

ALEXANDRIA

Exploring Web Archives: Challenges and Solutions

Wolfgang Nejdl

L3S Research Center
Hannover, Germany

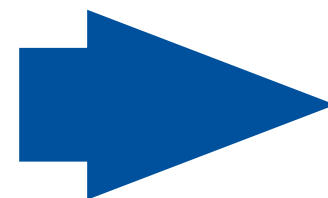


Web Science @ L3S

Computer Science and interdisciplinary research on all aspects of the Web

- Internet: Communication and Networks
- Information: Accessing information and knowledge on and through the Web
- Community: Supporting communities and groups on the Web, for research, education, production and entertainment
- Society: Requirements (technological, social, legal) for the Web

Selected projects



Real-time data processing
for finance predictions



Cross-media analysis
and interpretation



Retrieval, Exploration and
Analytics for Web
Archives (ERC)



ForgetIT

ForgetIT: Concise
Preservation via
Managed Forgetting



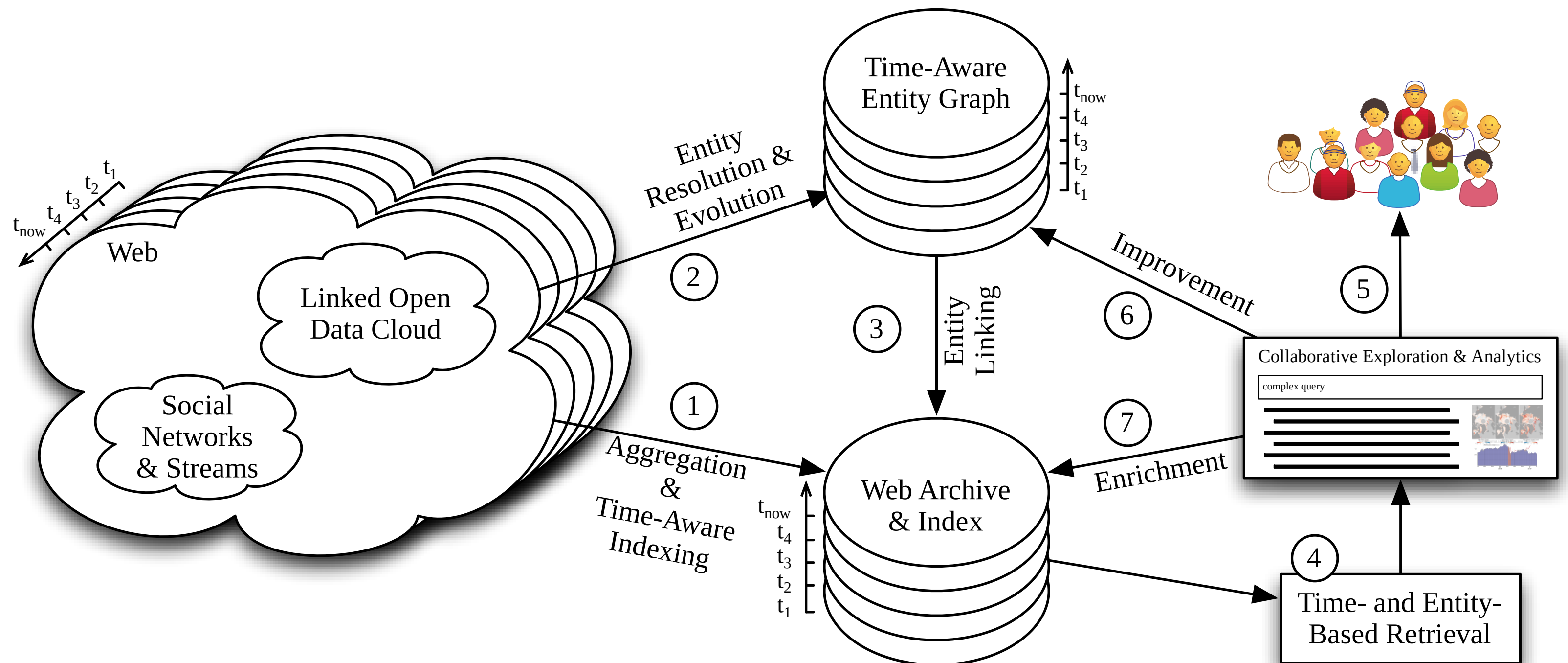
CUBRIK: Searching by
computers and humans

MAPPING

Privacy, Property and
Internet Governance

How can we provide better access to Web archives?

ALEXANDRIA Vision and Research Questions



Evolution-Aware Entity-Based Enrichment and Indexing

Q1: How to link web archive content against multiple entity and event collections evolving over time?

Ioannou, E., Nejdl, W., Niederée, C. and Velegrakis, Y. 2011. LinkDB: A Probabilistic Linkage Database System. *SIGMOD* (New York, New York, USA, Jun. 2011)

Q2: How to maintain entity and event information and indexes for web-scale archives?

Papadakis, G., Ioannou, E., Niederée, C., Palpanas, T. and Nejdl, W. 2012. Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data. *WSDM* (New York, NY, USA, 2012), 53–62.

Papadakis, G., Ioannou, E., Palpanas, T., Niederée, C. and Nejdl, W. 2012. A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces. *TKDE*. (2012).

Aggregating Social Networks and Streams

Q3: How to archive complex and dynamic network structures from social media?

Siersdorfer, S., Chelaru, S., Nejdl, W. and San Pedro, J. 2010. How useful are your comments? Analyzing and Predicting YouTube Comments and Comment Ratings. *WWW* (New York, New York, USA, Apr. 2010), extended for *TWEB* (2014)

Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y. and Senellart, P. 2012. Exploiting the Social and Semantic Web for guided Web Archiving. *TPDL* (Sep. 2012)

Q4: How to aggregate social media streams for archiving?

Minack, E., Siberski, W. and Nejdl, W. 2011. Incremental diversification for very large sets: a streaming-based approach. *SIGIR* (New York, New York, USA, Jul. 2011)

Diaz-Aviles, E., Drumond, L., Schmidt-Thieme, L. and Nejdl, W. 2012. Real-time top-n recommendation in social streams. *RecSys* (New York, New York, USA, 2012)

Temporal Retrieval and Ranking

Q5: How to support time-sensitive and entity-based query formulation?

Kanhabua, N. and Nørnvåg, K. 2010. Exploiting time-based synonyms in searching document archives. *JCDL* (New York, New York, USA, Jun. 2010)

Nguyen, T., and Kanhabua, N. 2014. Leveraging dynamic query subtopics for time-aware search result diversification. *ECIR* (Amsterdam, April 2014)

Q6: How to improve result ranking and clustering for time-sensitive and entity-based queries?

Kanhabua, N., Blanco, R. and Matthews, M. 2011. Ranking related news predictions. *SIGIR* (New York, New York, USA, Jul. 2011)

G. Demartini, C. Firan, T. Iofciu, R. Krestel, W. Nejdl: Why finding entities in Wikipedia is difficult, sometimes. *Inf. Retr.* 13(5): 534-567 (2010)

Collaborative Exploration and Analytics

Q7: How to support collaborative and complex search and analysis processes?

Ivana Marenzi and Sergej Zerr. Multiliteracies and Active Learning in CLIL - The Development of LearnWeb2.0 - IEEE Transactions on Learning Technologies (2012)

Q8: How to leverage (user) search and analysis processes to improve the web archive?

K. Bischoff, C. Firan, W. Nejdl, R. Paiu: Bridging the gap between tagging and querying vocabularies: Analyses and applications for enhancing multimedia IR. J. Web Sem. 8(2-3): 97-109 (2010)

M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, S. Siersdorfer: Extracting Event-Related Information from Article Updates in Wikipedia. ECIR 2013: 254-266

Events: What is going on in the world?



CRISIS IN SYRIA
Assad Denies Gas Attack
White House Presses a Case
By MICHAEL S. SCHMIDT
1:41 PM ET
As the White House staff appeared on morning news show, President Bashar al-Assad said in a CBS interview that his government was not responsible for a chemical attack.

As the World Watched, Syria Supply of News
By DAVID E. SANGER, W. LEHREN and RICK GLADSTONE
Syria's leaders exploit loopholes and lax enforcement to amass the world's largest stockpiles of chemical weapons, while few

CNN Breaking News
@cnnbrk

Breaking News from CNN, via the CNN.com homepage team. Now 18M strong. Check @cnn for all things CNN breaking and more.

📍 Everywhere
🌐 cnn.com
🕒 Joined January 2007

[Tweet to CNN Breaking News](#)

14 Followers you know



theguardian

BBC
WORLD
NEWS



BBC News

Home Videos Playlists Channels Discussion About

YouTube

Subscribe 265,491



Why should you subscribe to BBC News?

145,607 views 7 months ago

Here are several reasons why you should subscribe to the official BBC News YouTube channel.

Subscribe to BBC News HERE <http://bit.ly/1rbfUog>
Check out our website: <http://www.bbc.com/news>
Facebook: <http://www.facebook.com/bbcnews>
Twitter: <http://www.twitter.com/bbcnews>
Instagram: <http://instagram.com/bbcnews>

BBC News on YouTube

BBC Trending
Subscribe

BBC HARDtalk
Subscribe

BBC Newsnight
Subscribe

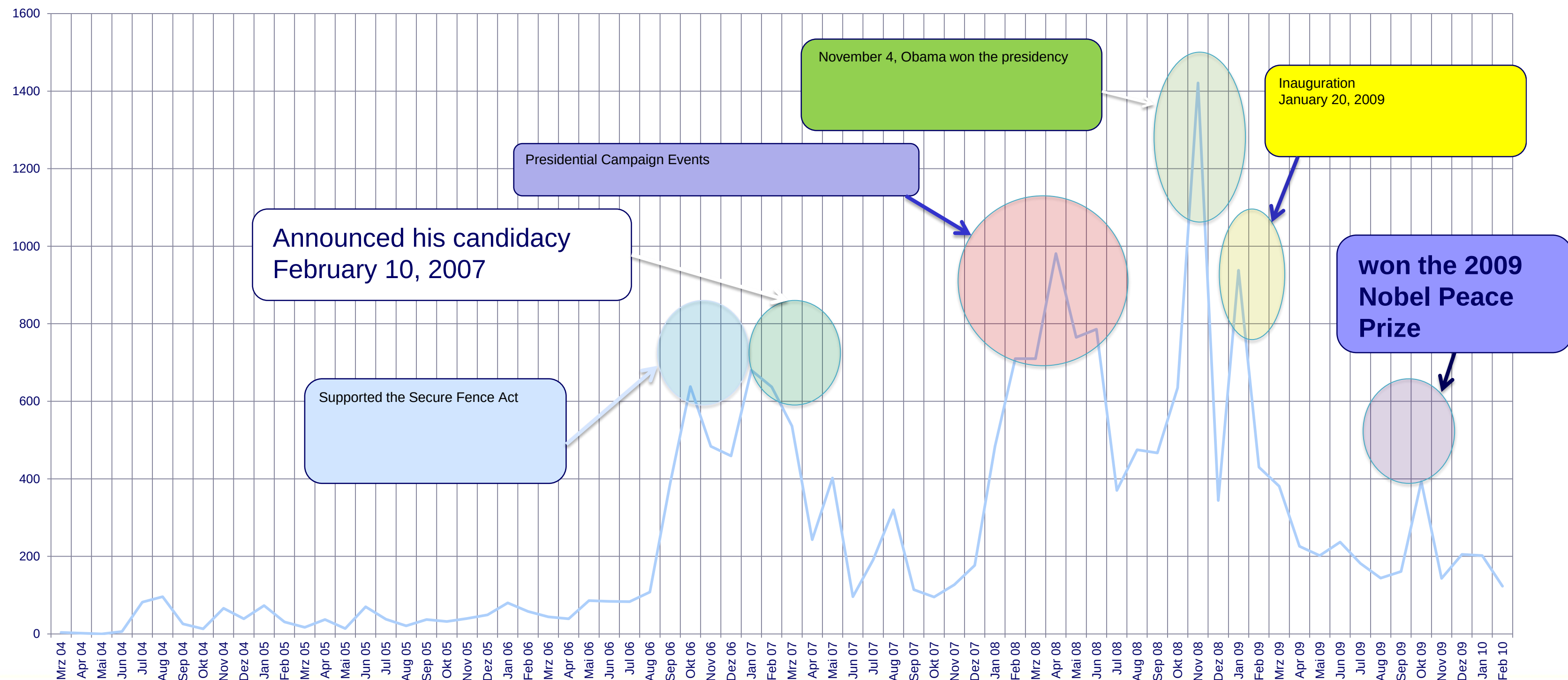
BBC
Subscribe

BBCWorldwide
Subscribe

BBC Earth

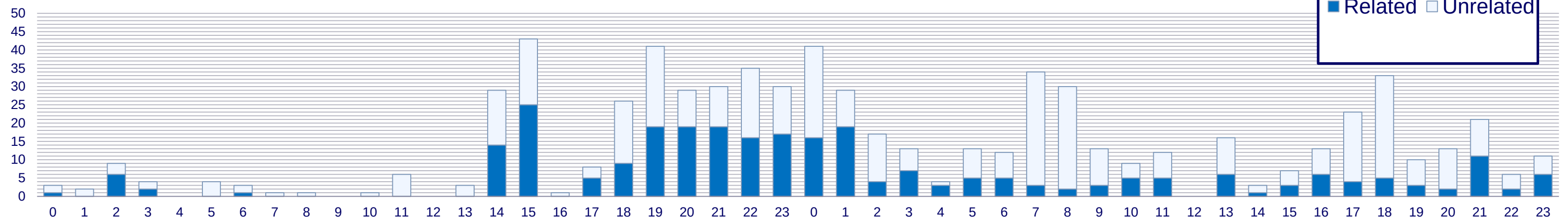
Peaks in Wikipedia update activity correlate with events

Edit history for the Barack Obama article (monthly)

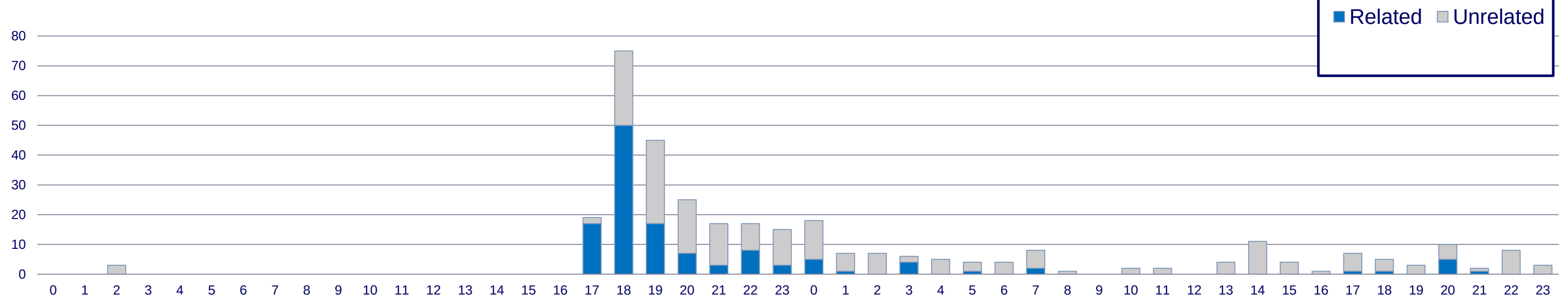


Update activity: controversy- or event-related?

Kosovo: Independence Declaration



Donald Rumsfeld: Resignation



Temporal Summarization

- Timelines summaries provide an overview of relevant events
- Important for news stories that last months, or years:
 - e.g. Syrian civil war and Ukraine crisis
- News agencies: manual creation is expensive, limited, slow, if exists at all.
- Automatic generation of timelines is often of a low quality

Time

Anti-Gaddafi uprising

2011 February - Arrest of human rights campaigner sparks violent protests in eastern city of Benghazi that rapidly spread to other cities. Authorities use aircraft to attack protestors. Gaddafi insists that he will not quit, and remains in control of the capital, Tripoli.

2011 March - UN Security Council authorises a no-fly zone over Libya and air strikes to protect civilians, over which NATO assumes command.

Backed by extensive NATO air raids, Libyan rebels initially capture territory but are then forced back by better-armed pro-Gaddafi forces. Rebels ask West for arms.

2011 July - The international Contact Group on Libya formally recognises the main opposition group, the National Transitional Council (NTC), as the legitimate government of Libya.

2011 August - Rebels swarm into Col Gaddafi's fortress compound in Tripoli, six months after the uprising began. With only a few remaining strongholds under his control, Col Gaddafi goes into hiding. His wife and three of his children flee to neighbouring Algeria.

2011 August-September - African Union joins 60 countries which have recognised the NTC as the new Libyan authority.

2011 20 October - Col Gaddafi is captured and killed.

Three days later, the NTC declares Libya to be officially "liberated" and announces plans to hold elections within eight months.

2011 November - Saif al-Islam, the fugitive son of former Libyan leader Muammar Gaddafi, is captured, becoming the last key Gaddafi family member to be seized or killed.

BBC
WORLD
NEWS

Col Gaddafi killed



The colonel's ruthless and quixotic 42-year rule came to an ignominious end

Obituary: Muammar Gaddafi
How Gaddafi died

News Stories on Wikipedia

2014 Ukrainian revolution

From Wikipedia, the free encyclopedia

Euromaidan

From Wikipedia, the free encyclopedia

2014 Crimean crisis

From Wikipedia, the free encyclopedia

2014 pro-Russian unrest in Ukraine

From Wikipedia, the free encyclopedia

This article is about the unrest throughout Ukraine. For the Russian military intervention there, see [2014 Russian military intervention in Ukraine](#). For the crisis in Crimea, see [2014 Crimean crisis](#). For the armed conflict, see [War in Donbass](#).

Since the end of February 2014, demonstrations by pro-Russian and anti-government groups have taken place in major cities across the eastern and southern regions of Ukraine, in the aftermath of the Euromaidan movement and the 2014 Ukrainian revolution. During the first stage of the unrest, Crimea was annexed by the Russian Federation after a crisis in the region, Russian military intervention and an internationally criticized referendum. Protests in Donetsk and Luhansk oblasts escalated into an armed separatist insurgency.^{[56][57]} This led the Ukrainian government to launch a military counter-offensive against the insurgents, which resulted in the ongoing War in Donbass.^[58]

Contents [hide]

- 1 Background
 - 1.1 Public opinion
 - 1.2 Anti-Maidan and paid protesters
 - 1.3 Media portrayal
- 2 Timeline
- 3 Unrest by region
 - 3.1 Crimea
 - 3.2 Donetsk Oblast
 - 3.2.1 Government building seizures
 - 3.2.2 Second counter-offensive
 - 3.2.3 Referendum
 - 3.2.4 Continued fighting

2014 pro-Russian unrest in Ukraine

Part of the aftermath of the 2014 Ukrainian revolution



Map of protests by region, indicating severity of the unrest at its peak

Date 23 February 2014–ongoing
(6 months, 3 weeks and 3 days)

Location Eastern Ukraine
Southern Ukraine
Crimea

Current Events on Wikipedia

news stories news updates of the day

May 2, 2014 (Friday) edit history watch

Armed conflicts and attacks

- **Syrian Civil War (Siege of Homs):**
 - A 24-hour ceasefire deal is reached in Syria's city of [Homs](#) allowing rebels to withdraw from their last stronghold in the central city. [\(BBC\)](#) [\(Reuters\)](#)
- **Islamist unrest in Egypt (Sinai insurgency):**
 - Attacks by suicide bombers on a security checkpoint and a civilian bus in [Egypt's Sinai Peninsula](#) as well as a separate bombing near a courthouse in [Cairo](#) killed at least four people. [\(The Washington Post\)](#)
 - Two people were shot dead when supporters of deposed President [Mohamed Morsi](#) clashed with residents in [Alexandria](#). [\(Reuters\)](#)
- At least nine students are killed in protests in [Ethiopia's Oromia Region](#). [\(BBC\)](#)
- Security tightens in the [People's Republic of China](#) after Wednesday's bombing at a train station left three dead and 79 injured. [\(Daily Mail\)](#)
- 2014 pro-Russian unrest in Ukraine and the [Siege of Sloviansk](#)
 - The [Armed Forces of Ukraine](#) launch a military operation against separatists in the city of [Sloviansk](#) with claims of many deaths and two Ukrainian Army helicopters downed. In the port city of [Odessa](#), violent clashes kill at least 42 people, as pro-Ukrainians set the pro-Russian headquarters on fire. [\(CNN\)](#), [\(BBC\)](#), [\(Express\)](#)

Disasters and accidents

- A landslide strikes the town of Hobo Barik in [Afghanistan's Badakhshan](#) province with at least 350 people dead and over 2000 people missing. [\(USA Today\)](#), [\(AP\)](#)

Law and crime

- **Murder of Jean McConville and Arrest of Gerry Adams:**
 - [Sinn Féin](#) claims that party president [Gerry Adams](#) is a victim of "political policing" are dismissed by the Taoiseach (Prime Minister of Ireland) [Enda Kenny](#), British prime minister [David Cameron](#) and Northern Ireland First Minister [Peter Robinson](#). [\(Irish Times\)](#)
 - Helen, the eldest daughter of IRA murder victim Jean McConville, is prepared to name the people responsible for her mother's death, saying she no longer fears reprisals by republicans. [\(RTE News\)](#)



WikiTimes Dataset

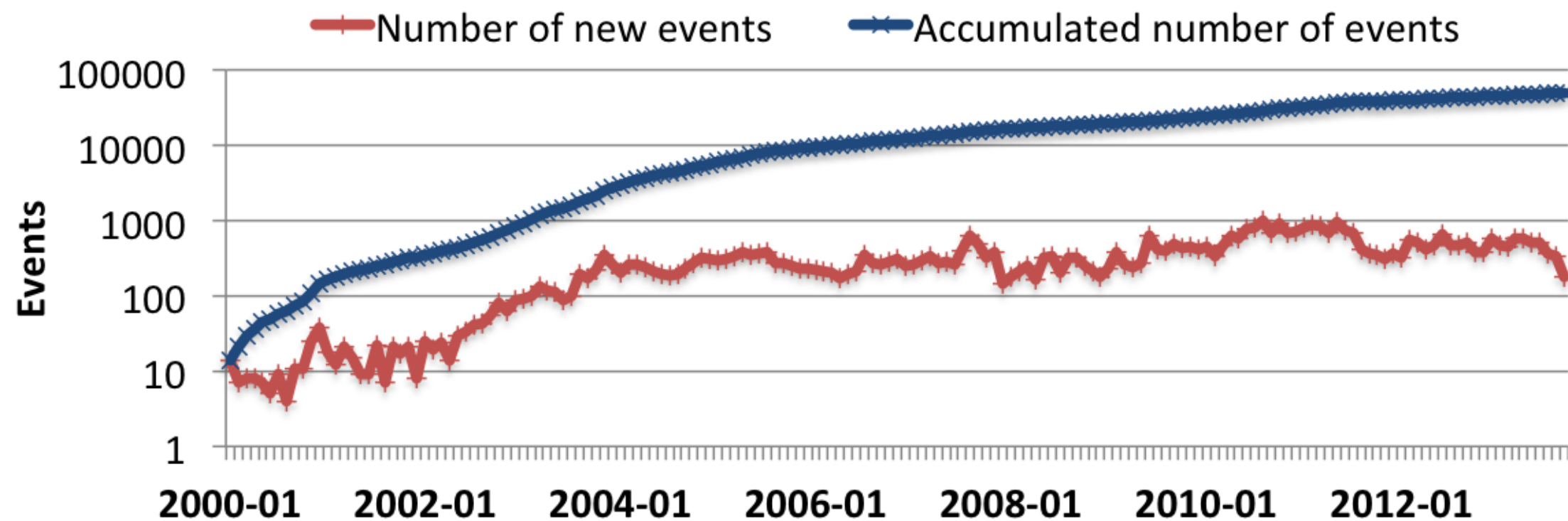
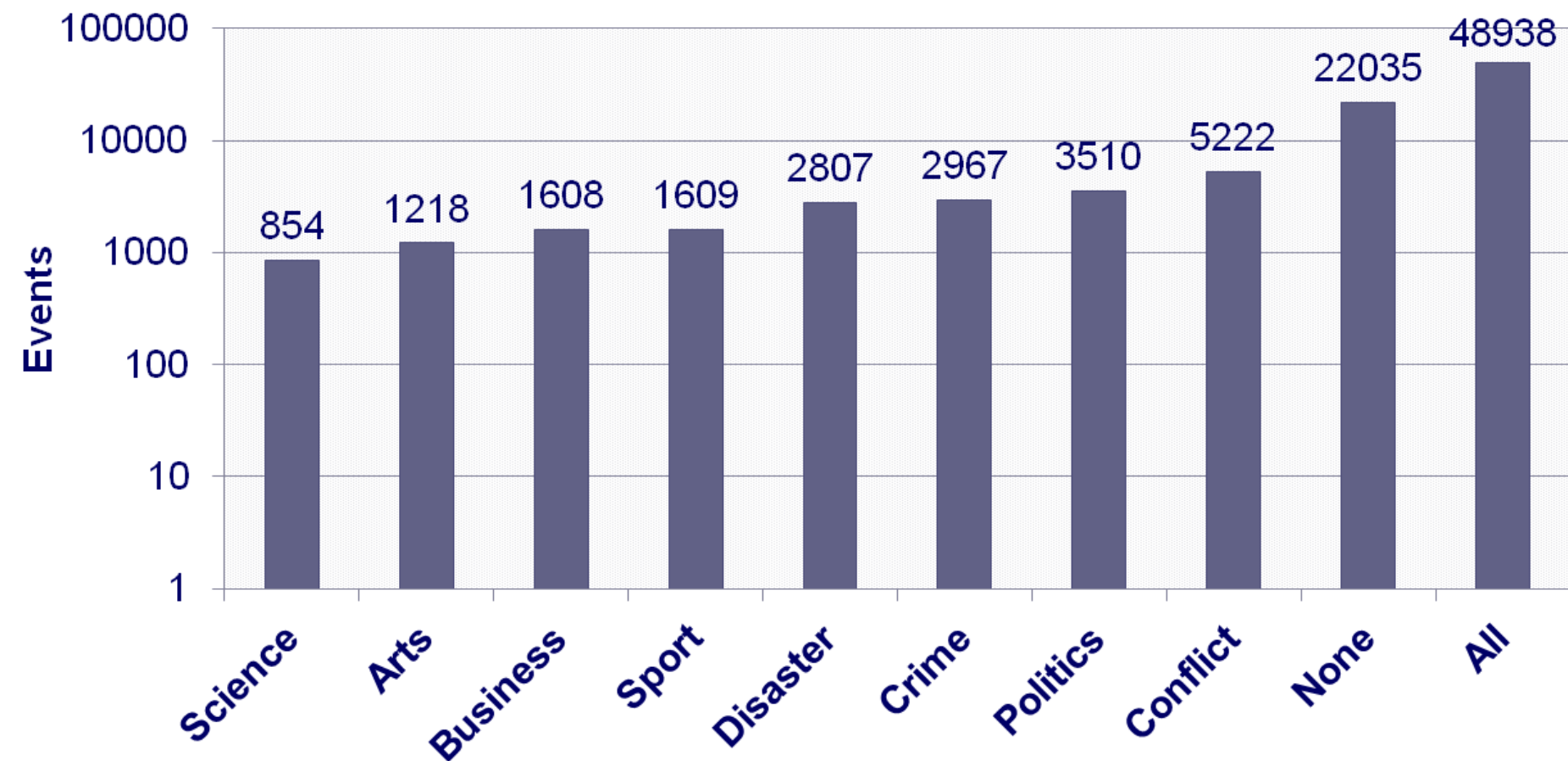
Collected events from the past
15 years

#stories in total	882
#events in total	51378
#events with story link	8206
#events without story link	43172
#events with topical category	24371
#entities	199788

Event Relation			
Relation	#Total facts	#Evaluated facts	Accuracy
happenedOnDate	51378	512	0.973 +- 0.014
belongsToStory	8206	100	1.000 +- 0.000
linksTo (Entity)	170001	512	0.982 +- 0.011
category	24371	300	0.979 +- 0.015

Quality assessment

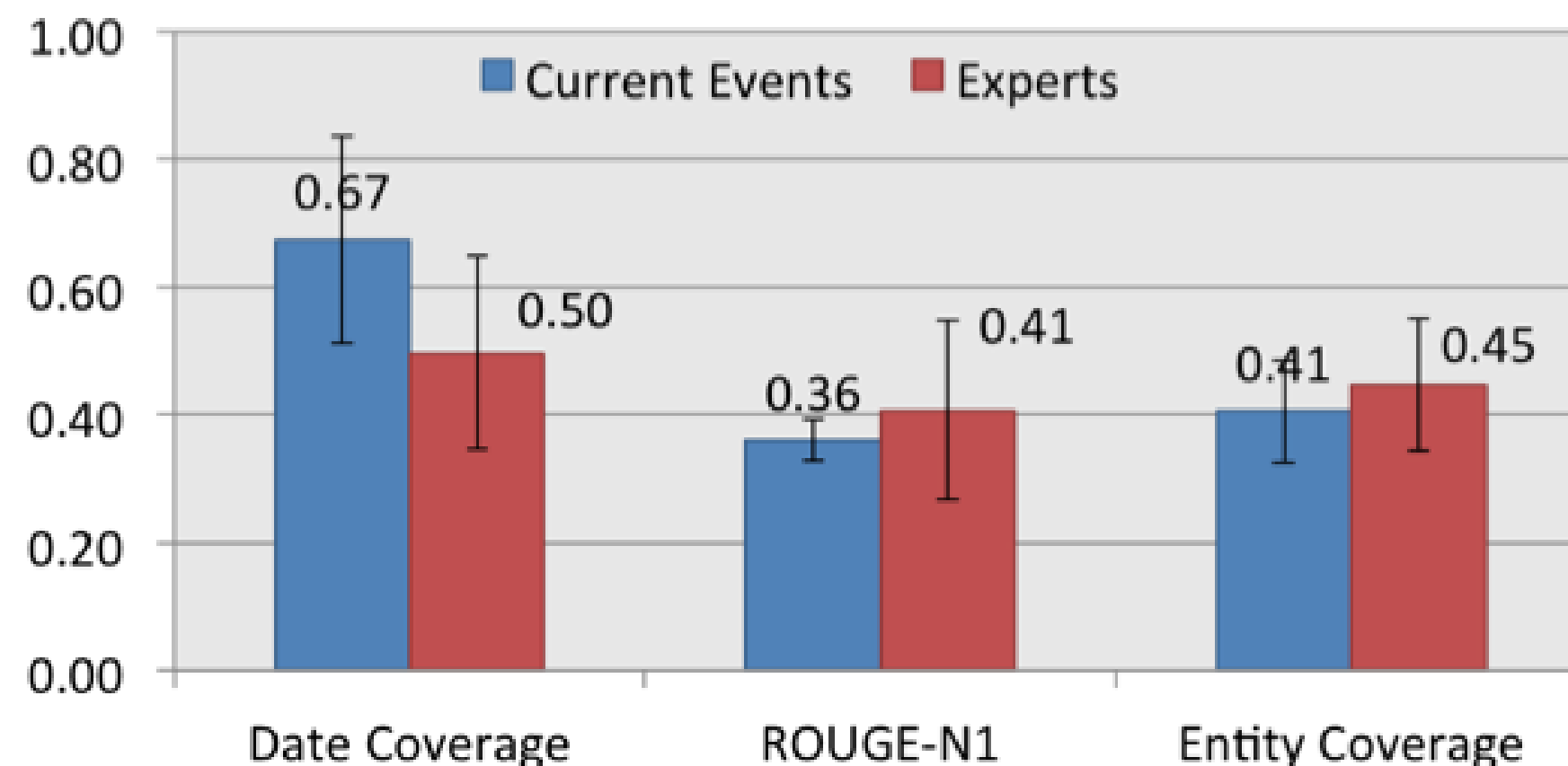
WikiTimes Dataset



WikiTimes Dataset

Comparison with timelines published by news agencies

- 21 timeline summaries on four example news stories:
 - e.g. Egypt Revolution, Syria Civil War, Yemen Crisis and Libyan Civil War
- Comparison with timelines published by 25 popular news agencies:
 - e.g. CNN, BBC, Reuters, New York Times, etc.
 - between January 2011 and July 2013.



- Giang Binh Tran and Mohammad Alrifai, "Indexing and Analyzing *Wikipedia's Current Events, the Daily News Summaries by the Crowd*, the 23rd International World Wide Web Conference, Web Science Track, 2014. <http://wikitimes.L3S.de/>

Learning from BUDDAH

- Combining qualitative and quantitative analysis.
- Corpus creation is very important, even more so is the documentation.
- We need much better ranking and exploration functionalities.
- Access is still problematic:
 - Keyword search is an improvement but advanced queries are needed to get less spam. Keyword generation is not fun when you're a scholar.
 - User intent is slightly different from traditional search. Users want documents to help them study a topic's history/evolution/development over time.

Filter by public_suffix:

- [co.uk \(5241\)](#)
- [org.uk \(254\)](#)
- [ac.uk \(53\)](#)
- [gov.uk \(28\)](#)
- [nhs.uk \(5\)](#)
- [me.uk \(2\)](#)
- [police.uk \(1\)](#)

Filter by content_type_norm:

- [html \(5412\)](#)
- [text \(93\)](#)
- [other \(58\)](#)
- [pdf \(17\)](#)
- [image \(4\)](#)

Filter by crawl_year:

- [2006 \(1600\)](#)
- [2007 \(1225\)](#)
- [2008 \(539\)](#)
- [2002 \(513\)](#)
- [2003 \(502\)](#)
- [2004 \(425\)](#)
- [2005 \(389\)](#)
- [2001 \(197\)](#)
- [2009 \(102\)](#)
- [2000 \(71\)](#)
- [1999 \(12\)](#)
- [1997 \(6\)](#)
- [1998 \(3\)](#)

Search

Enter your search terms

Sort by

Crawl Date ▼

Order

Asc ▼

Search

Clear

1. **Title:** [Hey folks, it's Rosie O'Donnell](#)

Crawl Date: 1997-04-11T17:57:09Z [\[more\]](#)

Host: webs.demon.co.uk (html)

Access Problems? [Use our access helper.](#)

2. **Title:** [Hey folks, it's Rosie O'Donnell](#)

Crawl Date: 1997-05-09T03:07:46Z [\[more\]](#)

Host: webs.demon.co.uk (html)

Access Problems? [Use our access helper.](#)

3. **Title:** [\[untitled\]](#)

Crawl Date: 1997-06-30T01:33:26Z [\[more\]](#)

Host: vibes.vossnet.co.uk (text)

Access Problems? [Use our access helper.](#)

4. **Title:** ["Things To Avoid At A Classical Concert"](#)

Crawl Date: 1997-07-30T11:17:14Z [\[more\]](#)

Host: wkweb4.cableinet.co.uk (html)

Access Problems? [Use our access helper.](#)

5. **Title:** [Vogue Daily](#)

Crawl Date: 1998-02-01T14:21:16Z [\[more\]](#)

Host: vitalwoman.co.uk (html)

Access Problems? [Use our access helper.](#)

6. **Title:** [Vogue Daily](#)

Crawl Date: 1998-02-01T22:03:15Z [\[more\]](#)

Host: vogue.co.uk (html)

Access Problems? [Use our access helper.](#)

Filter by links_public_suffixes:

- [co.uk \(5138\)](#)
 - [com \(3636\)](#)
 - [net \(425\)](#)
 - [org.uk \(364\)](#)
 - [org \(308\)](#)
 - [us \(133\)](#)
 - [ca \(112\)](#)
 - [gov \(88\)](#)
 - [de \(84\)](#)
 - [edu \(58\)](#)
 - [ac.uk \(57\)](#)
 - [gov.uk \(56\)](#)
 - [fr \(50\)](#)
 - [co.jp \(40\)](#)
 - [it \(38\)](#)
 - [ny.us \(38\)](#)
 - [ie \(33\)](#)
 - [tv \(28\)](#)
 - [com.au \(23\)](#)
 - [se \(23\)](#)
- [Show more](#)

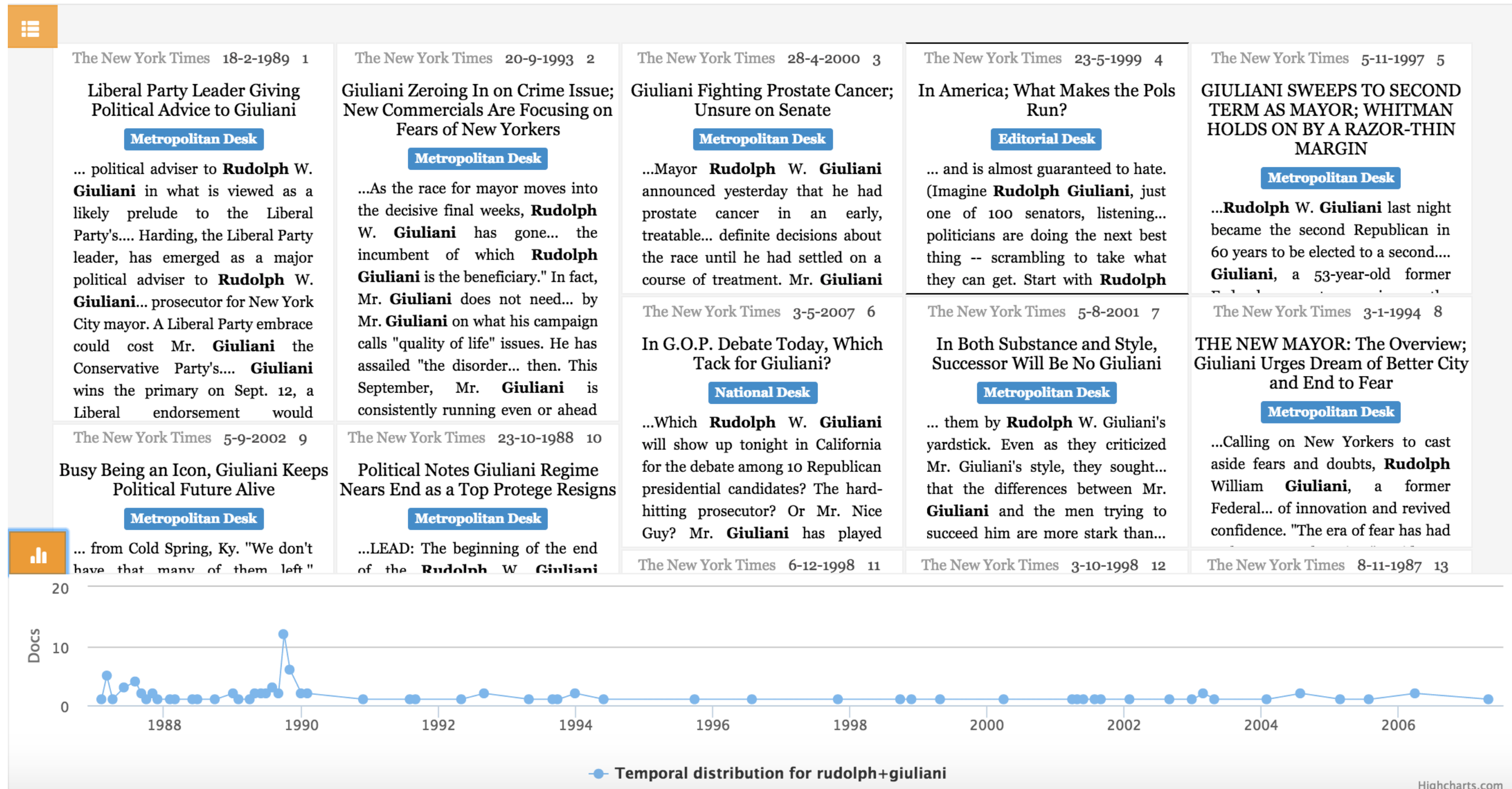
Filter by author:

- [ERNŐ SIMON \(5\)](#)
- [BBC \(2\)](#)
- [HO \(2\)](#)
- [jonathanharley \(2\)](#)
- [Katy \(2\)](#)
- [Mike Segar \(2\)](#)
- [bryan \(1\)](#)
- [Mark \(1\)](#)
- [Sarah Charman \(1\)](#)
- [www \(1\)](#)

History by Diversity

- Diversify just aspects: no guarantee we will get temporal diversity.
- Diversify just time: no guarantee we will get aspect diversity.
- HQI: I want documents from the most important aspects and from the most important time windows. (New Retrieval Task)

Rudolph Giuliani – ordered by histdiv



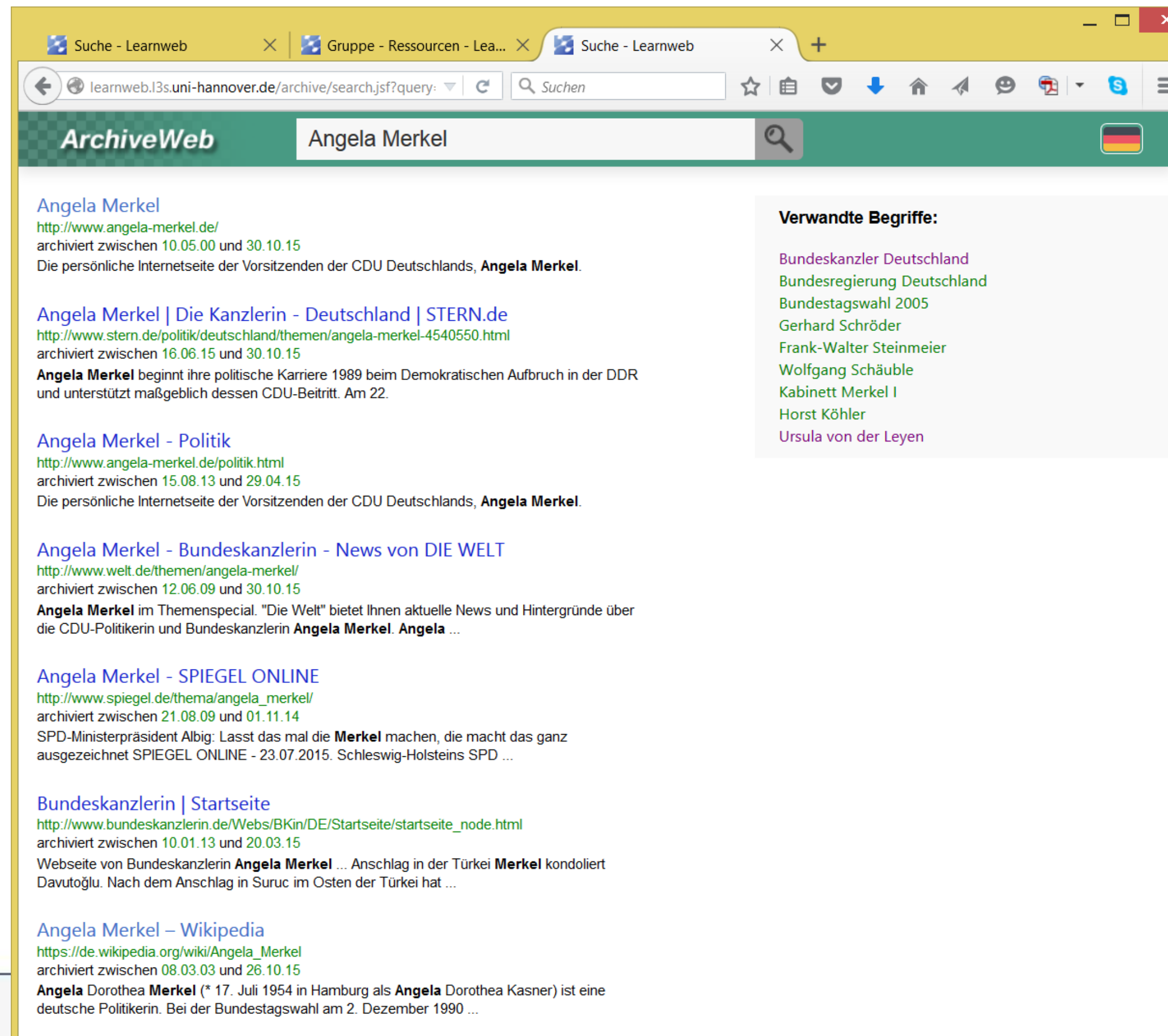
Exploring the Internet Archive

Search for important pages
about entities and
get related entities

with high precision and
appropriate ranking

(assumption:
recall is not important)

without full text index ☺
based on Bing ...



The screenshot shows a web browser window with three tabs: 'Suche - Learnweb', 'Gruppe - Ressourcen - Lea...', and 'Suche - Learnweb'. The address bar shows 'learnweb.l3s.uni-hannover.de/archive/search.jsf?query:'. The search results are displayed on the 'ArchiveWeb' page, with the search term 'Angela Merkel' entered in the search bar. The results list several links to Angela Merkel's website and news articles, each with a date range indicating when the page was archived. A sidebar on the right titled 'Verwandte Begriffe:' (Related Concepts) lists related terms such as 'Bundeskanzler Deutschland', 'Bundesregierung Deutschland', 'Bundestagswahl 2005', 'Gerhard Schröder', 'Frank-Walter Steinmeier', 'Wolfgang Schäuble', 'Kabinett Merkel I', 'Horst Köhler', and 'Ursula von der Leyen'.

Suche - Learnweb | Gruppe - Ressourcen - Lea... | Suche - Learnweb

learnweb.l3s.uni-hannover.de/archive/search.jsf?query: Suchen

ArchiveWeb Angela Merkel

Angela Merkel
<http://www.angela-merkel.de/>
archiviert zwischen 10.05.00 und 30.10.15
Die persönliche Internetseite der Vorsitzenden der CDU Deutschlands, **Angela Merkel**.

Angela Merkel | Die Kanzlerin - Deutschland | STERN.de
<http://www.stern.de/politik/deutschland/themen/angela-merkel-4540550.html>
archiviert zwischen 16.06.15 und 30.10.15
Angela Merkel beginnt ihre politische Karriere 1989 beim Demokratischen Aufbruch in der DDR und unterstützt maßgeblich dessen CDU-Beitritt. Am 22.

Angela Merkel - Politik
<http://www.angela-merkel.de/politik.html>
archiviert zwischen 15.08.13 und 29.04.15
Die persönliche Internetseite der Vorsitzenden der CDU Deutschlands, **Angela Merkel**.

Angela Merkel - Bundeskanzlerin - News von DIE WELT
<http://www.welt.de/themen/angela-merkel/>
archiviert zwischen 12.06.09 und 30.10.15
Angela Merkel im Themenspecial. "Die Welt" bietet Ihnen aktuelle News und Hintergründe über die CDU-Politikerin und Bundeskanzlerin **Angela Merkel**. **Angela** ...

Angela Merkel - SPIEGEL ONLINE
http://www.spiegel.de/thema/angela_merkel/
archiviert zwischen 21.08.09 und 01.11.14
SPD-Ministerpräsident Albig: Lasst das mal die **Merkel** machen, die macht das ganz ausgezeichnet SPIEGEL ONLINE - 23.07.2015. Schleswig-Holsteins SPD ...

Bundeskanzlerin | Startseite
http://www.bundeskanzlerin.de/Webs/BKin/DE/Startseite/startseite_node.html
archiviert zwischen 10.01.13 und 20.03.15
Webseite von Bundeskanzlerin **Angela Merkel** ... Anschlag in der Türkei **Merkel** kondoliert Davutoğlu. Nach dem Anschlag in Suruc im Osten der Türkei hat ...

Angela Merkel – Wikipedia
https://de.wikipedia.org/wiki/Angela_Merkel
archiviert zwischen 08.03.03 und 26.10.15
Angela Dorothea Merkel (* 17. Juli 1954 in Hamburg als **Angela Dorothea Kasner**) ist eine deutsche Politikerin. Bei der Bundestagswahl am 2. Dezember 1990 ...

Verwandte Begriffe:

- Bundeskanzler Deutschland
- Bundesregierung Deutschland
- Bundestagswahl 2005
- Gerhard Schröder
- Frank-Walter Steinmeier
- Wolfgang Schäuble
- Kabinett Merkel I
- Horst Köhler
- Ursula von der Leyen

Evolution and Semantic Enrichment

- Maarten de Rijke: Monitoring Vocabulary Shifts over Time.
- Avishek Anand: Wikipedia Enrichment using News Archives.
- Besnik Fetahu: Suggesting news Articles for Enriching and Expanding Wikipedia Entity Pages. (Demo)
- Tuan Tran: Semantic Annotation for Microblog Topics using Wikipedia Temporal Information.
- Binh Giang Tran: WikipEvent: How do we get events from the crowd? (Demo)

Social Networks and Relations

- Sergej Zerr, Miroslav Shaltev: Temporal Analysis of Social Networks in Archives. (+ Demo)

Temporal Retrieval, Ranking, Exploration

- Jaspreet Singh: History by Diversity: Helping Historians Search news Archives. (+ Demo)
- Katja Markert: Predicting the social media popularity of news articles from headline text.
- Yannis Velegrakis: Coping with the User-Data Mismatch Problem.
- Helge Holzmann: Tempas – Temporal Archive Search Based on Delicious and Twitter. (Demo)
- Tu Nguyen: A Bing-inspired Ranking for Entity-centric Web Archive Search. (Demo)

Collaborative Exploration, Summarization

- Ivana Marenzi, Trevor Zeon: Working with Web Archive Collections. (+ Demo)
- Binh Giang Tran: Timeline Generation: A View from the Temporal Analytic Aspect.

Using Web Archives for Research

- Elisabeth Niggemann: Archiving-on-demand – cure against link rot in scholarly citations?
- Niels Brügger: Web History, Web archives, and Web research infrastructure – between close and distant read.
- Thomas Risse: SoBigData – Social Mining and Big Data Ecosystem.
- Richard Rogers: The crisis in Web archive use, and approaches that address it.
- Gerhard Gossen: The iCrawl Wizard – Supporting Interactive Focused Crawl Specification. (Demo)
- Simon Gottschalk: Analyzing Language-specific Differences in Multilingual Wikipedia. (Demo)