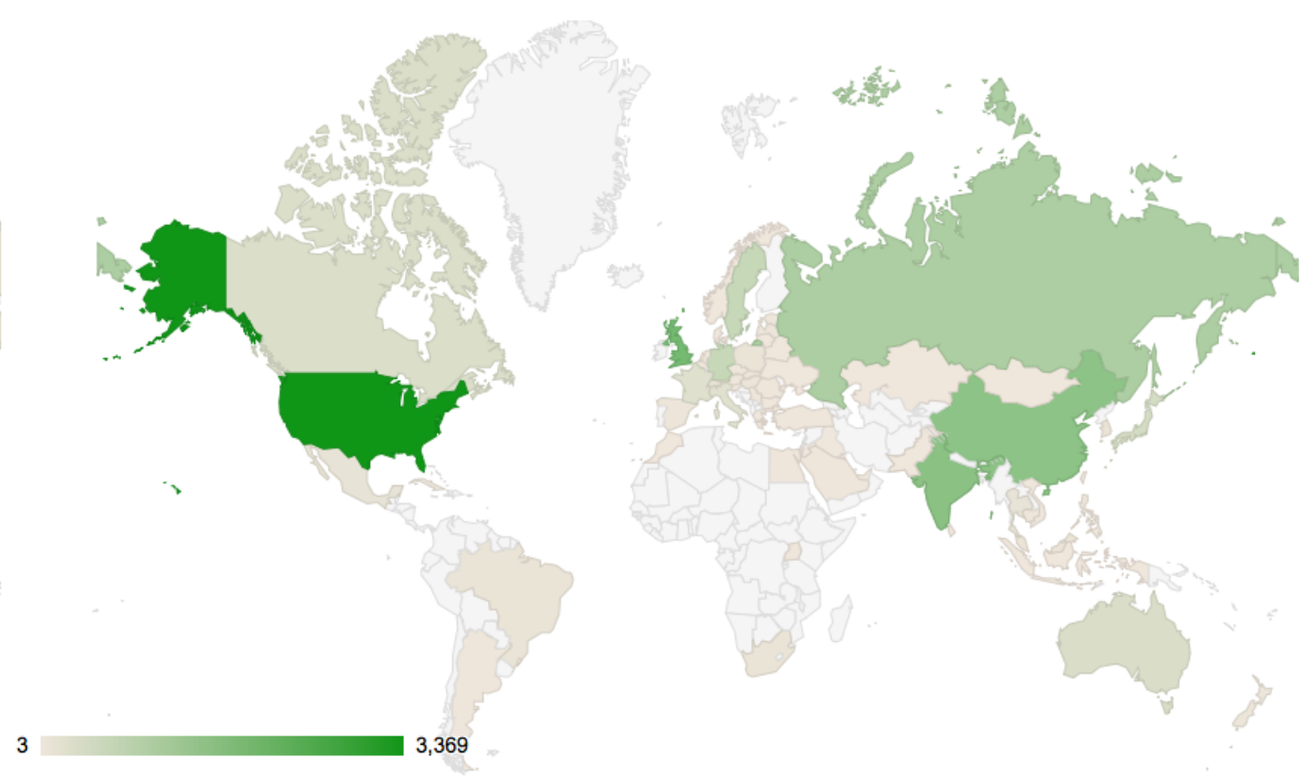

15000
10000
5000
0
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Number of Events with Actors
BMW
Chrysler
Fiat
Ford Motor Company
General Motors
Honda
Land Rover
Nissan Motor Company
Toyota
Volkswagen

CAR NEWS, WHERE & WHEN

2003



2008



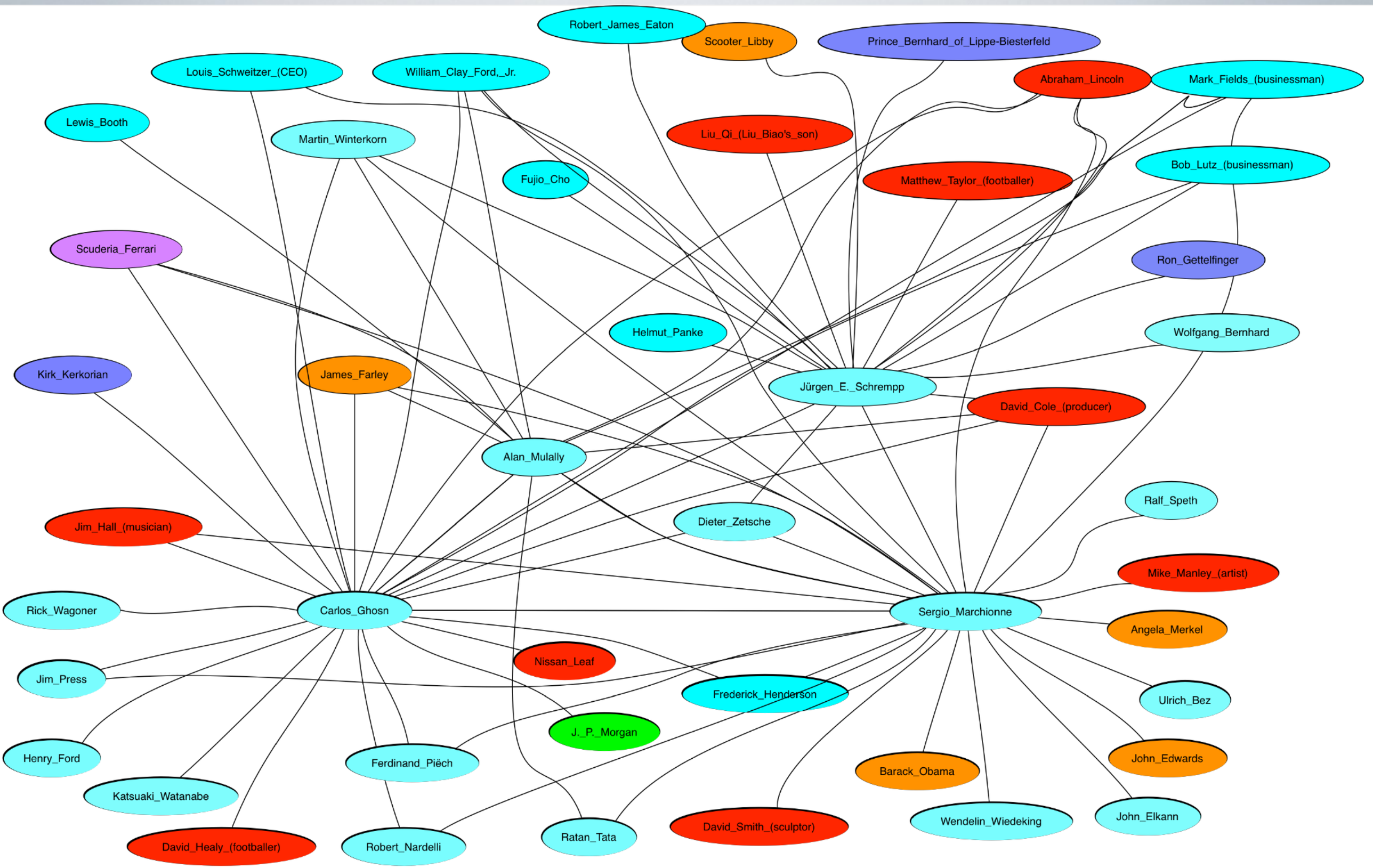


Fig. 3. The key actors' network in the global automotive industry domain

PROVENANCE STATISTICS

Source owner	Triples
Automotive_News	321,321
PR_Newswire	201,399
Detroit_Free_Press_(Michigan)	193,420
Just_-_Auto	167,735
Automotive_News_Europe	162,424
The_Associated_Press	160,911
just-auto_global_news	158,493
Associated_Press_Financial_Wire	151,971
The_Detroit_News_(Michigan)	150,383
The_Associated_Press_State_&_Local	129,248
etc.	…
TOTAL	12,851,504

CONCLUSIONS

- Derived large event graphs from massive streams of news in 4 languages.
- Handling the dynamic growth of information, separating the new from the old, the factual and the speculation.
- Represent provenance of information
- Storage in a KnowledgeStore
- Visualization of complex graph structures evolving in time

FUTURE WORK

- Benchmarking and evaluation of modules such as NER, event detection, coreference
- Extract relations between events and storylines
- Cross-lingual event extraction and interpretation
- End-user evaluation using interfaces to deal with large complex knowledge graphs
- Kick Off '@NewsReader' and Hack 100,000 World Cup Articles, Tuesday, June 10, 2014 LONDON

MORE NEWSREADER AT LREC

May 29:

- 12:25-12:45: Piek Vossen, German Rigau, Luciano Serafini, Pim Stouten, Francis Irving and Willem Van Hage. NewsReader: Recording History from Daily News Streams (O23 – Text Mining, Silfurberg B)
- 18:20 – 19:20: Marieke van Erp, Gleb Satyukov, Piek Vossen and Marit Nijssen. Discovering and Visualising Stories in News (P46 – Information Extraction and Information Retrieval, Poster area 1)
- 18.20 – 19.20: Christian Girardi, Manuela Speranza, Rachele Sprugnoli and Sara Tonelli. CROMER: a Tool for Cross-Document Event and Entity Coreference (P45 – Anaphora and Coreference, Poster area 1)

May 30:

- 11:45-13:25: Rodrigo Agerri, Josu Bermudez and German Rigau. IXA Pipeline: Efficient and Ready to Use Multilingual NLP Tools (P59 – Language Resource Infrastructures, Poster Area 1)
- 11:45 – 13:25: Chantal van Son, Marieke van Erp, Antske Fokkens and Piek Vossen. Hope and Fear: How Opinions Influence Factuality (P61 – Opinion Mining and Sentiment Analysis, Poster Area 1)
- 14.55 – 15:15: Giuseppe Rizzo, Marieke van Erp and Raphaël Troncy. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web (O48 – Information Extraction and Text Structure, Kaldalon)
- 15.55 – 16:15: Agata Cybulska and Piek Vossen. Using a Sledgehammer to Crack a Nut? (O46 – Event Extraction and Event Coreference, Silfurberg A)

http://www.newsreader-project.eu