

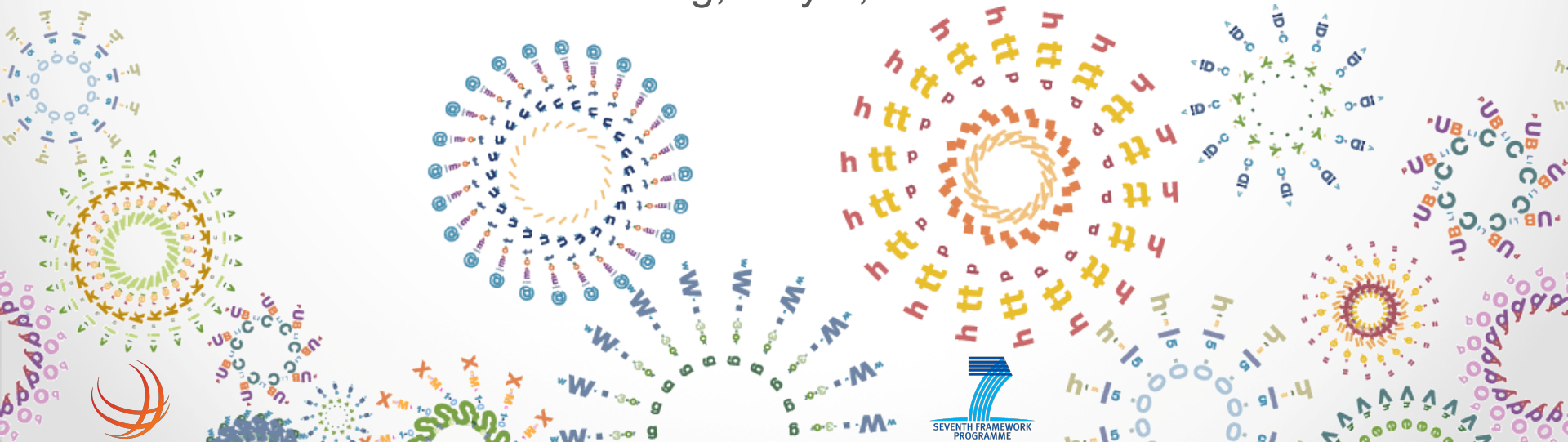


Longitudinal Analytics of Web Archive Data

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



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

The screenshot shows a web browser window titled "Internet Memory Extraction" with the URL `http://localhost:8080/hbaseadmin/views/editview/25?removeExtractor=1&extractorId=17`. The page header includes "Internet Memory" and "Web Extraction Platform". A navigation menu contains: Home, Collections, Views, Filters, Extraction, HDFS, Tools, Search, Sign in.

The main content area is titled "Edit a view". It includes a form with the following fields:

- View name:
- Column family:
- name:

Below the form is a table of extractors:

Rank	Extractor name	Projected fields	Filter name	Action
<input type="checkbox"/>	Detection of MIME types	detectedMime	ImageFilter	(remove)
<input type="checkbox"/>	Plain text extraction based on Tika	text_content text_title	PdMimeFilter	(remove)

Below the table is an "Add:" section with a dropdown menu labeled "Select One...".

At the bottom, there is a "Description:" field and a "Save" button.

On the right side of the page, there is a section for "HBase clusters" with a "+" icon. It lists:

- Current cluster: local
- Cluster: local. IP: localhost. [Edit](#) | [Choose](#)
- Cluster: Atrium. IP: 37.16.72.24. [Edit](#) | [Choose](#)

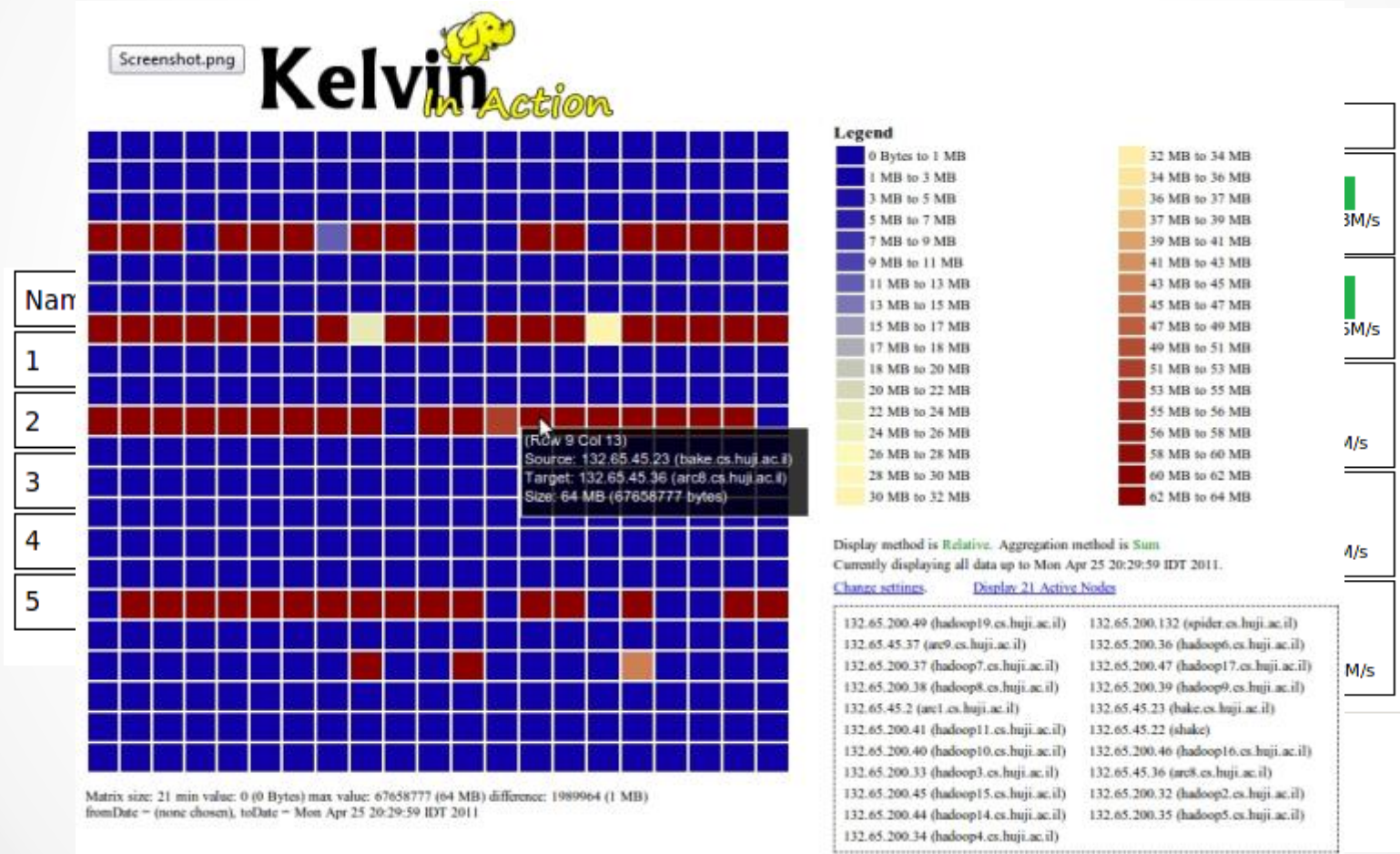
Below this is a section for "Collections [+]" with a "+" icon. It lists:

- Collection: `collection_schema`.
- Collection: `myColl`.



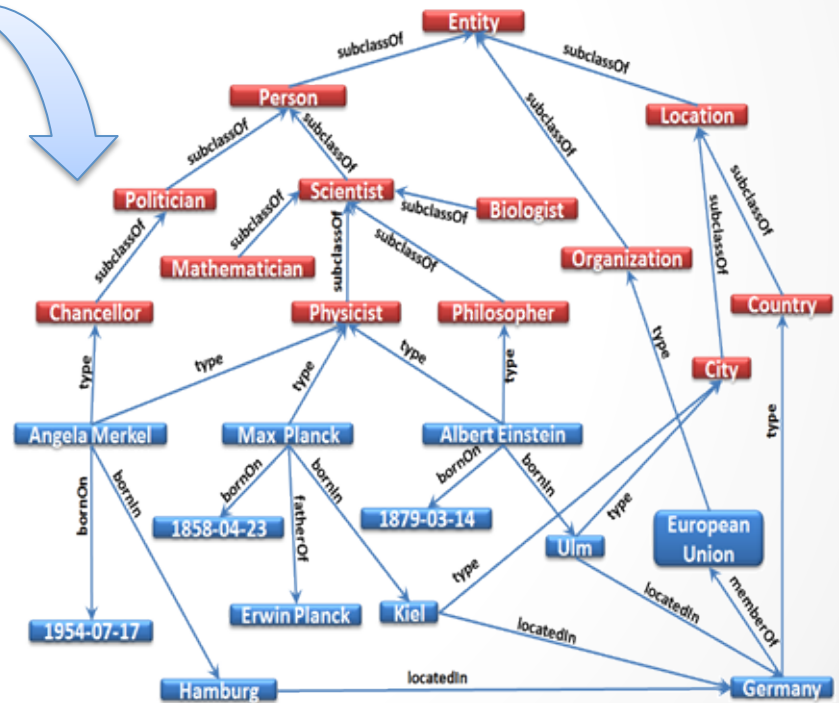
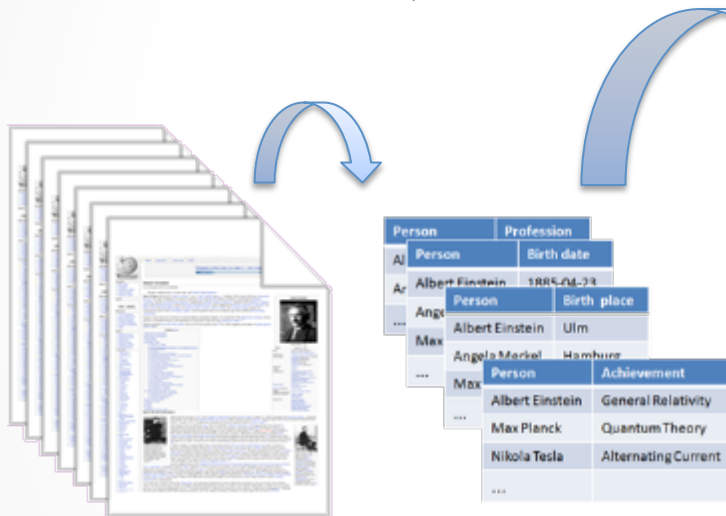
LAWA Collection Management

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)

Query

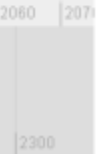
Id	Subject	Property	Object	Time	Location	Keywords	
?id0:	?x	isA	company		nearby	Stuttgart, 100	Software
?id1:	?x	wasCreatedOnDate	?d	after	1970		
?id2:							
?id3:							
?id4:							

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	-	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	-	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	-	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 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
- ...



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



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

Prior	Similarity	Coherence
58%	0.2	Harlan Ellison
14%	7.1	Larry Ellison Neelie Kroes

How good do entities like "Sun" link to disambiguation on Wikipedia?



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

Disambiguation Method:
 prior prior+sim prior+sim+coherence

Parameters: (default should be OK)
 Prior-Similarity-Coherence balancing ratio:
 prior VS. sim. balance = 0.4
 (prior+sim.) VS. coh. balance = 0.6
 Ambiguity degree: 5
 Coherence robustness test threshold: 0.9

Entities Type Filters:
 Enter the types here

Mention Extraction:
 Stanford NER Manual
 You can manually tag the mentions by putting them between [and] HTML Tags are automatically disambiguated in the manual mode.

Input Type: TEXT Overall runtime: 1s, 789ms

News Snippet:
 [Neelke Kroes] Kroes attacked [Larry Ellison] Ellison about the [Sun Microsystems] Sun deal threatening open source software and free competition

Candidate Entity List:

Candidate Entity	ME S
Larry_Ellison	0.03889792
Harlan_Ellison	0.00413171
Keith_Ellison_yu0028spolldan_yu0029	0.01080220
Abiyah_Ellison	0.0
Ralph_Ellison	0.0
Bervis_Ellison	0.0
Brady_Ellison	0.0
Kevin_Ellison_yu0028American_football_yu0029	0.01192484
Matt_Ellison	0.0
Keith_Ellison_yu0028American_football_yu0029	0.00647876
Katherine_Ellison	0.0
Andrew_Ellison	0.0
Jennifer_Ellison	0.00397071
Jason_Ellison	0.02298966
Chris_Ellison	0.0
Ellison_Oreizuka	0.0
Melania_Ellison	0.0
Ellison_yu0028band_yu0029	0.0
Rik_Ellison	0.0
David_Ellison_yu0028British_actor_yu0029	0.0
Larry_Ellison_yu0028baseball_yu0029	0.0
Ellison_yu0028crater_yu0029	0.0
James_Ellison_yu0028polygamist_yu0029	0.0
Bill_Ellison	0.0
David_Ellison	0.0
James_Ellison_yu0028footballer_born_1991_yu0029	0.0
Bob_Ellison_yu0028screenwriter_yu0029	0.0
Eddie_Ellison	0.0
Thomas_Ellison	0.0
Ellison_yu0028West_Virginia	0.0
Jane_Ellison	0.0
Anthony_Ellison	0.0
Kevin_Ellison_yu0028footballer_yu0029	0.0
Harold_John_Ellison	0.0
Robert_Ellison_yu0028Roman_Catholic_bishop_yu0029	0.0
Ellison_Caniers	0.0

YAGOTypes Ontology:
 person
 executive



LAWA Experimental Testbed

Web Content Classification

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

The screenshot shows the 'Assessment Interface' in Mozilla Firefox. The browser address bar shows the URL: <http://monster.lab.sztaki.hu:9076/lawa/pages/assessment.jsf#http://www.euromed-justice.eu>. The interface is divided into two main sections: a sidebar on the left for classification and a main content area on the right for the document being assessed.

Assessment Interface Labels:

Hosting Type	Normal
Language	English
Adult Content	No
Other Problem	No
Web Spam	No
News/Editorial	No
Commercial	No
Educational/Research	Yes
Discussion	No
Recreation/Personal	No
Media	No
Database	No
Readability-Vts	Good
Readability-Lang	Good
Neutrality	Facts
Bias	Not biased
Trustiness	Trustworthy
Unsure	No

Main Content Area:

The main content area displays the 'EUROMED JUSTICE II PROJECT' page. It features the European Union flag and the text 'Project funded by the European Union'. The page title is 'EUROMED JUSTICE II PROJECT' and 'PROJET EUROMED JUSTICE'. The language is set to 'english'. The page content includes an 'Introduction' section in English, French, and Arabic, and a section for 'Implemented by' with logos for the Spanish Government and FIAPP.



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