Linked Open Data & Delph-In

Antske Fokkens
VU University Amsterdam

Overview

- A short introduction to Linked Data
- What we are doing with Linked Data
  - Representation
  - Provenance & Perspectives
- What I would like to do with Delph-In

Linked Data

- The Semantic Web is a web of data
- Semantic Web technology allows us to query this data
- For this to be useful, we need a huge amount of data available on the Web in standard format
- If we want to increase knowledge about data, we should also make **relations** between data explicit (adapted from http://www.w3.org/standards/semanticweb/data)

Tim Berners-Lee's note on Linked Data

``The Semantic Web isn't just about putting data on the web. It's about making links, so that a person or machine can explore the web of data"

1. Use URIs for names of things

2. Use HTTP URIs so that people can look up those names

3. When someone looks up a URI, provide useful information (using the standards RDF*, SPARQL)

4. Include links to other URIs, so they can find more things

Linked data and NLP

- Linked Data provides datasets with unique identifiers, which can be useful in task such as named entity recognition

- Research on improving technology to extract information from text and provide RDF conform representations of this data

From text to Linked Data

- BiographyNet (VUA & Dutch partners):
  - gather information from biographical dictionaries and build an interface for historic research
- NewsReader (VUA, Basque University, FBK Trento, Lexis Nexis, Scraperwiki, Synerscope):
  - a ``history recorder" that monitors news and directly links new news items to ``old news" from the last decades

Goals

- Get more complete information about people, companies and events
- Find and represent alternative views on events
- Use what is known to improve textual interpretation
- Investigate methodological issues of NLP application used by researchers from other domains

The Grounded Annotation Framework (GAF)

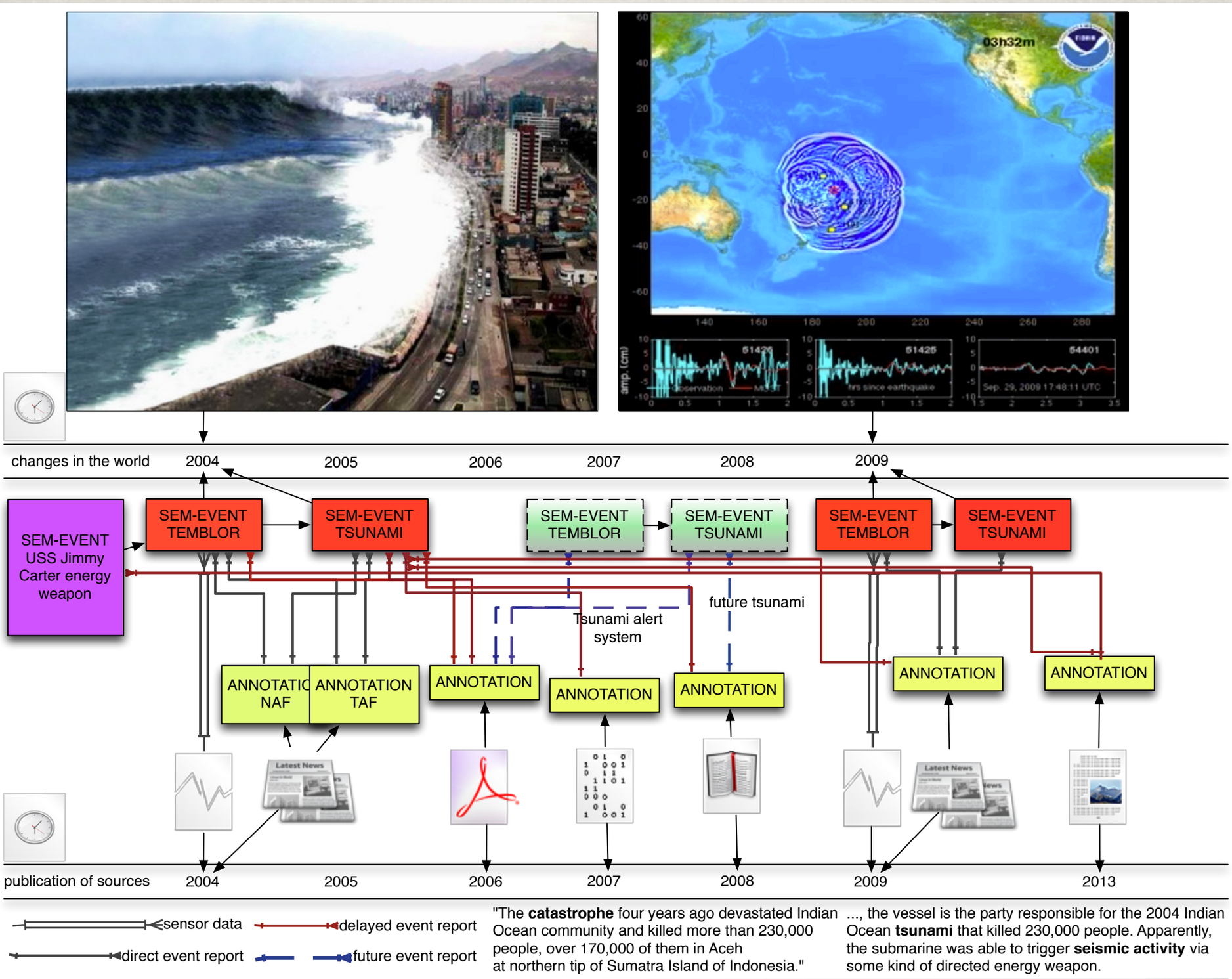
- Representation that can represent information from text and extra-textual sources

- Clear distinction between **mentions** in text and their formal representations as **instances** in a semantic layer

- Instances are represented in RDF compliant URIs

- Compatible with various linguistic annotations (we'll use the NLP Annotation Format)

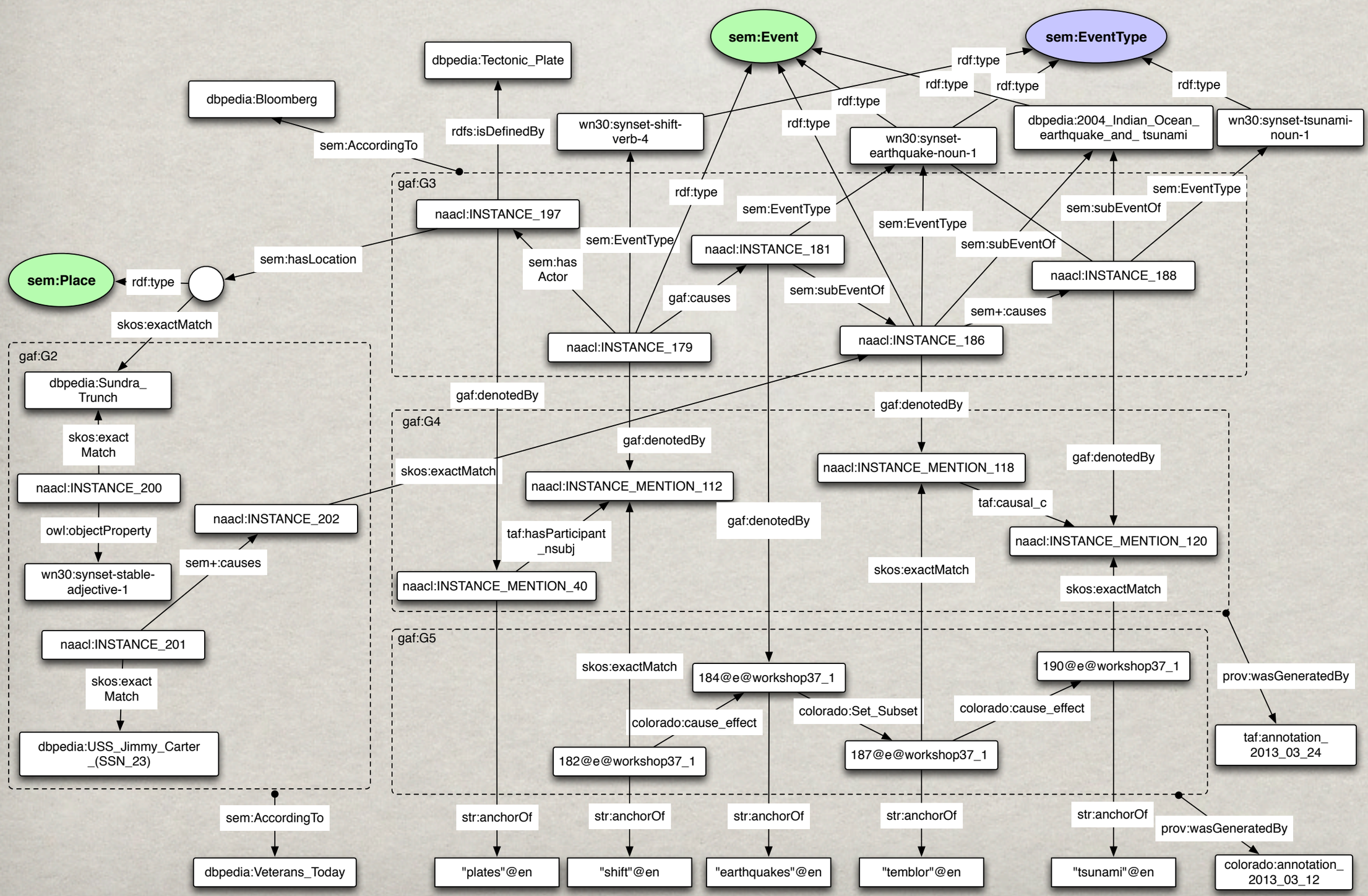
A GAF Illustration



"The **catastrophe** four years ago devastated Indian Ocean community and killed more than 230,000 people, over 170,000 of them in Aceh at northern tip of Sumatra Island of Indonesia."

..., the vessel is the party responsible for the 2004 Indian Ocean **tsunami** that killed 230,000 people. Apparently, the submarine was able to trigger **seismic activity** via some kind of directed energy weapon.

A GAF REPRESENTATION



Perspectives

- We want to represent different perspectives
  - How do different source represent the same event?
  - Can we monitor shifts in opinion (a governor of the Dutch Indies described in the 19th or 20th century)
  - What is the influence of the NLP approaches taken on the ultimate results?

Provenance

- If we represent different perspectives, we need to indicate where they came from

- We look at provenance at two (main) levels:

1. Who or what claimed or stated something

2. How did we derive this information?

Modeling NLP Provenance

- GAF is compatible with the PROV-DM
- This provenance model allows us to model:
  - which processes were involved
  - carried out by which agent
  - derived from which dataset
  - ....
  -

Research with provenance

- A clear indication of tools, versions, datasets and intermediate datasets helps to make research reproducible

- A clear overview of the overall process helps us to perform systematic tests on the performance of different setups

- Provenance information can be useful when combining the output of alternative tools

Delph-in & LOD (1)

- Can provenance modeling play a role in the work that has been done on:
  - Supertagging
  - Pruning
  - Parse ranking
  - CFG backbone
  - ...

Delph-in & LOD (2)

- Current approach to obtaining RDF conform representations from text:
    - PoS tag the text
    - Apply a NE recognition and NE determiner
    - Run a parser
    - All verbs are events, all NE that are related to a verb by a dependency parser are its participants, location or time

Delph-in and LOD (2)

- I think we can do better...
- I'd like to see if we can derive RDF conform URIs from the output of our parsers

  (we can, of course, but it is a bit of work to do this properly)
- Linking lexical entries to WordNet could be a nice start :-)

References

- http://groundedannotationframework.org/

- Fokkens, Antske, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli and Jesper Hoeksema (2013) GAF: A Grounded Annotation Framework for Events. *Proceedings of the first Workshop on EVENTS: Definition, Detection, Coreference and Representation*. Atlanta, USA. [poster][.bib]

- http://linkeddata.org/

- http://www.w3.org/standards/semanticweb/data

- http://www.w3.org/DesignIssues/LinkedData.html

Acknowledgements

- Supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404)

- Supported by the BiographyNet project (Nr. 660.011.308) funded by the Netherlands eScience Center (http://escience.center.nl). Partners in this project are the Netherlands eScience center, the Huygens ING Institute of the Royal Dutch Academy of Science and VU University Amsterdam