

Exploring Web Archives Through Temporal Anchor Texts

4th Int. Alexandria Workshop

Helge Holzmann, Wolfgang Nejdl, Avishek Anand

L3S Research Center, Hannover, Germany

19/10/2017

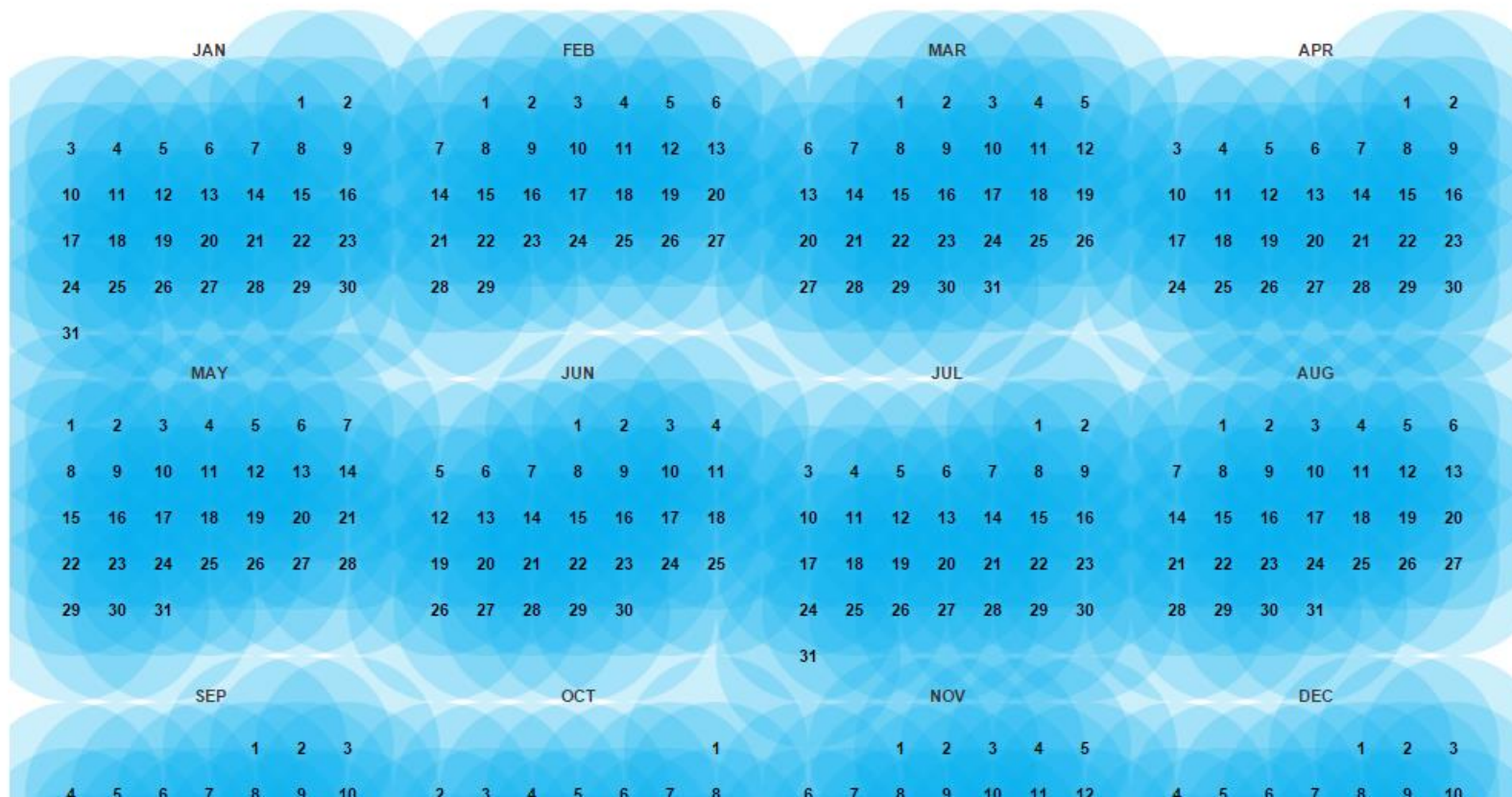


Helge Holzmann (holzmann@L3S.de)



Saved **49,342 times** between November 12, 1996 and February 12, 2017.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



Subscribe to The Times.



The New York Times

Wednesday, November 7, 2012 Last Update: 6:07 PM ET

Subscribe to The Times.

HOME DELIVERY

DIGITAL SUBSCRIPTION

Search

Follow Us f | | Personalize Your Weather

President »



Senate »



House »



FULL RESULTS

- President Big Board »
- Exit Polls »
- National Map »

WORLD
U.S.
POLITICS
NEW YORK
BUSINESS
DEALBOOK
TECHNOLOGY
SPORTS
SCIENCE
HEALTH
ARTS
STYLE
OPINION

Autos
Blogs
Books
Cartoons
Classifieds
Crosswords
Dining & Wine
Education
Event Guide
Fashion & Style
Home & Garden
Jobs
Magazine
Movies
Music
Obituaries
Public Editor
Real Estate
Sunday Review

Obama Wins a Clear Victory, but Balance of Power Is Unchanged in Washington

Boehner Strikes Conciliatory Tone in Talk of Fiscal Cliff

By JONATHAN WEISMAN and JACKIE CALMES 38 minutes ago

The speaker of the House said Wednesday he is ready to accept a budget deal that raises revenues if it is linked to an overhaul of the tax code and entitlements.

- Boehner's Offer May Bring Sides to Table 6:03 PM ET

NEWS ANALYSIS

Debate Over Size of Federal Government Left Unresolved

By PETER BAKER 3:15 PM ET

After \$6 billion, two dozen presidential primary days, four general election debates and more TV ads than anyone could watch, the two parties essentially fought to a standstill.

Comments



Today's TimesCast

Republicans search for a path forward. | Changes in store for Obama's cabinet.

Play Video »

G.O.P. Factions Grapple Over Meaning of Loss

By MICHAEL COOPER 4:15 PM ET

There was no shortage of theories from inside and outside the party about where it fell short and what to do next.

Day After Election, a Sharp Loss on Wall Street

By NELSON D. SCHWARTZ 4:50 PM ET

Stocks moved sharply lower in New York, with both major indexes down 2.4 percent.

Marriage Gains Cheer Gay Rights Advocates

By ERIK ECKHOLM 12:20 PM ET

Supporters called votes in Maine and Maryland a turning point. Results in Washington State were still being tallied.

MORE ELECTION NEWS

- Californians Back Taxes to Avoid Education Cuts
- Democrats Gain in Senate, Adding to Majority
- Republicans Shore Up Incumbents, Holding House
- Why The Times Was Slower in Calling Election
- Twitter Kills the Fall Whale, One Tweet at a Time

The Opinion Pages

EDITORIAL

Marriage Equality

The nation took a step toward full equality for all Americans with a string of victories for gay marriage.

FIXES

A Change in the Weather on Wall Street

Let's act on the renewed sense of urgency about climate change.



OP-ED COLUMNISTS

- Bruni: Maine and Maryland Say 'We Do'
- Douthat: The Obama Realignment
- Keller: Mo'bama
- Friedman: Hope, Part 2
- Brooks, Collins: Orca, Meet Obama

MORE IN OPINION

- Editorial: Obama's Win
- Op-Ed: Unwelcome at the Party

MARKETS » last week's change

JAPAN	HANG SENG	CHINA
Nikkei	Hang Seng	Shanghai
19,376.93	23,574.98	3,197.33
Closed for holiday	Closed for holiday	Closed for holiday

Data delayed at least 15 minutes

GET QUOTES My Portfolio »

Stock, ETFs, Funds Go

BIRTHDAY GIFT

THANK YOU GIFT

The Wayback Machine

- Replays Web resources with a temporal dimension
 - Identified by **URL** and **timestamp** (crawl time)

```
http://web.archive.org/web/20121107020708/http://www.nytimes.com/
```

```
http://web.archive.org/web/TIMESTAMP/URL
```

- **Challenges** for the user
 1. Find the relevant timestamp
 - At what date / time was the webpage / content of interest online / crawled?
 2. Discover the desired resources
 - What is the URL of the webpage / content of interest at the relevant date / time?

Temporal Search in Web Archives

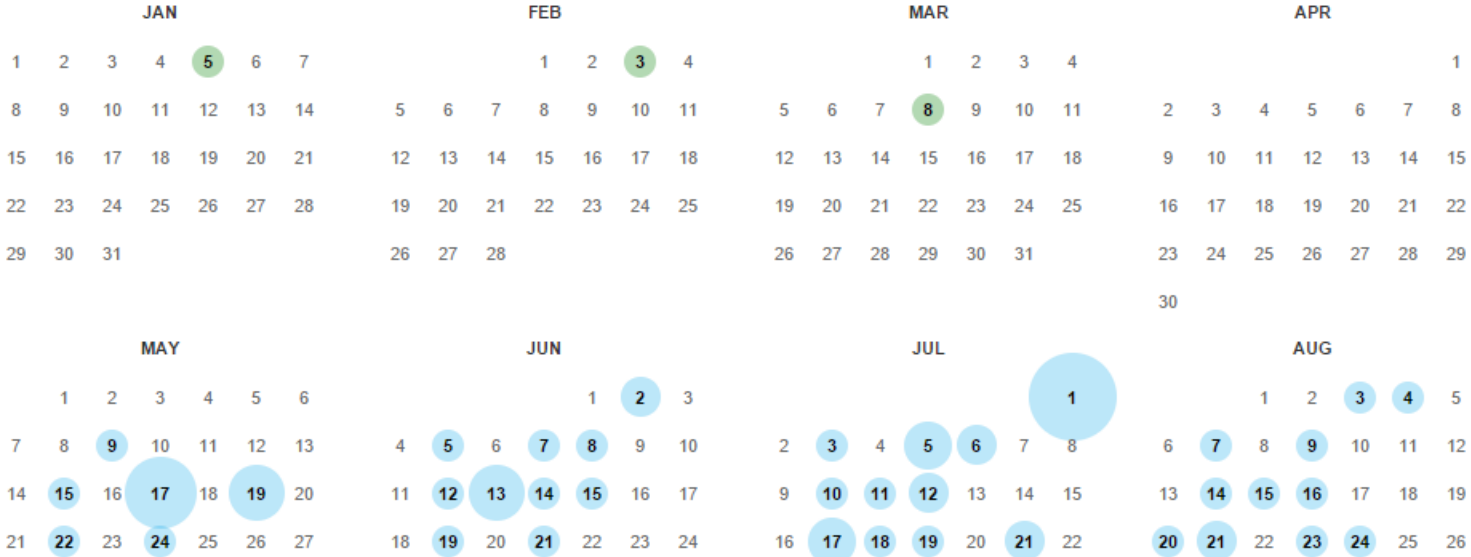
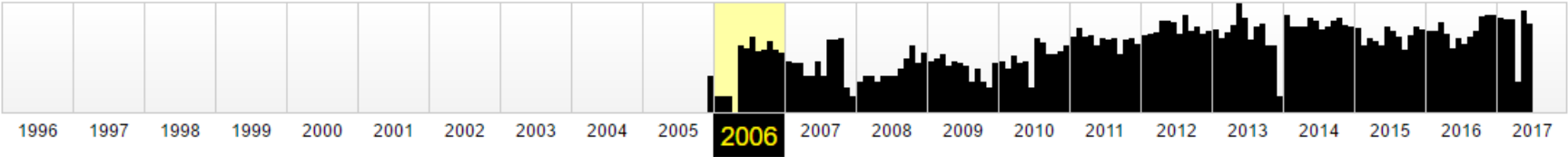
- Documents are temporal / consisting of **multiple versions**
 - Version / snapshot / capture represents a crawl
 - A version may be a duplicate of a previous one
 - Or it may contain slight or drastic changes (might be a completely new page)
- **Temporal relevance** in addition to textual relevance
 - Temporal relevance is not always encoded in the content
 - Very little text snippets or changes may be of high importance
- Resource identifiers (i.e., URLs) may **change over time**
 - A webpage moved to a new URL makes it hard to detect previous versions
- Information needs / **query intents are different** from traditional IR
 - There is no clear understanding of what is (temporally) relevant

Explore more than 284 billion web pages saved over time

Saved 3.992 times between December 6, 2005 and June 11, 2017.

Summary of europa.eu

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



INTERNET ARCHIVE

WayBackMachine

europa.eu.int

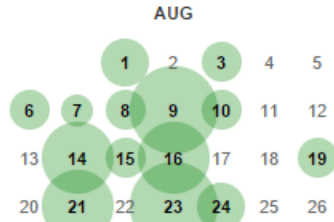
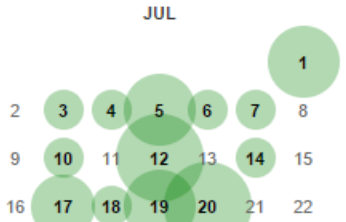
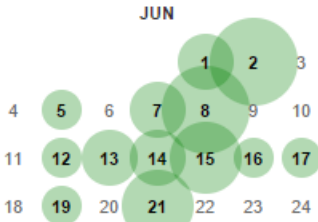
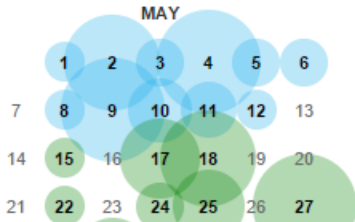
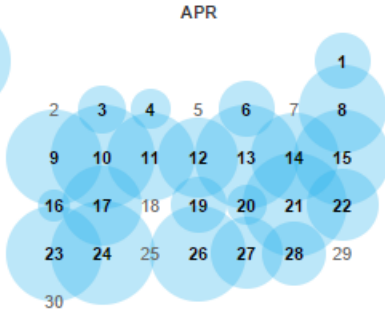
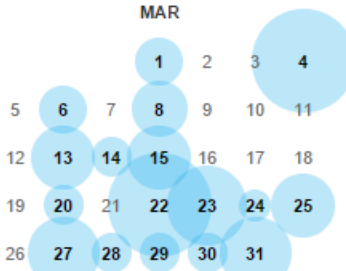
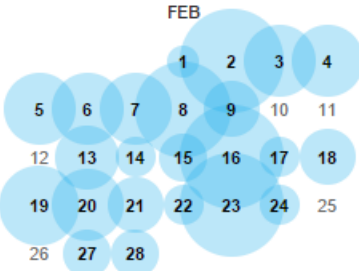
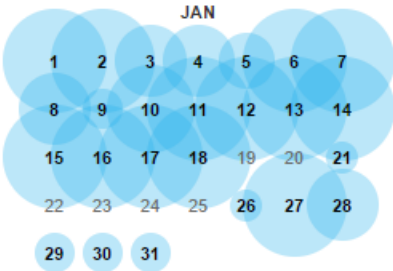
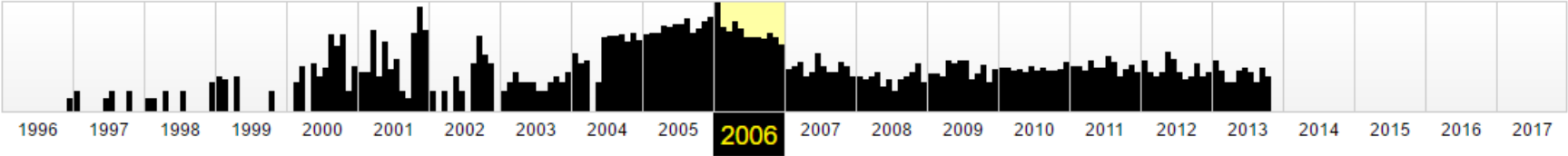


Explore more than 284 billion web pages saved over time

Saved 5,567 times between December 19, 1996 and October 16, 2013.

Summary of europa.eu.int

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



Temporal Archive Search (Tempas)

- **Goal:** find URLs / entry points / authority pages over time
 - most *central* URLs of an entity / topic in a given time
- **Idea:** exploit *external* information to detect temporal relevance
 - capture temporal keywords / descriptors from external data (surrogates)
- **v1:** based on **tags** from *Declicious* (tempas.L3S.de/v1)
 - uses temporal frequencies of social bookmarks as proxy for temp. importance
 - biased by Delicious users, only limited available data for 8 years
- **v2:** based on the **hyperlink graph of the Web** (tempas.L3S.de/v2)
 - less biased, more data, growing with the Web archive

Tempas v1 (tempas.L3S.de/v1)

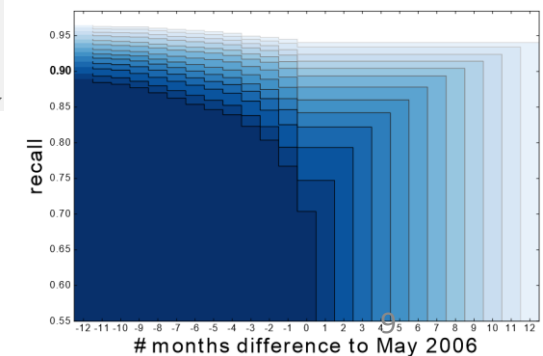
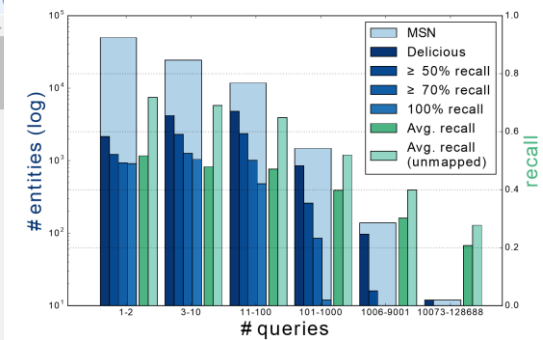
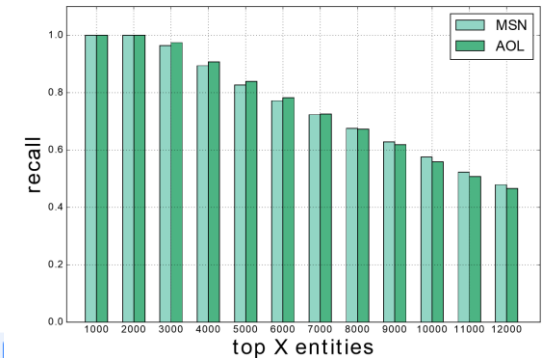


The screenshot shows the Tempas v1 web interface. At the top, the word "Tempas" is written in a large, stylized blue font. Below it, there is a search bar with the text "obama" and a "Go!" button. To the right of the search bar are two horizontal timeline sliders. The top slider spans from 2003 to 2011, and the bottom slider spans from January to December. Below the sliders is a horizontal navigation bar with various tags such as "politics", "election", "2008", "america", "afghanistan", "2009", "article", "art", "analysis", "activism", "articles", "2010", "abortion", "acorn", "2008election", "@home", "advertising", "administration", "american", "2011", "08", "08election", "and", "asshole", "apple". On the left side, there is a list of search results for "obama", including links to "obama government blog 2008 barack awesome america 2009 activism barack-obama" and "http://change.gov/". The main content area displays a screenshot of the "CHANGE.GOV" website, which is the official website of the Office of the President-Elect. The website features a quote from President-Elect Barack Obama: "Today we begin in earnest the work of making sure that the world we leave our children is just a little bit better than the one we inhabit today." Below the quote are navigation links: "HOME", "NEWSROOM", "BLOG", "LEARN", "AGENDA", "AMERICAN MOMENT", "AMERICA SERVES", "JOBS", "ABOUT", and a "SEARCH" button. At the bottom of the website screenshot is a banner for "AN AMERICAN MOMENT" with the text "The story of this campaign is your story."

- UI with support for temporal queries and temporal results

[Helge Holzmann, Avishek Anand. *Tempas: Temporal Archive Search Based on Tags*. WWW 2016]

[Helge Holzmann, Wolfgang Nejdl, Avishek Anand. *On the Applicability of Delicious for Temporal Search on Web Archives*. SIGIR 2016]

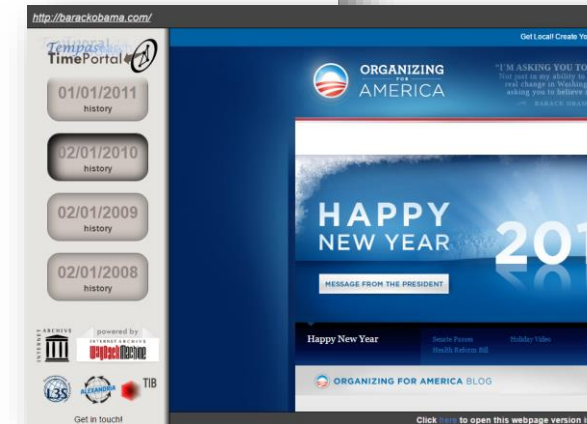
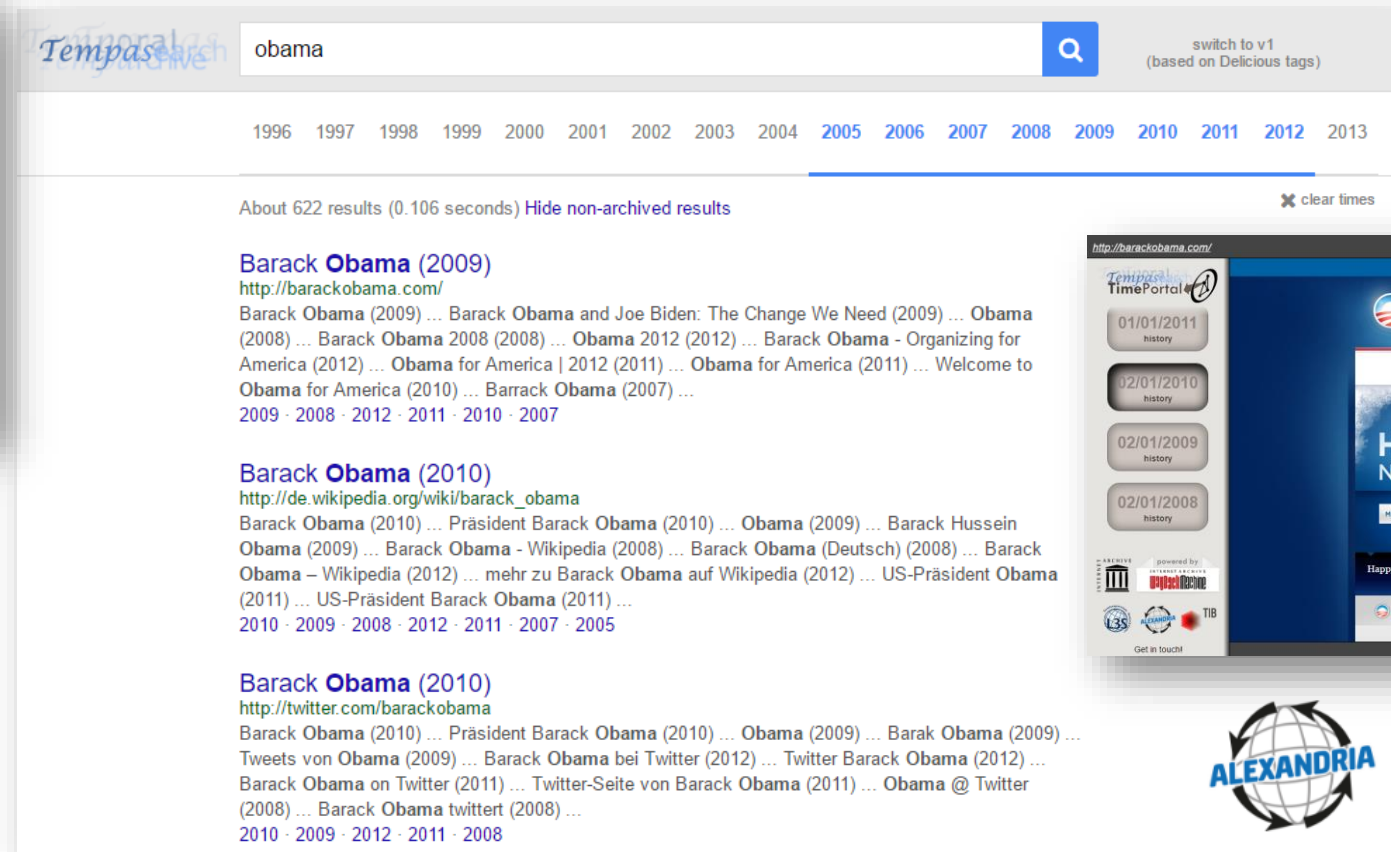


Tempas v2 (tempas.L3S.de/v2)



• From tags to **anchor texts**

- Delicious is biased
- Available data is limited
- Tags do not allow for **natural keyword search**
- Anchor texts serve well for the task of **temporal navigational queries**



Temporal Retrieval Model

- Based on **temporal hyperlink graphs** with yearly granularity
 - only emerging links in time period $[t_a, t_b]$ are considered

$$L_{\text{emergence}} = \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p \mid t_a \leq t \leq t_b\} \setminus \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p \mid t < t_a\}$$

- Relevance is determined by **hosts** linking to a URL v w.r.t. anchor text a

$$\text{freq}(v, a) = |\{host(u) \mid e = (u, v) \in E \wedge (e, a) \in L\}|$$

$$\text{rel}(v, a) = \log\left(\frac{\text{freq}(v, a)}{\max_{v \in V, a \in A} \text{freq}(v, a)} \cdot \gamma\right)$$

Index Construction and Retrieval

- Based on the **German Web archive** (TLD .de) from 1996 to 2013
 - > 2 billion distinct archived pages linking to 26,443,384,902 URLs (not only .de)
 - **319,574,156 URLs** indexed after cleaning infrequent and malformed links
- Anchor texts classified by **relevance class** per year: $\varphi = relc(v, a) = \lfloor rel(v, a) \rfloor$
 - Weighted by maximum frequency in a year: $boost(\varphi) = \varphi \cdot \log(\max_{v \in V, a \in A} freq(v, a))$
- Indexed into **Elastic Search**
 - One full-text index per year and relevance class
 - Scored by tf-idf with boost

```
{  
  "url": "http://.../wiki/World_Wide_Web",  
  "years": {  
    "2013": {  
      "8": ["World Wide Web", "WWW", ...],  
      "7": ["internet", ...]  
    },  
    "2012": {  
      ...  
    },  
    ...  
  }  
}
```

Tempas v2 Example Queries (1)

- Barack Obama

obama @ [2005, 2006]

1. <http://obama.senate.gov>
 2. http://de.wikipedia.org/wiki/barack_obama
-

obama @ [2005, 2007]

1. <http://myspace.com/barackobama>
 4. <http://obama.senate.gov>
 5. <http://youtube.com/profile?user=barackobamadotcom>
-

obama @ [2008, 2013]

1. <http://barackobama.com>
 2. http://de.wikipedia.org/wiki/barack_obama
 3. <http://twitter.com/barackobama>
-

- Angela Merkel

merkel @ [2000, 2004]

1. <http://merkel.de> (*university bookstore Merkel*)
 2. <http://angela-merkel.de>
-

angela merkel @ [2000, 2004]

1. <http://angela-merkel.de>
 2. <http://cdu.de/idx-merkel.htm>
 3. <http://cdu.de/ueber-uns/buvo/pv/pv.htm>
-

merkel @ [2005, 2010]

1. <http://angela-merkel.de>
 2. http://de.wikipedia.org/wiki/angela_merkel
-

merkel @ [2010, 2013]

1. <http://angela-merkel.de>
 2. <http://facebook.com/angelamerkel?...>
 3. http://de.wikipedia.org/wiki/angela_merkel
 4. <http://twitter.com/search?q=%23merkel>
-

Tempas v2 Example Queries (2)

• European Union

european union @ [1996, 2005]

1. <http://europa.eu.int>

european union @ [2005, 2013]

1. http://en.wikipedia.org/wiki/european_union
2. <http://europa.eu>
3. <http://europa.eu.int>

• Wikipedia

wikipedia @ [2001, 2002]

1. http://de.wikipedia.com/wiki.cgi?wikipedia_willkommen
2. <http://wikipedia.com>
3. <http://de.wikipedia.com>

wikipedia @ [2003, 2013]

1. <http://de.wikipedia.org>
2. <http://de.wikipedia.org/wiki/hauptseite>
3. <http://wikipedia.de>
3. <http://wikipedia.org>

• Creative Commons License

creative commons license @ [2002, 2003]

1. <http://creativecommons.org/licenses/by-nc-sa/1.0>
2. <http://creativecommons.org/licenses/by-nd-nc/1.0>

creative commons license @ [2004, 2006]

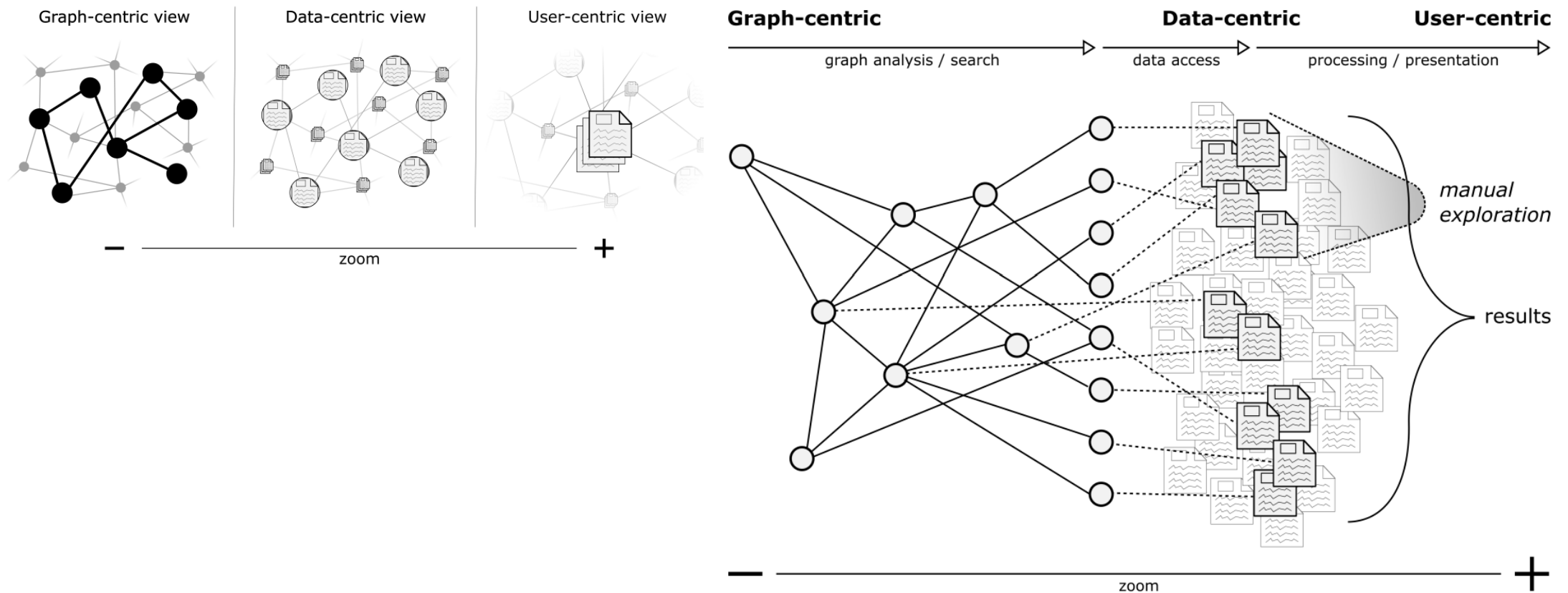
1. <http://creativecommons.org/licenses/by-nc-sa/2.0>
2. <http://creativecommons.org/licenses/by-nc-nd/2.0>

creative commons license @ [2007, 2013]

1. <http://creativecommons.org/licenses/by/2.5>
2. <http://creativecommons.org/licenses/by/3.0>
3. <http://creativecommons.org/licenses/by-nc-sa/3.0>

Tempas Results as Entry Points

- From temporal search to **temporal data analysis**



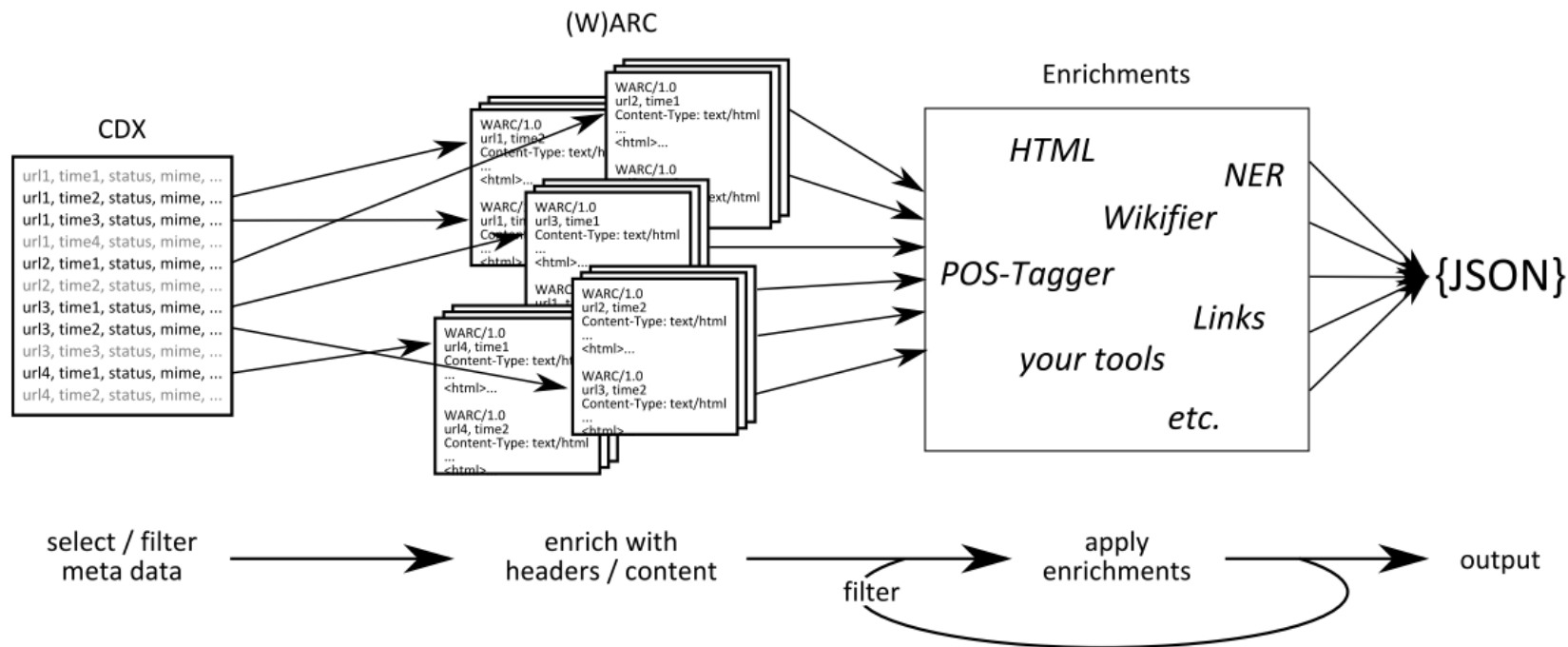
ArchiveSpark

- **Expressive** and **efficient** Web archives data access / processing
- Joint work with the **Internet Archive**
- **Open source**
 - Fork us on **GitHub**: <https://github.com/helgeho/ArchiveSpark>
 - Star, contribute, fix, spread, **get involved!**
 - Easily **extensible**
- More details in:
 - *Helge Holzmann, Vinay Goel, Avishek Anand.*
ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation.
In Proceedings of JCDL, Newark, New Jersey, USA, 2016.



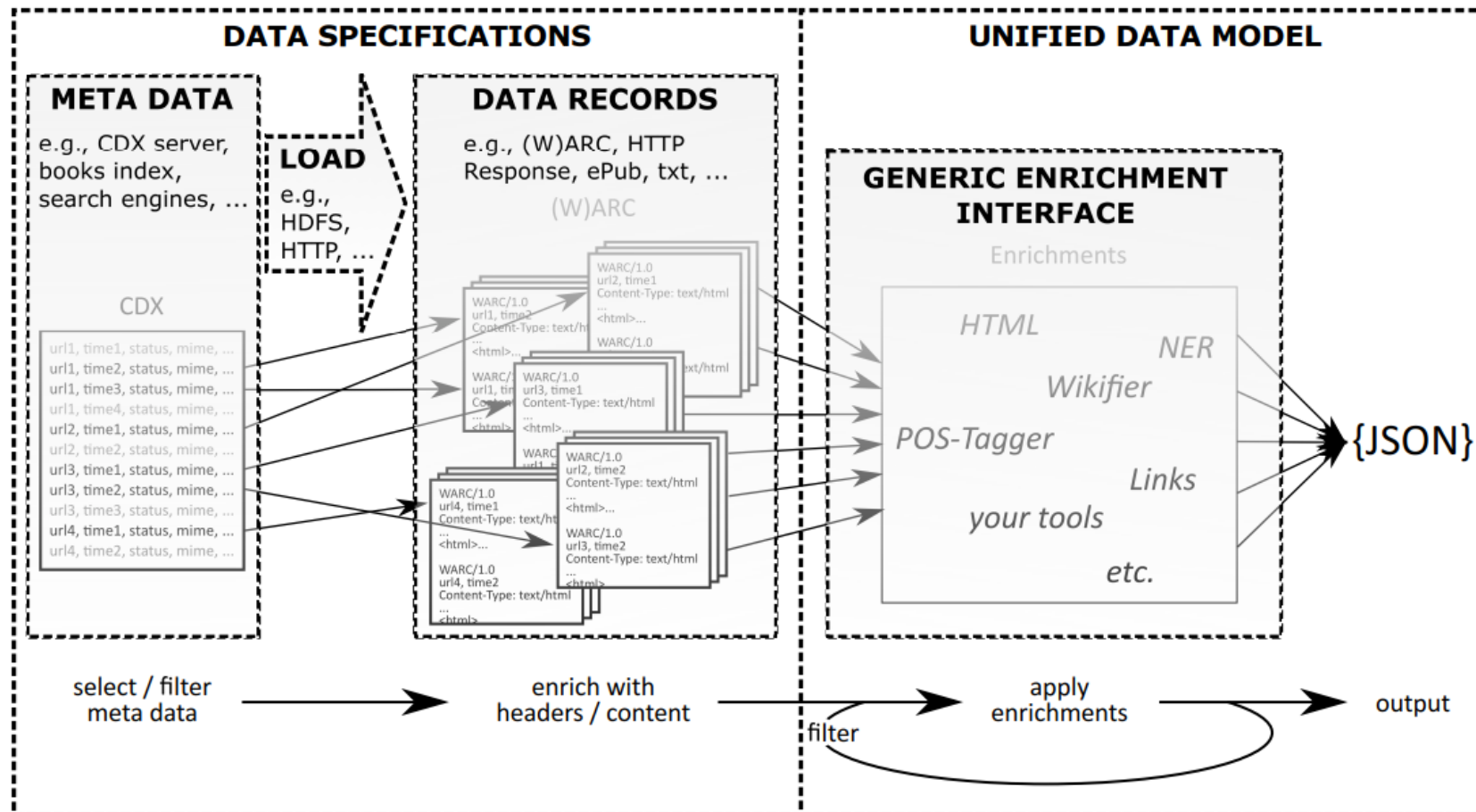
Efficient Processing with ArchiveSpark

- Seamless **two step loading** approach
 - Filter as much as possible on metadata before touching the archive



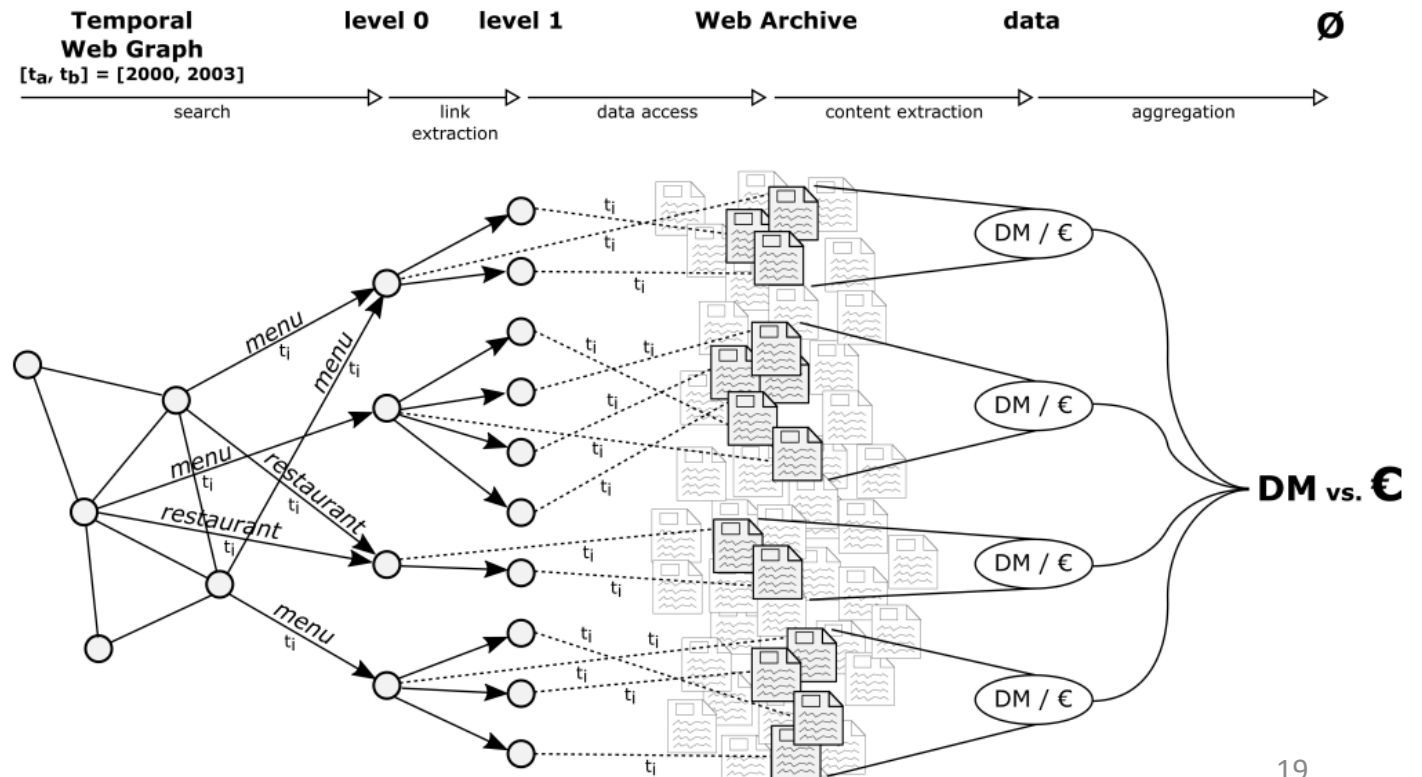
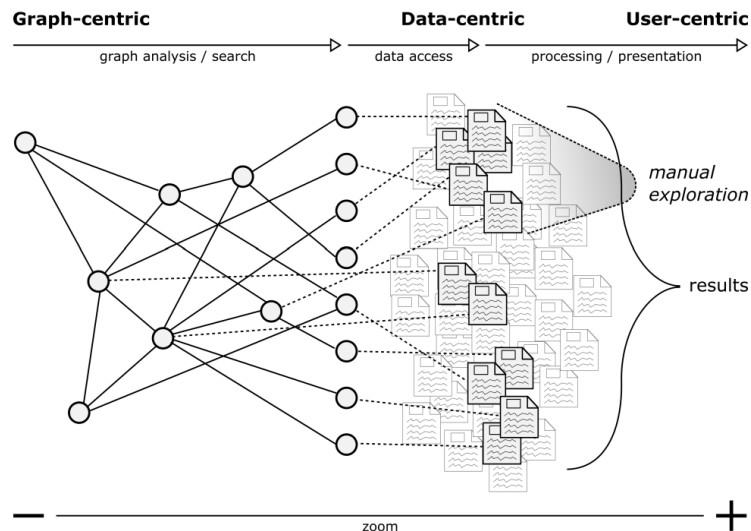
ArchiveSpark 2.0

- **Data specifications** allow for more data sources



Data Analysis with Tempas

- Integrating Tempas search results with ArchiveSpark
 - Identifying **entry points** for efficient Web archive access
- Example study
 - € restaurant prices
 - Average increase of 23%**



Conclusion and Future Work

- **Efficient access** is crucial when working with Web archives
 - Specialized retrieval methods are required to find suitable entry points
- Anchor texts are **valuable (temporal) pointers** into Web archives
 - Good (lightweight) alternative to full-text search for navigational queries
 - To find URLs of interest for an entity / object in a given time period
 - As entry points for a more detailed temporal exploration / analysis (at scale)
- **Quantitative evaluation** is difficult
 - Available judgements not suitable for this (new?) task
 - e.g., phoenix.edu irrelevant for query 'university of phoenix' in TREC 2012 Web Track
 - We are working on it, **stay tuned**



Thank you!

- www.L3S.de
- www.ALEXANDRIA-project.eu
- tempas.L3S.de/v2
- github.com/helgeho/Tempas2ArchiveSpark
- github.com/helgeho/ArchiveSpark

Questions?

Tempas search
Temporal archive

