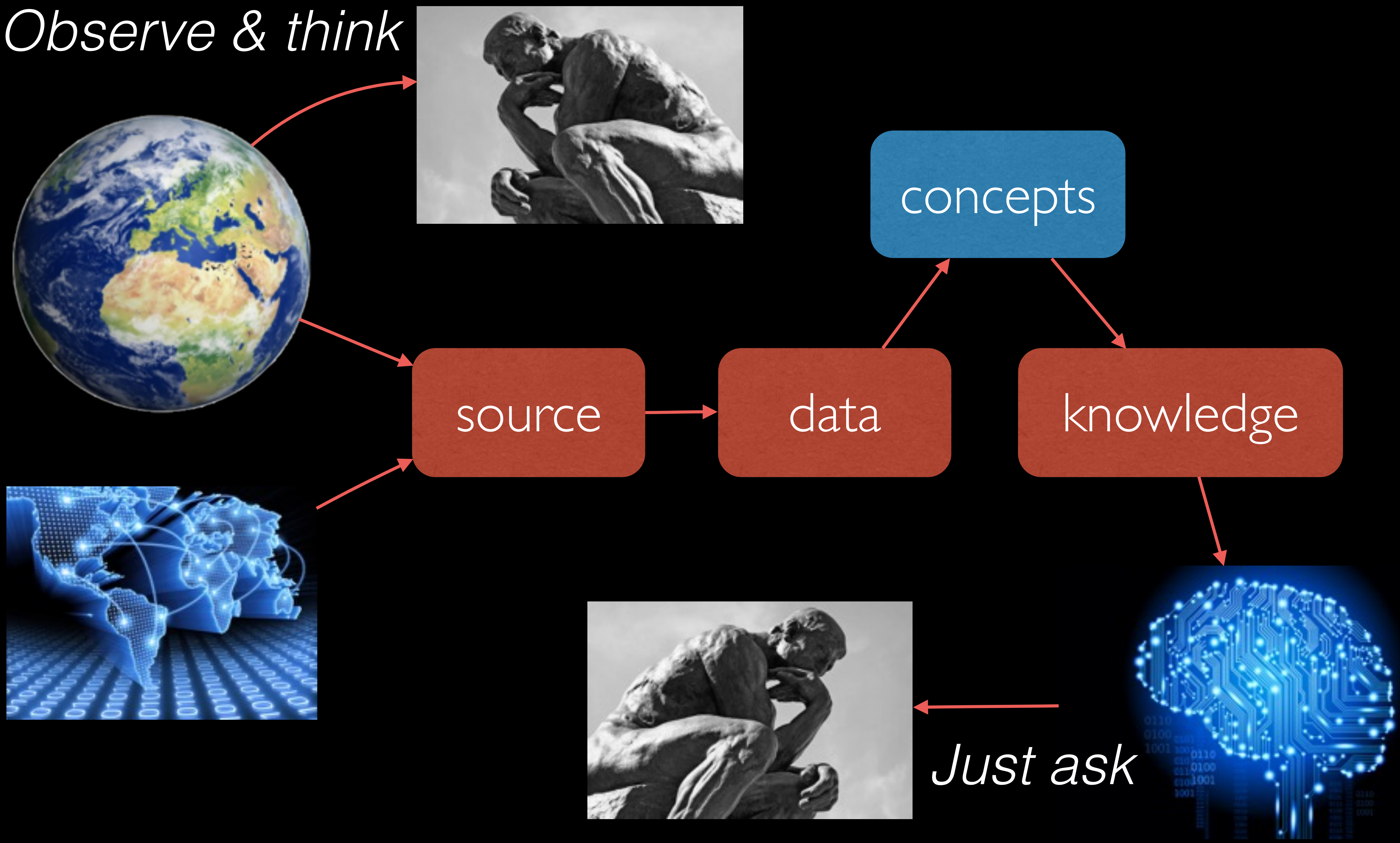
NewsReader
extracting event-centric knowledge graphs from massive news streams

FORUM LINKED DATA, SEP-2015, HILVERSUM

Piek Vossen

ICT 316404
FP7-ICT-2011-8
www.newsreader-project.eu

The changing world



Observe & think

source

data

concepts

knowledge

Just ask

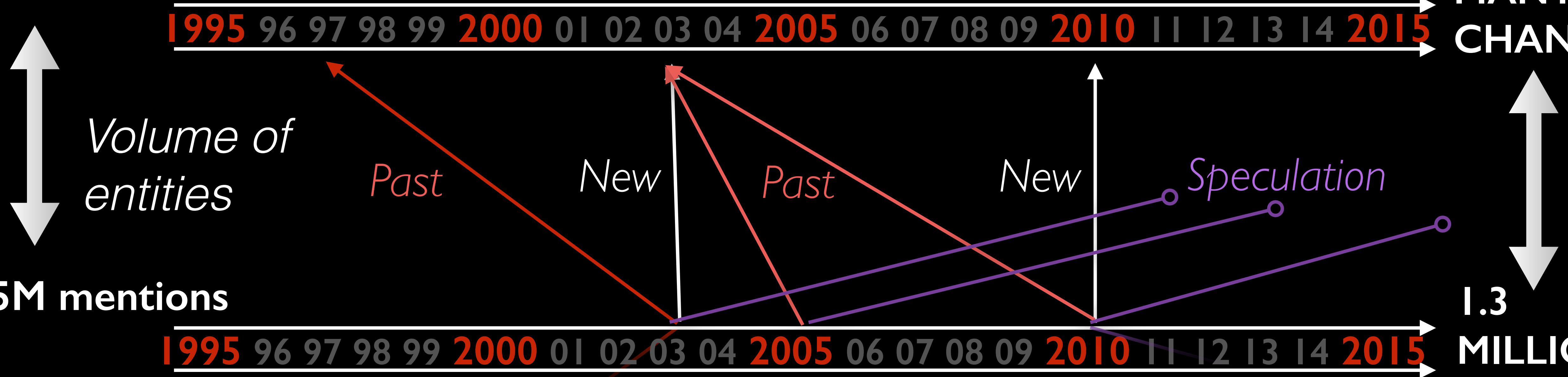
To learn about the changing world …read the news…

- But can we handle the news?
- Information broker LexisNexis:
  - 1.5 million news articles on a single working day
  - 30,000 different sources

VOLUME OF CHANGE

27M entities

HOW MANY CHANGES?



205M mentions

1.3 MILLION ARTICLES

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.

What if computers could read the news?

Who is Ford?

- President Woodrow Wilson asked ***Ford***[?] to run as a Democrat for the United States Senate from Michigan in 1918.

- Gerald Ford
- Ford, the motor company
- Henry Ford

Language:
how difficult can it be?

President [6] Woodrow Wilson [10] asked [7] Ford [8] to run [41] as a Democrat [2] for the United States [4] Senate [2] from Michigan [3] in 1918.

6*10*7*8*41*2*4*2*3= 6,612,480 combinations of word senses and entities

How to know Ford?

- President Woodrow Wilson asked ***Ford***[?] ***to run as*** a Democrat for the United States Senate from Michigan in 1918.
- Semantic knowledge:
  - meaning of ***to run as***
- World knowledge:
  - run as senator: +human, >18 years old, US citizen
  - Gerald Ford born in 1917, so 1 year old
  - Henry Ford the founder of Ford Motor Company, born in 1863, died in 1947, so 56 years old

Types of data

Concept centric

- All men are mortal
- Socrates is a man

Entity centric

- Socrates is mortal

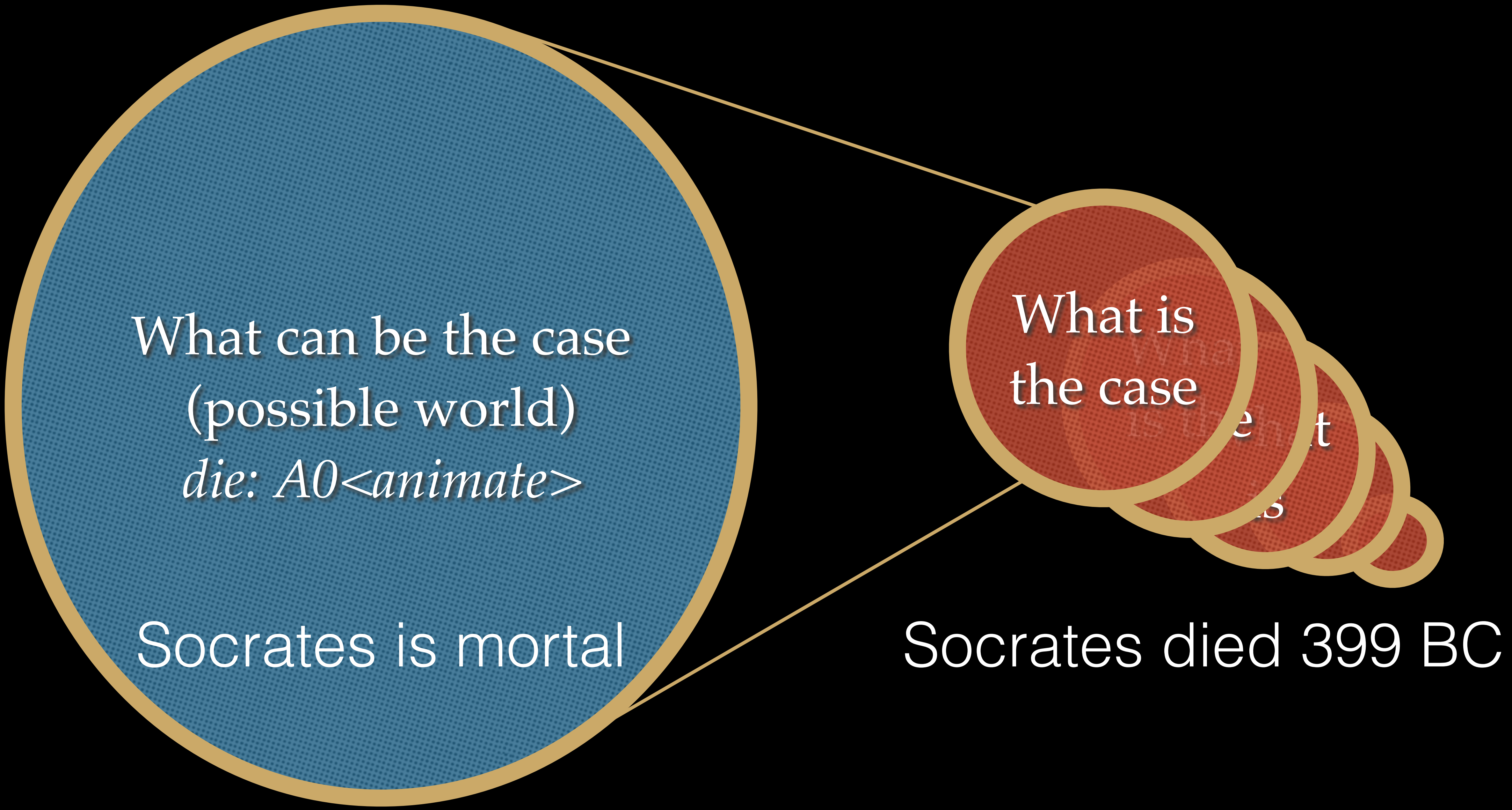
Event centric

- Socrates died in the year 399 BC

Semantic & Episodic knowledge

Concepts and relations
small data define a big world

Instantiations
many small worlds make big data

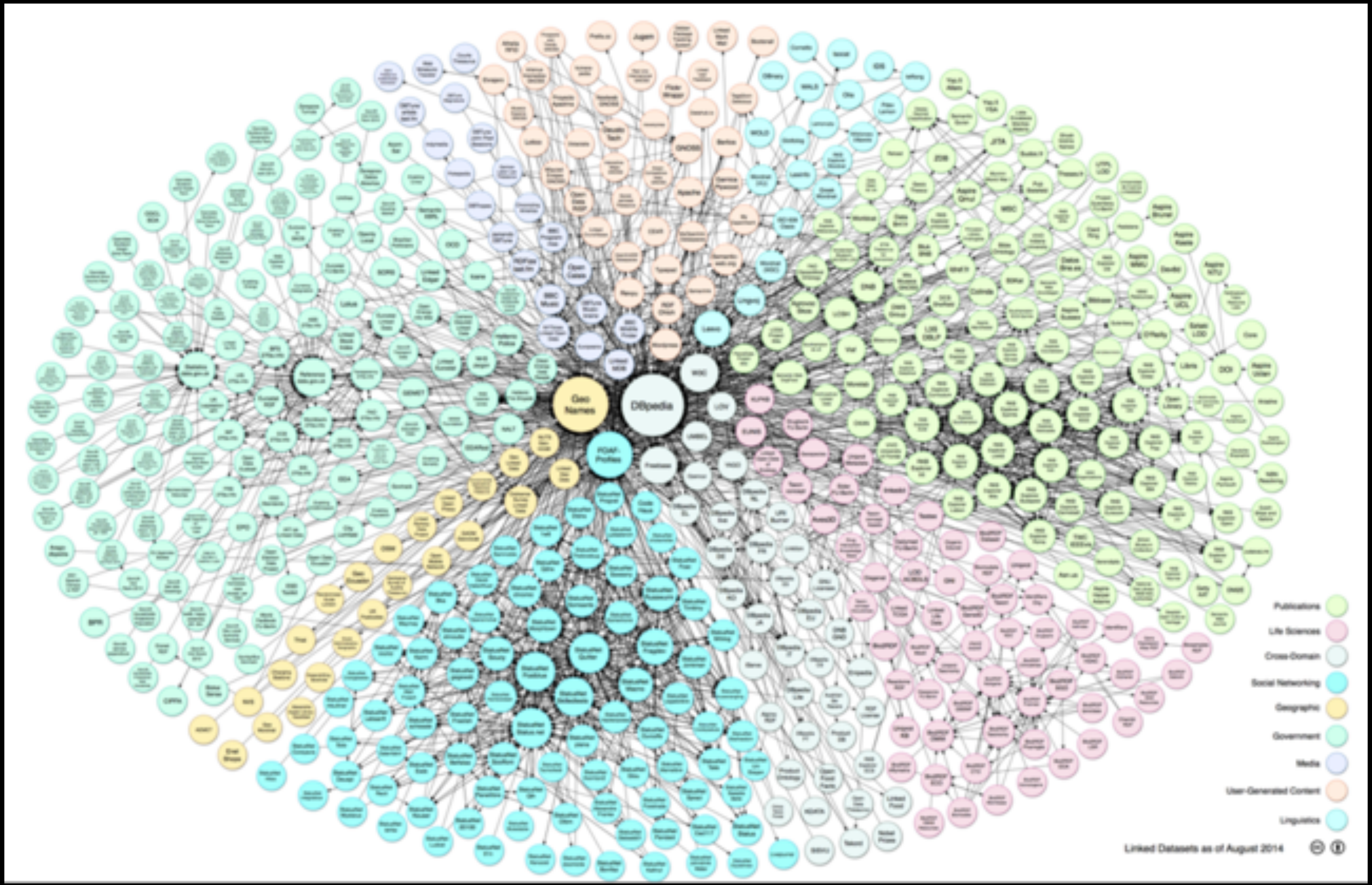


What can be the case
(possible world)
die: A0<animate>

Socrates is mortal

What is
the case

Socrates died 399 BC



World
of
Data

Real
World

LOD helps NLP

NLP helps LOD

NewsReader (ict316404)

- ***Reading 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.
  - Who made what statement, where do sources agree and disagree: ***provenance***!

http://www.telegraph.co.uk/finance/newsbysector/industry/engineering/10125280/Porsche-family-buys-back-10pc-stake-from-Qatar.html

17 Jun 2013

Porsche family buys back 10pc stake from Qatar

Descendants of the German car pioneer Ferdinand Porsche have bought back a 10pc stake in the company that bears the family name from Qatar Holding, the investment arm of the Gulf State’s sovereign wealth fund.

http://www.telegraph.co.uk/finance/newsbysector/industry/engineering/10125280/Porsche-family-buys-back-10pc-stake-from-Qatar.html

17 Jun 2013

Porsche family buys back 10pc stake from Qatar

MENTIONS

Descendants of the German car pioneer Ferdinand Porsche have bought back a 10pc stake in the company that bears the family name from Qatar Holding, the investment arm of the Gulf State's sovereign wealth fund.

INSTANCES

WHO:
http://dbpedia.org/page/Category:Porsche_family
http://dbpedia.org/page/Category:Qatarholding
http://dbpedia.org/page/Category:PorscheSE

http://www.telegraph.co.uk/finance/newsbysector/industry/engineering/10125280/Porsche-family-buys-back-10pc-stake-from-Qatar.html

17 Jun 2013

Porsche family buys back 10pc stake from Qatar

Descendants of the German car pioneer Ferdinand Porsche have bought back a 10pc stake in the company that bears the family name from Qatar Holding, the investment arm of the Gulf State’s sovereign wealth fund.

WHAT:
buy_back

MENTIONS

INSTANCES

http://english.alarabiya.net/en/business/banking-and-finance/2013/06/17/Qatar-Holding-sells-10-stake-in-Porsche-to-family-shareholders.html

Monday, 17 June 2013

Qatar Holding sells 10% stake in Porsche to founding families

Qatar Holding, the investment arm of the Gulf state’s sovereign wealth fund, has sold its 10 percent stake in Porsche SE to the luxury carmaker’s family shareholders, four years after it first invested in the firm.

http://english.alarabiya.net/en/business/banking-and-finance/2013/06/17/Qatar-Holding-sells-10-stake-in-Porsche-to-family-shareholders.html

Monday, 17 June 2013

Qatar Holding sells 10% stake in Porsche to founding families

Qatar Holding, the investment arm of the Gulf state’s sovereign wealth fund, has sold its 10 percent stake in Porsche SE to the luxury carmaker’s family shareholders, four years after it first invested in the firm.

MENTIONS

INSTANCES

WHO:
http://dbpedia.org/page/Category:Porsche_family
http://dbpedia.org/page/Category:Qatarholding
http://dbpedia.org/page/Category:PorscheSE

http://english.alarabiya.net/en/business/banking-and-finance/2013/06/17/Qatar-Holding-sells-10-stake-in-Porsche-to-family-shareholders.html

Monday, 17 June 2013

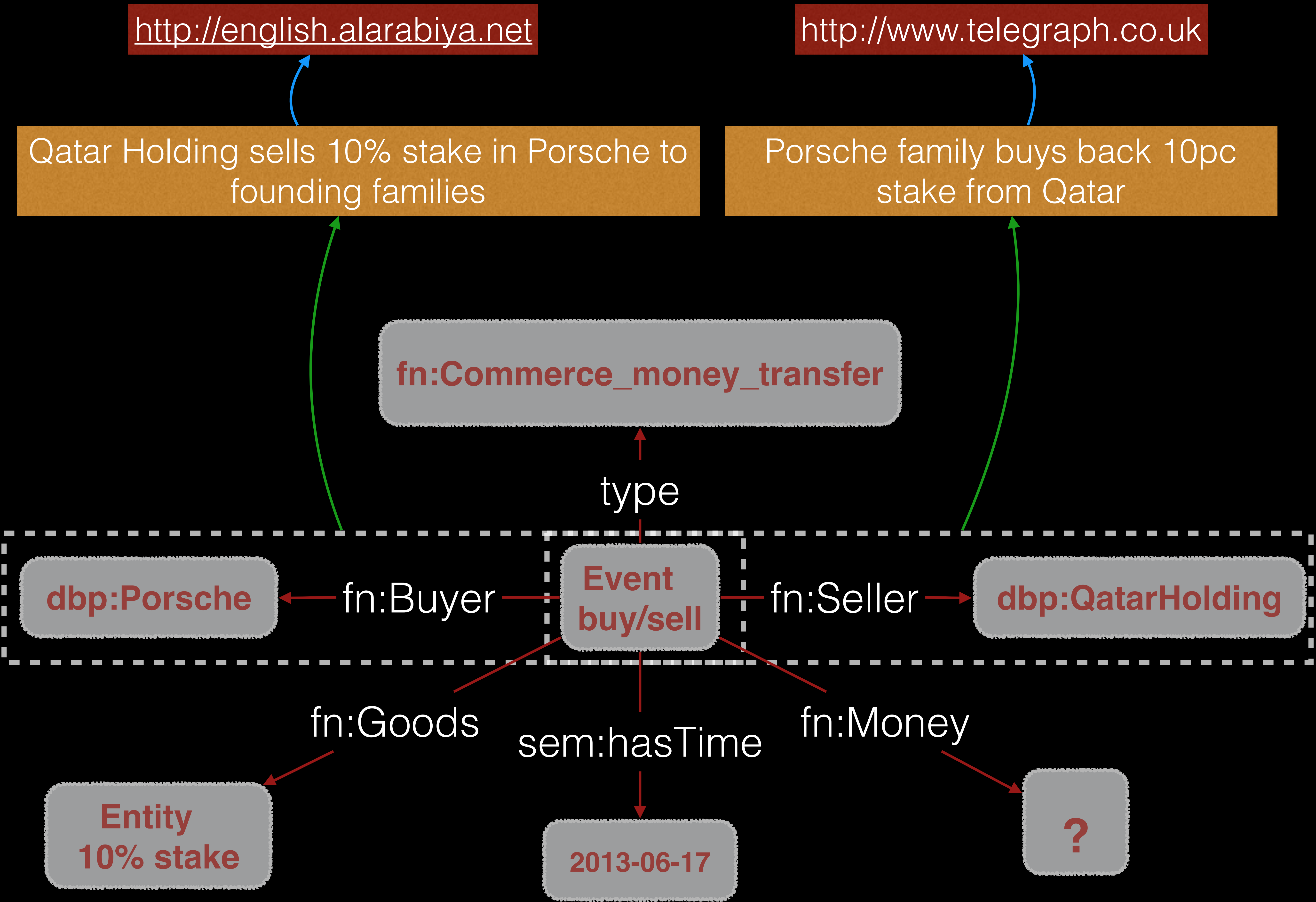
Qatar Holding sells 10% stake in Porsche to founding families

Qatar Holding, the investment arm of the Gulf state’s sovereign wealth fund, has sold its 10 percent stake in Porsche SE to the luxury carmaker’s family shareholders, four years after it first invested in the firm.

MENTIONS

INSTANCES

WHAT:
sell



http://english.alarabiya.net

http://www.telegraph.co.uk

Qatar Holding sells 10% stake in Porsche to founding families

Porsche family buys back 10pc stake from Qatar

fn:Commerce_money_transfer

type

dbp:Porsche

fn:Buyer

Event buy/sell

fn:Seller

dbp:QatarHolding

fn:Goods

sem:hasTime

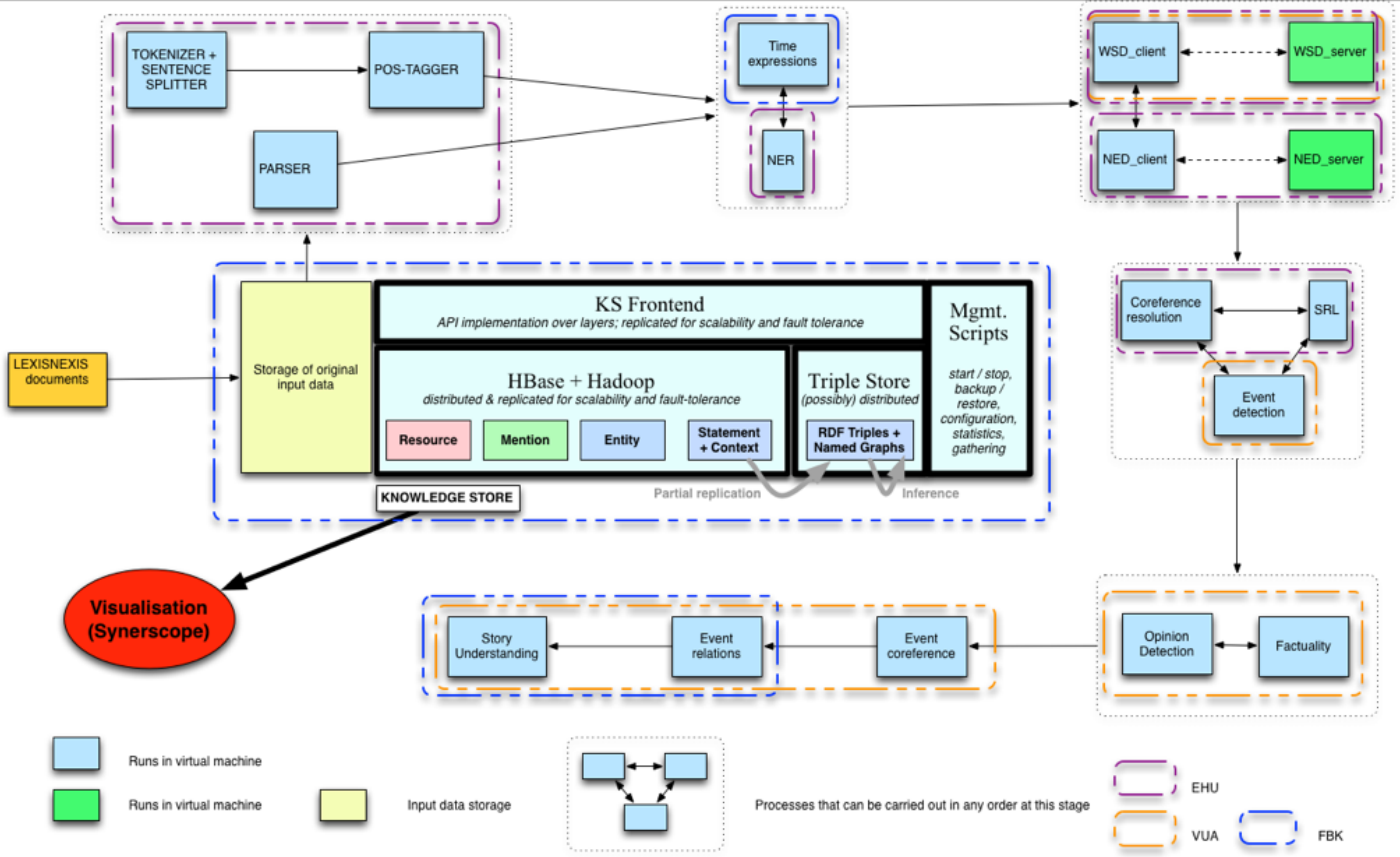
fn:Money

Entity 10% stake

2013-06-17

?

System Architecture



TOKENIZER + SENTENCE SPLITTER

POS-TAGGER

PARSER

Time expressions

NER

WSD_client

WSD_server

NED_client

NED_server

LEXISNEXIS documents

Storage of original input data

KS Frontend

API implementation over layers; replicated for scalability and fault tolerance

HBase + Hadoop

distributed & replicated for scalability and fault-tolerance

Resource

Mention

Entity

Statement + Context

Triple Store

(possibly) distributed

RDF Triples + Named Graphs

Mgmt. Scripts

start / stop, backup / restore, configuration, statistics, gathering

Partial replication

Inference

KNOWLEDGE STORE

Coreference resolution

SRL

Event detection

Visualisation (Synerscope)

Story Understanding

Event relations

Event coreference

Opinion Detection

Factuality

Runs in virtual machine

Runs in virtual machine

Input data storage

Processes that can be carried out in any order at this stage

EHU

VUA

FBK

SEM in RDF-TriG format

dbo:Agent,Company
dbr:Privately_held_company
schema.org/Organization

ENTITY INSTANCE

```
<http://dbpedia.org/resource/PorscheSE>
        rdfs:label    "Porsche" , “Porsche company" ;
        gaf:denotedBy
            <nwr:data/cars/2013/1/1/5760-PM51-JD34-
            P4RM.xml#char=98,104> ,
            <nwr:data/cars/2013/1/1/57K5-FKK1-
            DYBW-2534.xml#char=44934,44940> .
```

SEM in RDF-TriG format

EVENT INSTANCE

```
<nwr:data/cars/2013/1/1/5758-BPN1-F0J6-D2T2.xml#sellEvent>
        a               fn:Commerce_sell , fn:Commerce_buy;
        rdfs:label      "sell" , “buy”;
        gaf:denotedBy
          <nwr:data/cars/2013/1/1/5758-BPN1-F0J6-
          D2T2.xml#char=12,15> ,
          <nwr:data/cars/2013/1/1/5758-BPN1-F0J6-
          D2T2.xml#char=1352,1356> ,
          <nwr:data/cars/2013/1/1/5760-PM51-JD34-
          P4H7.xml#char=1536,1540>.
```

Semantic 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/Qatar_Holding> .
}
<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/Germany> .
}
<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:hasFutureTime  <nwr:time/2016> .
}
```

predicate *object* *subject*

Properties of relations

ATTRIBUTION

```
<nwr:data/cars/2013/1/1/57K5-FKK1-DYBW-2534.xml#pr27,rl57,rl56> {
   <nwr:data/cars/2013/1/1/57K5-FKK1-DYBW-2534.xml#sellEvent>
         gaf:hasAttribution
            "UNCERTAIN-NEG-FUTURE" .}

<nwr: data/2013/06/23/58T2-K531-JCDY-Y0X2.xml#pr27,rl57,rl56>
        prov:wasAttributedTo  <nwr:data/cars/entities/ErikGottfried> .
```

PROVENANCE

```
<nwr:data/cars/2013/1/1/57R8-5451-F0J6-D2GH.xml#pr27,rl57,rl56>
     gaf:denotedBy
        <nwr:data/cars/2013/1/1/57R8-5451-F0J6-D2GH.xml#char=1352,1356> ;
     prov-o:wasAttributedTo
        <nwr:sourceowner/Peru_Autos_Report> .
```

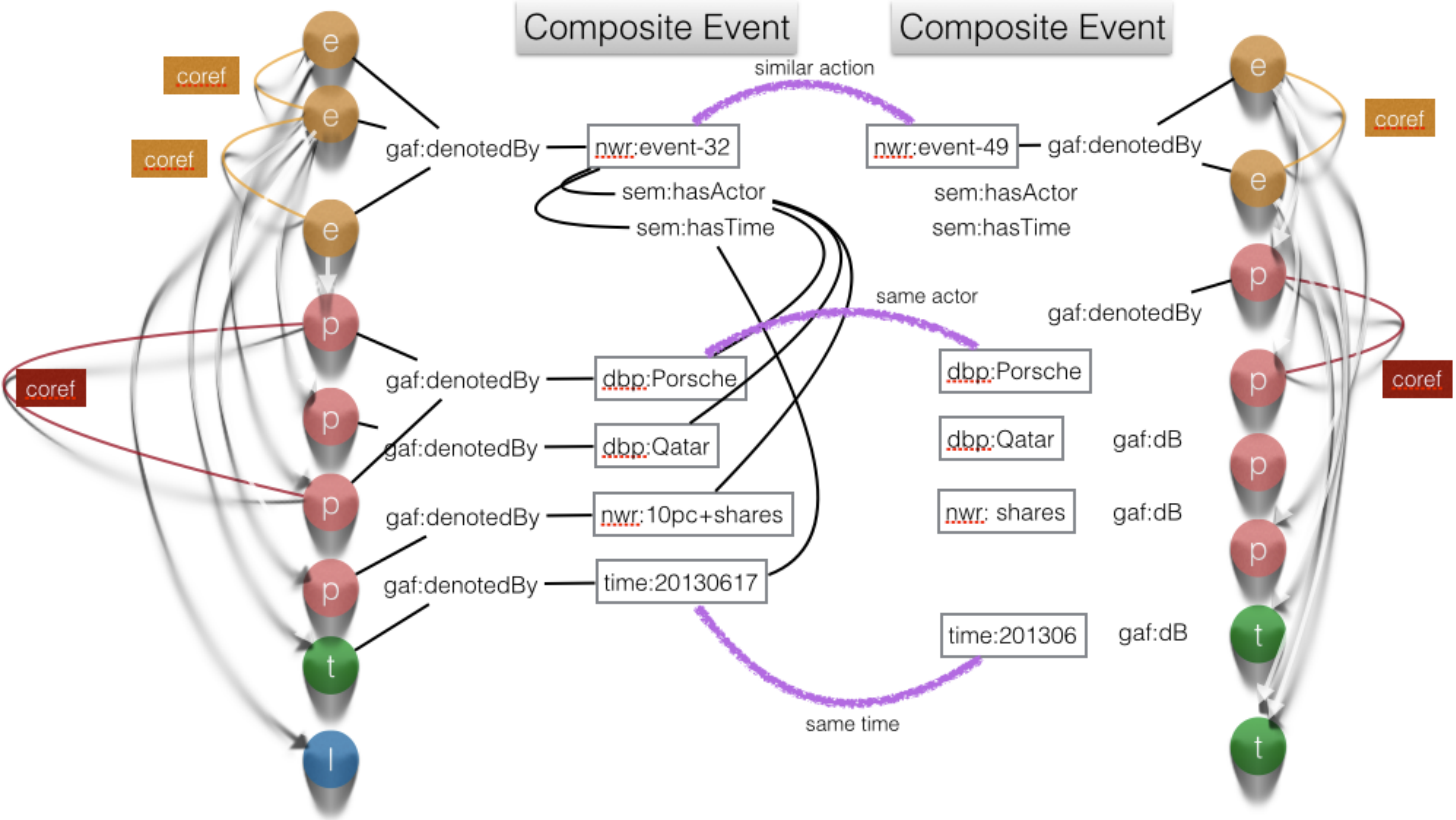
A *maiden flight* in the news

- The Airbus A380, the world's largest passenger plane, was set to land in the United States of America on Monday after a test flight. One of the A380s is flying from Frankfurt to Chicago via New York; the airplane will be carrying about 500 people.

- It is being billed as the first time it has carried a near-normal number of passengers, though most will be staff of Airbus and German airline Lufthansa.

- A second A380 is also travelling to the U.S. on Monday, but without passengers. This will be branded as a Qantas flight and fly from Frankfurt to Los Angeles LAX airport. The first leg of the flight going towards New York will be travelling under a Lufthansa flight number, and is due to arrive at New York's John F. Kennedy airport at 12:30 EST (16:30 UTC).

NAF mentions SEM instances NAF mentions



Composite Event

Composite Event

similar action

coref

coref

gaf:denotedBy — nwr:event-32

nwr:event-49 — gaf:denotedBy

sem:hasActor

sem:hasTime

sem:hasActor

sem:hasTime

same actor

gaf:denotedBy

coref

gaf:denotedBy — dbp:Porsche

dbp:Porsche

coref

gaf:denotedBy — dbp:Qatar

dbp:Qatar gaf:dB

gaf:denotedBy — nwr:10pc+shares

nwr: shares gaf:dB

gaf:denotedBy — time:20130617

time:201306 gaf:dB

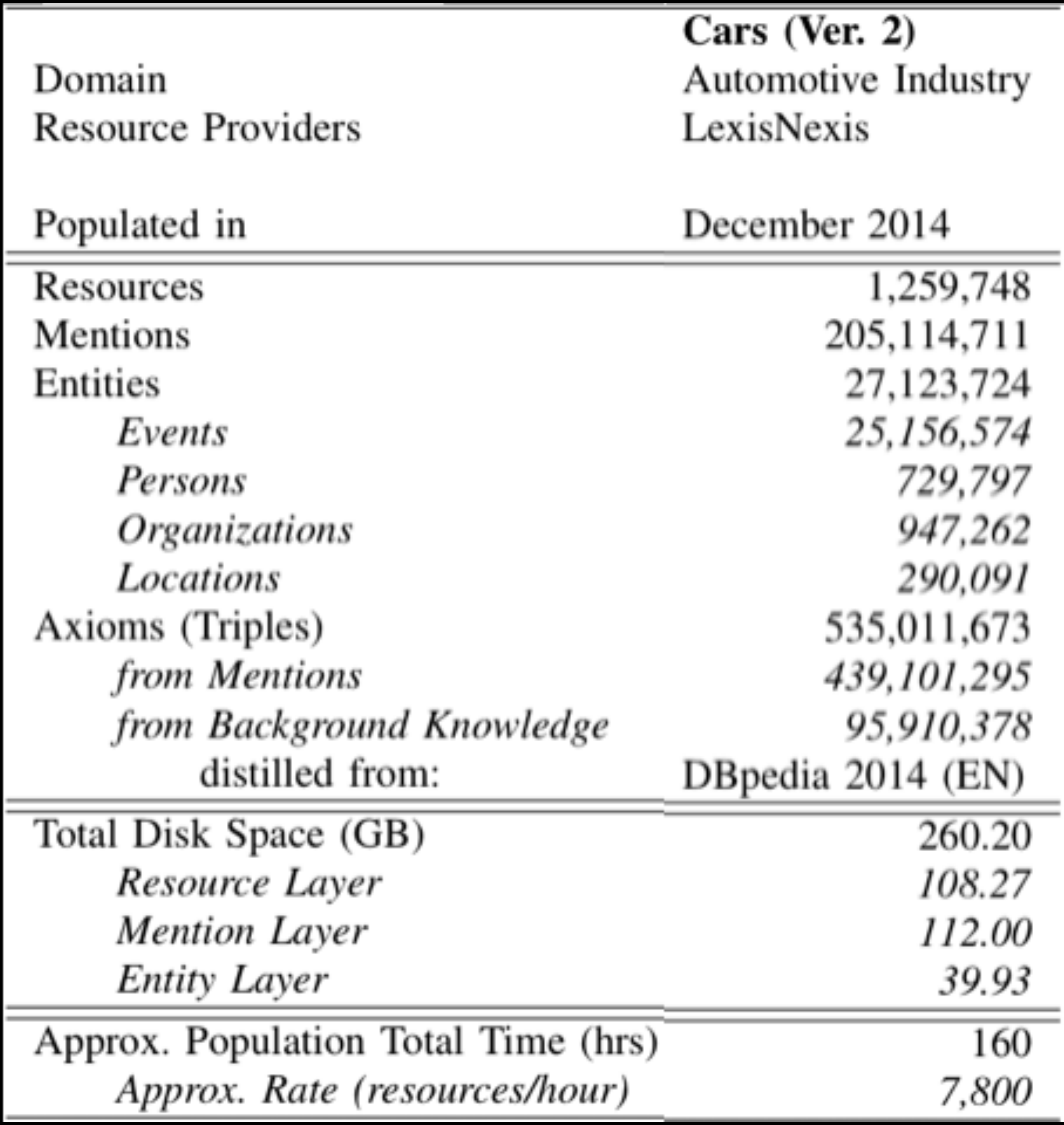
same time

buy/acquire/acquisition **sell/sales**

Use cases

- Automotive industry, English, 10 years, 1.3M articles
- Wikinews 18K English, 8k Spanish, 7k Italian, 1k Dutch
- Fifa world-cup, 212K English
- Human, drug, animal trafficking, 30 years, 900K English
- Dutch House of Representatives, Bank crisis enquiry, 700K Dutch
- Spanish Ministry: ???K Spanish

Automotive industry 2003-2013



	Cars (Ver. 2)
Domain	Automotive Industry
Resource Providers	LexisNexis
Populated in	December 2014
Resources	1,259,748
Mentions	205,114,711
Entities	27,123,724
Events	*25,156,574*
Persons	*729,797*
Organizations	*947,262*
Locations	*290,091*
Axioms (Triples)	535,011,673
from Mentions	*439,101,295*
from Background Knowledge	*95,910,378*
distilled from:	DBpedia 2014 (EN)
Total Disk Space (GB)	260.20
Resource Layer	*108.27*
Mention Layer	*112.00*
Entity Layer	*39.93*
Approx. Population Total Time (hrs)	160
Approx. Rate (resources/hour)	*7,800*

KnowledgeStore demos

- Video: Video: http://youtu.be/if1PRwSII5c
- https://knowledgestore2.fbk.eu/nwr/wikinews/ui
- SELECT ?s WHERE {?s rdf:type framenet:Commerce_sell . ?s sem:hasActor dbpedia:Airbus .} LIMIT 10
- Reasoning Module: https://knowledgestore2.fbk.eu/reasoner/

NewsReader platform

- GAF, NAF & SEM-RDF to model semantic interpretation across documents and across languages
- Pipelines in 4 languages (15 NLP modules each) available through source code (Apache license) and virtual machines
- Storm and Hadoop installations for fast and robust parallel processing, tested on millions of news articles and different uses cases
- KnowledgeStore for storing and querying sources, NAF and resulting triples in combination with background knowledge and supported reasoning
- Public benchmark and evaluation corpora in 4 languages manually annotated for rich and complex semantics