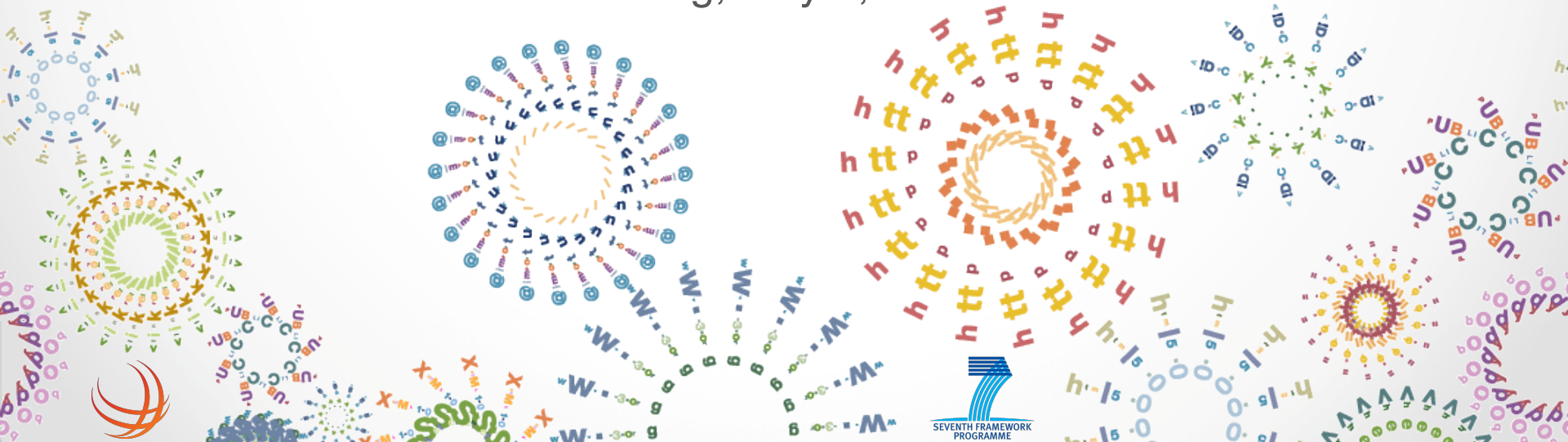
Longitudinal Analytics of Web Archive Data

# Methods and Tools for Temporal Web Analytics

Marc Spaniol

Ålborg, May 9, 2012



# LAWA in a Nutshell

- Key objectives
  - Web-scale data provisioning and access
  - Providing an infrastructure required to make research at Internet scale
  - Federation of distributed FIRE facilities with a centralized Web repository
- Consortium
  - Max-Planck-Institut für Informatik (Gerhard Weikum)
  - Hebrew University of Jerusalem (Scott Kirkpatrick)
  - Internet Memory Foundation (Julien Masanès)
  - Hungarian Academy of Science (Andras Benczur)
  - University of Patras (Peter Triantafillou)
  - Hanzo Archives Ltd. (Mark Middleton)
- Start: 9/2010
- Duration: 36 month
- [www.lawa-project.eu](http://www.lawa-project.eu)

# Big Data Analytics in LAWA

## Scientific:

- Online media, open knowledge
- Web history as an asset

- Needs **Web-scale analytics** along **temporal** dimension
  - New **algorithms**
  - Easy-to-use **tools**
  - **Scalable platform**

- Needs **instrumentation** and **performance** studies

## Infrastructure:

- Virtual Web observatory
- Web archives easily accessible

- Pursued via **public testbed** for **experiment**-driven research:
  - **Reference collection**
  - **Added-value** content
  - **Services** for analysis
  - **User group**

- Geared for **extensibility**, own user data, new tools

⇒ Benefits for science, society, business

# LAWA Methods and Tools for Temporal Web Analytics

- Distributed access to large scale data sets
  - Wide area operations
  - Heterogeneous distributed indices

→ LAWA Reference Collection

- Distributed temporal Web analytics
  - Distributed Web data aggregation, querying and ranking
  - Entity detection, disambiguation and tracking → AIDA
  - Interesting phrase mining
  - Text-entity-time analytics → YAGO
  - Community detection and analytics

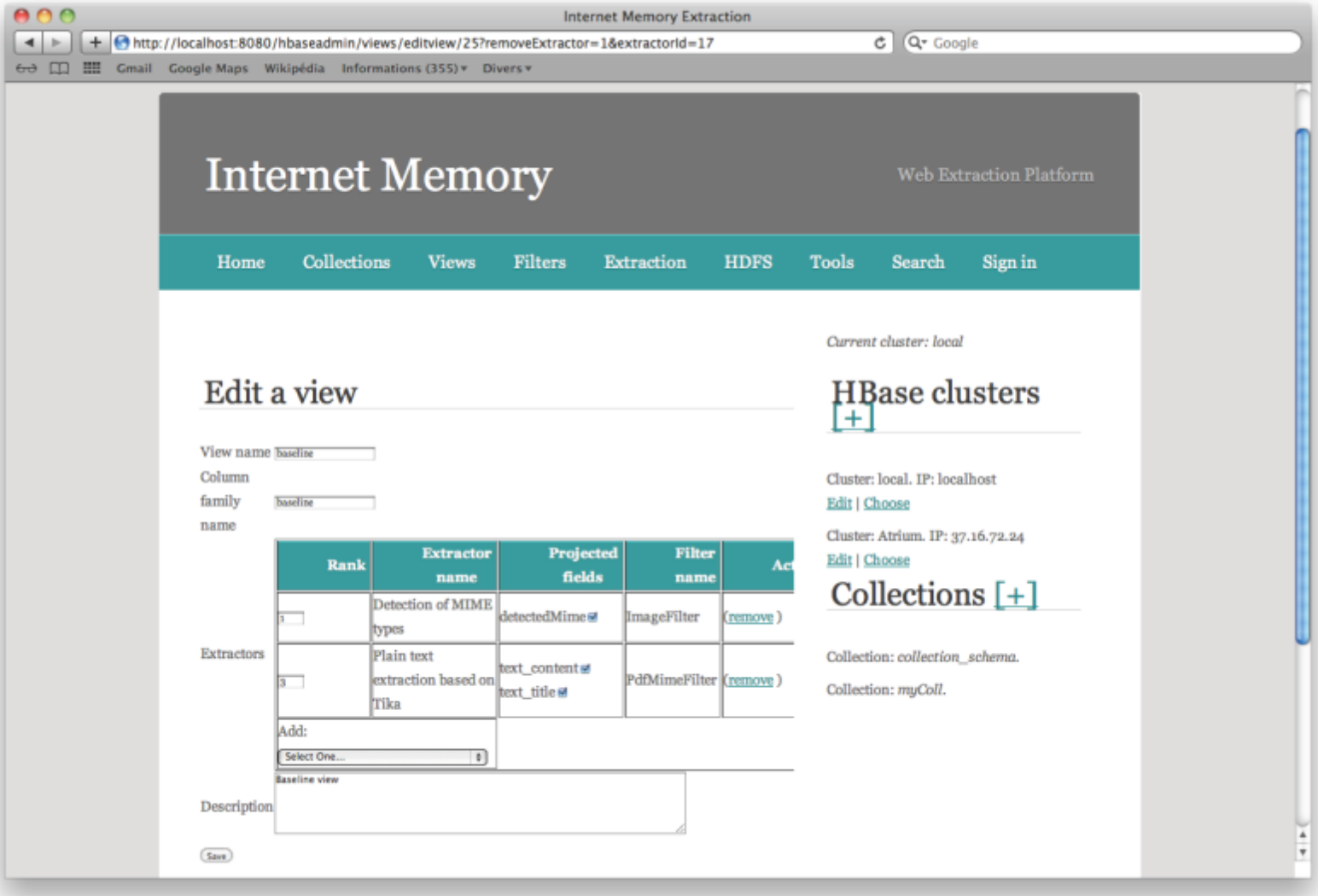
- Studies on Web-scale data
  - Measurement, mining, and classification services
  - Experiment-driven research on large-scale

→ LAWA Experimental Testbed

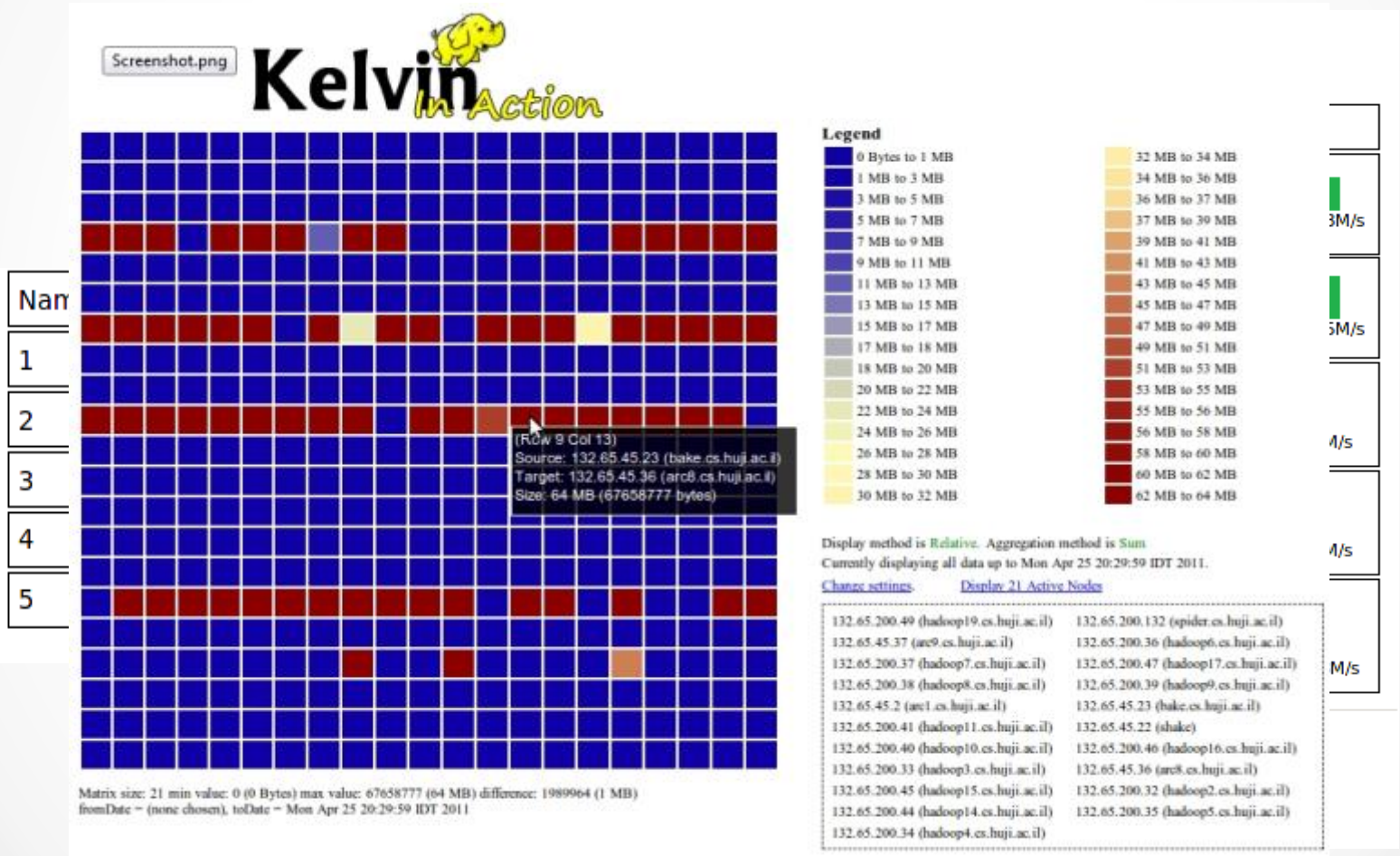
# LAWA Reference Collection

- UK Web content collection
  - Dating back until 1998
  - More than 100 crawl instances for selected sites
- Broad coverage
  - Government sites
  - Health portals
  - Education
  - Science
  - Finance
- Deployed in HBase for analytics within LAWA's testbed

# LAWA Collection Management

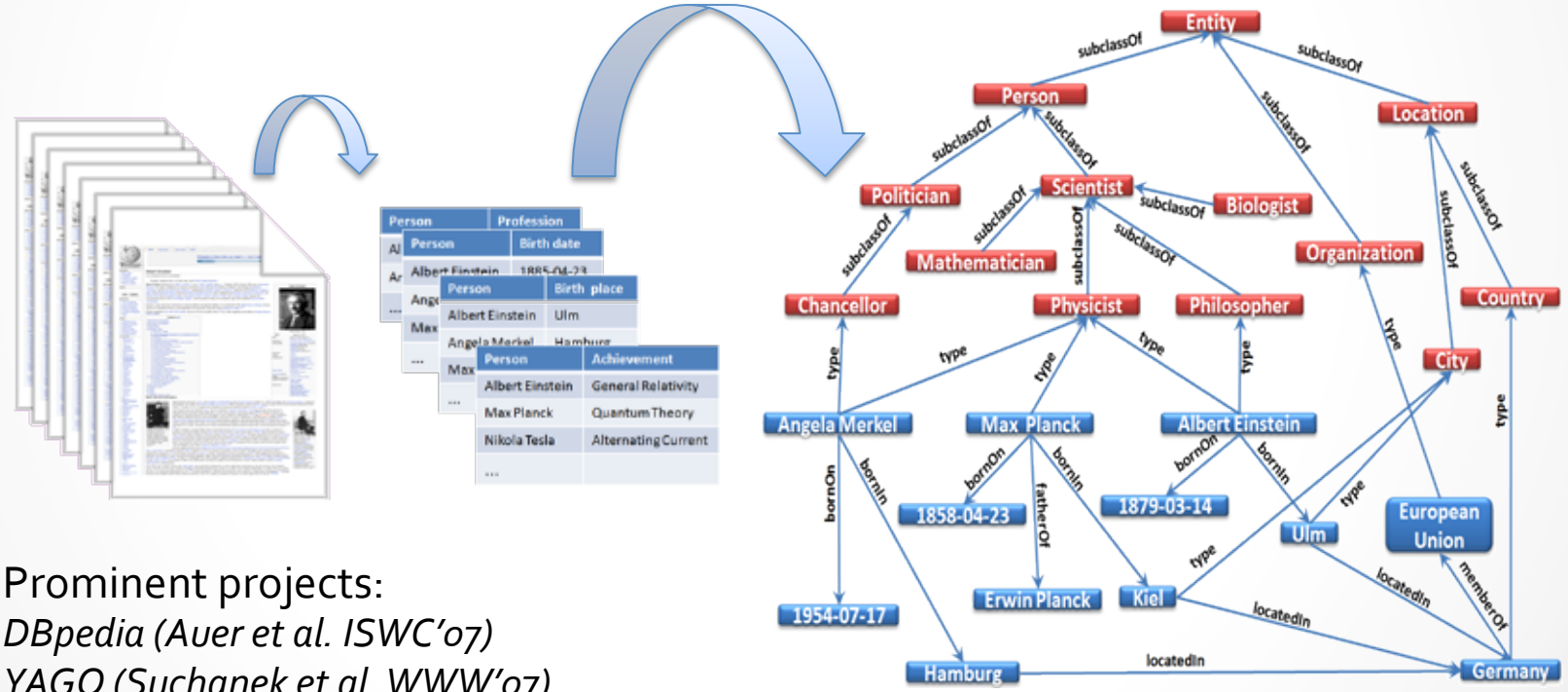


# LAWA Collection Management
# Wide-Area Hadoop Operations



# The YAGO Knowledge Base

- Knowledge harvesting from Web sources
- 1.9 Mio. entities, 19 Mio. facts




Knowledge graph

Prominent projects:

*DBpedia (Auer et al. ISWC'07)*
*YAGO (Suchanek et al. WWW'07)*
*Linking Open Data Initiative (Sem. Web)*
*Cyc, Freebase (commercial systems)*

# YAGO https://d5gate.ag5.mpi-sb.mpg.de/webyagospotlx/WebInterface

## Query

| Id | Subject | Property | Object | Time | | Location | | Keywords |
|---|---|---|---|---|---|---|---|---|
| ?id0: | ?x | isA | company | | | nearby | Stuttgart, 100 | Software |
| ?id1: | ?x | wasCreatedOnDate | ?d | after | 1970 | | | |
| ?id2: | | | | | | | | |
| ?id3: | | | | | | | | |
| ?id4: | | | | | | | | |

query

## Results

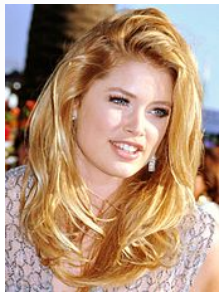
>>

| | Id | Subject | Property | Object | Time | Location | Keywords |
|---|---|---|---|---|---|---|---|
| 1 | #47711796 | SAP AG | type | company | 1972-01-01 ↓↑, - | Walldorf | supply chain management ... |
| | #47711740 | SAP AG | wasCreatedOnDate | 1972-##-## | 1972-01-01 ↓↑, 1972-12-31 ↓↑ | - | supply chain management ... |
| | #792992 | company | means | company | - | - | - |
| | #9794348 | Stuttgart | means | Stuttgart | - | Stuttgart | King Wilhelm I ... |
| 2 | #234539048 | Bechtle | type | company | 1983-01-01 ↓↑, - | Neckarsulm | Information technology consulting ... |
| | #234539080 | Bechtle | wasCreatedOnDate | 1983-##-## | 1983-01-01 ↓↑, 1983-12-31 ↓↑ | - | Information technology consulting ... |
| | #792992 | company | means | company | - | - | - |
| | #9794348 | Stuttgart | means | Stuttgart | - | Stuttgart | King Wilhelm I ... |
| 3 | #428975822 | Netviewer | type | company | 2001-01-01 ↓↑, - | Karlsruhe | Collaborative software software-as-a-service ... |
| | #428975794 | Netviewer | wasCreatedOnDate | 2001-##-## | 2001-01-01 ↓↑, 2001-12-31 ↓↑ | - | Collaborative software software-as-a-service ... |
| | #792992 | company | means | company | - | - | - |
| | #9794348 | Stuttgart | means | Stuttgart | - | Stuttgart | King Wilhelm I ... |

# AIDA

## Accurate online Disambiguation of Named Entities

Neelie Kroes



Doutzen Kroes

Larry Ellison

Jennifer Ellison

Kroes attacked Ellison about the Sun deal threatening open source software and free competition.


The Sun (UK)

Sun microsystems

Sun Microsystems

Sun Myung Moon

M. A. Yosef et al. [VLDB 2011]
C. I. Sidló et al. [QDDB2011]

# Features for Disambiguation

- Wallace and Gromit
- The Guardian
- The Independent
- …


| Prior | Similarity | Coherence |
|---|---|---|
| 58% | 0.2 | Harlan Ellison |

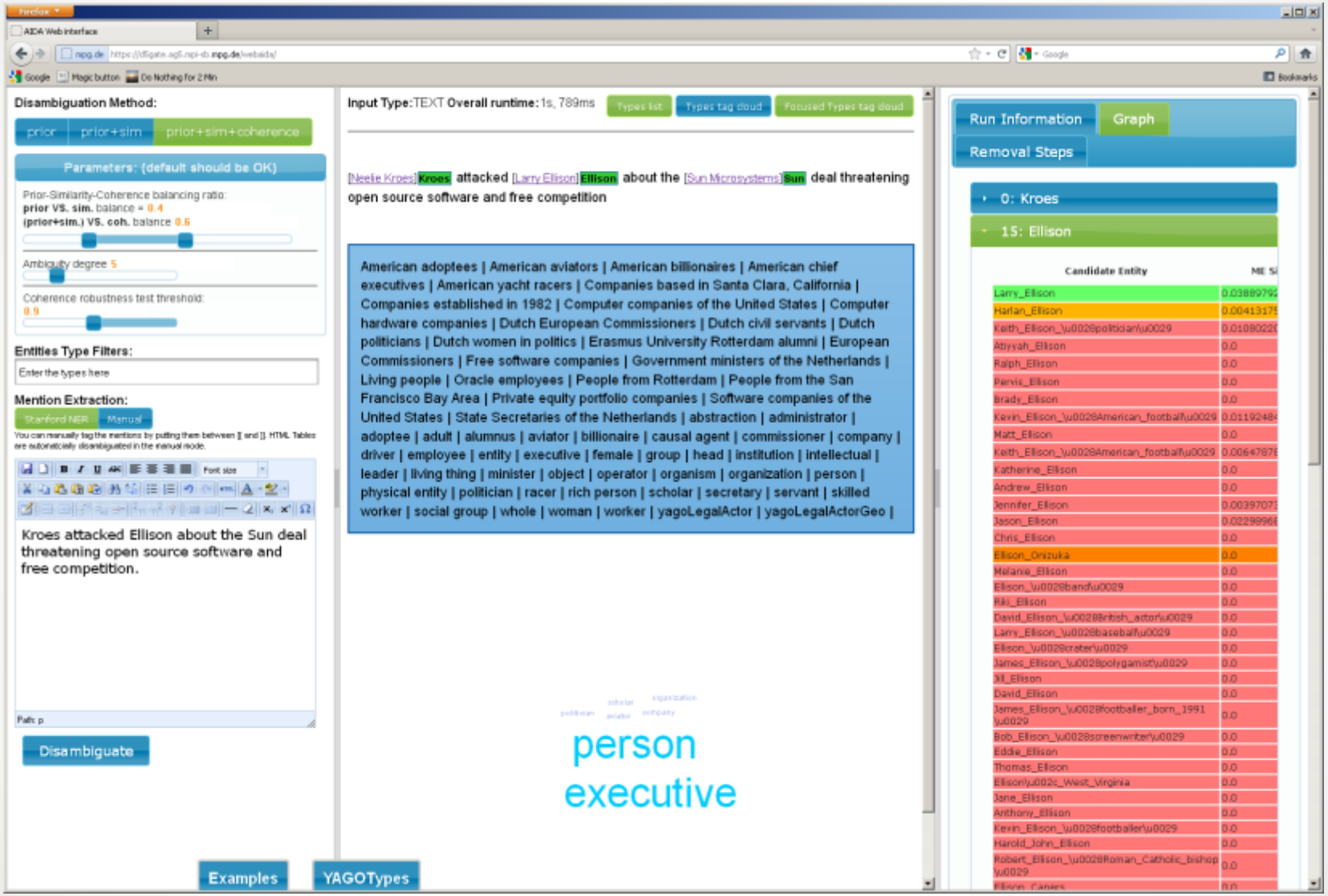
Kroes attacked Ellison about the Sun deal threatening open source software and free competition.

| Prior | Similarity | Coherence |
|---|---|---|
| 14% | 7.1 | Larry Ellison<br>Neelie Kroes |

- Larry Ellison
- Open Source
- European Union Microsoft competition case
- Computer software
- Oracle software
- …

How good do entity disambiguation links need to be?

# AIDA https://d5gate.ag5.mpi-sb.mpg.de/webaida/



# LAWA Experimental Testbed

# Web Content Classification

- Advanced query interface
- Multidimensional indices
- Entity resolution
- Ranked information retrieval
- Large-scale Web graph analytics
- Time-travel facilities

# LAWA Experimental Testbed Measurement and Tools for FIRE

- Platform for analytic tools
- Showcase: “Virtual Web Observatory”
- Initial measurement facilities and tools
  - LAWA reference collection
  - Wide-Area Hadoop Operations (Hadoop Kelvin)
  - Temporal knowledge base (YAGO)
  - Named entity disambiguation services based on AIDA
  - Service(s) for Web content classification

⇒ The *Web of the Past* “connects” the *Web of the Future*

# Conclusions & Next Steps

- "Moving" Web data up the semantic value chain
  - Semantic enrichment of Web archive data
  - Entity-level support for Web archive analytics
- Creation of a reference collection for Web analytics
  - Longitudinal Web data collection
  - Multilingual Wikipedia history
- Scaling Web analytics technologies up
  - More data
  - More Hadoop
  - More statistics

⇒ LAWA adds (textual) Web data analytics to FIRE

# LAWA User Community

- "Links" LAWA with
  - Researchers interested in Web (archive) analytics / Web science
  - The FIRE community
- Intended to
  - Get feedback and recommendations
  - Help identifying potential features of the "Virtual Web Observatory"
- Organizes annual user workshops
  - LAWA and guest presentations
  - Next workshop: November 2012

⇒ Open to everybody!

www.lawa-project.eu