Challenges and Approaches for Web Archive Creation and Usage

Thomas Risse
L3S Research Center

Congress of Catalan Archivists 2015
Lleida, 29. May 2015
The Web is a core part of our daily life

World Wide Web = 50 Bill. Pages + 1 Bill. Users

The Web and the Social Web play a crucial role

- Information and services for all domains
- Allows contributions by every citizen
- Giving room for the articulation for a multitude of stakeholders
- Reflects all types of events, opinions, developments within society, science, politics, environment, business, ...
Are we loosing the past of the web?

Spam

Gun running from Sudan

EgyTweets

RT @RSSEgypt: Egypt Jan25

آحكام عادلة ومن 3 لـ 10 سنوات ضد 25

البلجية والسينية بالإكراد وخرو حظر التجوال

http://dlvr.it/KCHz8 #tahrir

about 1 hour ago via EgyTweets

RT @RSSEgypt: Egypt Jan25

على حديث الأمة رقم 25

http://dlvr.it/KCH84j #tahrir

about 1 hour ago via EgyTweets

RT @SofufunkLA: comes rough, tough like an elephant tusk. Ya head rush fly like Egyptian music...

about 1 hour ago via EgyTweets

RT @DavoodDw: DailyNewsEgypt: Egypt广泛的国际间和国际间

arms from Sudan http://tinyurl.com/4x066x

about 1 hour ago via EgyTweets

RT @AlMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders http://tinyurl.com/4erDv4 #Attech #Egypt

about 1 hour ago via EgyTweets

RT @echyeyt: Intel buys Egypt-Based SySDSat to boost its 4G

LTE efforts. http://tinyurl.com/4brn4h

about 1 hour ago via EgyTweets

Conservative party deletes archive of speeches from internet

Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh and Alex Hern

The Guardian, Wednesday 13 November 2013 15.40 GMT

Jump to comments (1284)

A speech in which David Cameron said the internet would help people hold politicians to account was among those deleted. Photograph: Barcroft Media
The Web is Changing and Forgetting

The Web is a quickly changing, ever growing information space [1]
– It’s growing by >8% per week
– After 1 year only 40% of the pages are still accessible while 60% of the pages are new

→ A Web Archive as a Collective Memory is a cultural necessity for the future
→ But „Archive and Store Everything“ is not a practical approach

Where are we now?

- A number of tools are available (e.g. Heritrix)
- Crawl descriptions are currently lists of URLs
- >42 world wide Web Archives initiatives with different scopes [2]
- Still a lot manual effort is necessary but only some 100 people are involved world wide [2]
- Increasing interest from Digital Humanities, Journalism, etc.

More support is necessary

- Crawl by Events, Topics and Entities
- Using the “Wisdom of the Crowds” for selection and appraisal

Agenda

- **Motivation**
- **The Crawl Process and its Challenges**
  - Process Overview
  - Dynamic Pages
  - Basic Crawl Strategies
  - Termination of Crawls
- **Next Generation Web Archiving**
  - Requirements
  - Topical Crawling
  - The ARCOMEM Approach
  - iCrawl - Integrated Crawling
- **Web Archive Access and Usage Examples**
  - Current methods
  - The ARCOMEM Approach
  - Time and Topic aware diversification
  - Language Evolution
- **Conclusions**
Agenda

- **Motivation**
- **The Crawl Process and its Challenges**
  - Process Overview
  - Dynamic Pages
  - Basic Crawl Strategies
  - Termination of Crawls
- **Next Generation Web Archiving**
  - Requirements
  - Topical Crawling
  - The ARCOMEM Approach
  - iCrawl - Integrated Crawling
- **Web Archive Access and Usage Examples**
  - Current methods
  - The ARCOMEM Approach
  - Time and Topic aware diversification
  - Language Evolution
- **Conclusions**
Web Archiving and some challenges

Contributions the European Projects

LiWA
Living Web Archives

arcomem

Deep Web

Archiving

Data Volume
Temporal Aspects

Index
Storage

Dynamic Pages
JavaScript, Flash

Link Extraction
Fetching

Multimedia Content

Temporal Coherence of Crawls

Social Media
API Crawling

Temporal Aspects

User

Content Appraisal

Selection
Preparation
Discovery
Filtering
Capture
Link Extraction
Extraction
Fetching

Access

Quality Review

Archivist
Web Archiving and some challenges

- **Selection**
  - Preparation
  - Discovery
  - Filtering
- **Content Selection**
- **Content Appraisal**
- **Access**
- **Deep Web**
- **Archiving**
  - Index
  - Storage
- **Capture**
  - Link Extraction
  - Fetching
- **Dynamic Pages**
  - Javascript, Flash
- **Temporal Coherence of Crawls**
- **Social Media API Crawling**
- **Data Volume**
- **Temporal Aspects**
- **Quality Review**
- **Noise Filtering**
- **User**
- **Archivist**

- **Web Science – Investigating the Future of Information and Communication**

Congress of Catalan Archivists 2015

Thomas Risse

25/11/15
Dynamic Pages
Follow the links....

What is GNU?

GNU is an operating system that is free software. That is, it respects users' freedom. The development of GNU made it possible to use a computer without software that would trample your freedom.

We recommend installable versions of GNU (more precisely, GNU/Linux distributions) which are entirely free software. More about GNU below.

<a href="http://www.gnu.org/philosophy/free-sw.html">free software</a>
But what about these pages?

```javascript
function() {
    var fcb_referrers = [
        "fcbarcelona.com", "fcbarcelona.cat", "fcbarcelona.es",
        "fcbarab.com", "fcbarcelona.fr", "fcbarcelona.jp",
        "fcbarcelona.cn", "fcbarcelona.qq.com"
    ];
    var current_language = "en";
    var routes = {"ca": "http://www.fcbarcelona.cat"};
    var navigator_language = navigator.language || navigator.userLanguage;
    var language = (navigator_language || "").split("-")[0];
    var s = "^https?:\/[\^\]/\[^\]/]*\(\+fcb_referrers.join("\|\")\).replace(\./g, "\." + ")");
    var fcb_referrer = document.referrer.match(new RegExp(s));
    var redirect_url = routes[language];
    if (language && language != current_language && !fcb_referrer && redirect_url) {
        window.location = redirect_url;
    }
}();
```
“Endless” Pages

Automatically reload content while browsing the page

- Requires execution of JavaScript
- When can the fetching be stopped?
- Examples: Facebook, Twitter, Tumblr, etc.
Embedded Social Media

What Do You Think?

Embedded Social Media Content

- JavaScript embeds
- Dynamically integrated content on the browser side
- Often standard scripts
  → Possibility to extract parameters
- Content fetching requires API Crawler
Handling of Dynamic Pages

The Problem

• Embedded “Programs”
  • Adobe Flash impossible to handle
  • JavaScript is readable
• Dynamic creation of links and content

Approaches

• Guessing of links
  ■ “guessing” by assembling any fragments that look like links into URLs
  ■ Can be very noisy - lots of wrong URL’s
• Extraction of parameters from program code
  ■ Applicable for known code libraries e.g. Facebook, Twitter
• Execution of JavaScript
  ■ Simulate user activities- “pressing” the links and see what comes out
  ■ Execute code in a Javascript engine (e.g. WebKit, Firefox Browser)
  ■ Extract links from resulting DOM tree

Status

• Some implementations exist (e.g. LiWA Service, Browser Monkey)
• API Crawler exist
• No integration into standard Web Crawler
Basic Crawl Strategies
Crawl Strategy
Crawl Strategy: Depth-first
Crawl Strategy: Breadth-first

- www.news.de
- about.htm
- journalists.htm
- ukraine_crisis.htm
- sports.htm
- ukraine0105.htm
- ukraine0205.htm
- spain_sports0205.htm
- thomas.htm
- joe.htm
- ukraine_sports0105.htm
Termination of a Crawl

When is the harvesting of Web Pages complete?

Problem: The total amount of pages is unknown

Typical Approach

• Crawler Queue is empty (only for very focused crawls)
• Termination after number of pages/amount of data
• Termination after time

→ Most often incomplete crawls depending on the strategy
Possible Consequences after stopping the Crawl

Depth-first

Breadth-first
Alternative Crawl Strategies

- Selection by Popularity
  - Ranking of the waiting queue according to the popularity of the content
  - Requires knowledge about the Web Graph
    - For example: PageRank
  - Mainly useful to optimize regular crawls

- Content based selection
  - Topics, Events, Entities
  - Requires semantic crawl specification

- The selection of the right strategy depends on the crawl intention

Will be discussed afterwards
Agenda

- **Motivation**
- **The Crawl Process and its Challenges**
  - Process Overview
  - Dynamic Pages
  - Basic Crawl Strategies
  - Termination of Crawls
- **Next Generation Web Archiving**
  - Requirements
  - Topical Crawling
  - The ARCOMEM Approach
  - iCrawl - Integrated Crawling
- **Web Archive Access and Usage Examples**
  - Current methods
  - The ARCOMEM Approach
  - Time and Topic aware diversification
  - Language Evolution
- **Conclusions**
Growing Scientific Interest in Web Archive Content (1/2)

**Historians**
- Official Publications (e.g. Government)
- Journalistic Resources
- Important topics and events with a high media coverage
- Multi-cultural or controversial topics
- Optimal: continues observation of topics in the Web

**Social Sciences**
- Observations of topics and events on major sites are good starting points
- Identified Topic
  - Official publications, journalistic and social media sources
  - Changes on the topic should be identified
- Metadata / Context (e.g. Author, Organizations and their interests, gender, location)
- Demographic information about social sites
- Provenance: Transparent and detailed documentation of content selection
Growing Scientific Interest in Web Archive Content (2/2)

Law

- Research is based on official publications and protocols of parliaments or comments released by publishers

- Social media (especially blogs) are increasingly used
  - Only used as background information
  - Reason: missing citability and authenticity of resources

- Genesis of laws
  - Used to understand original intention of laws
  - A democratic system requires a complete documentation of the law genesis.
  - Currently different degrees of documentation
    - Official publications: parliament and committee meetings
    - Public discourse
Derived Requirements

Topical Dimension
- Crawl intention are mainly focused around events and rarely around entities
- What is the intention of the researcher?
- Easy monitoring by the researcher and possibility to correct

Flexible Crawling Strategies
- Shallow observation crawls
- Focused crawls with prioritization (e.g. PageRank and/or semantics)

Social Web Crawling
- General interest with different media focus
- Integrated with Web crawler to capture the full context

Authenticity
- See a web page as the user saw the page (e.g. including ads and tweets at that time point)

Context and Provenance
- Demographics of sites
- Documentation of crawl specification and history
Topical / Event Crawling

Crawl Specification
- Terms
  - Ukraine
  - Crisis
  - ...
- Seed List
  - www.news.de
  - ...

Graph showing the crawling process with web page nodes and links.
ARCOMEM Crawling Phases

Entities
Obama, Romney, Biden, Ryan, Republicans, Democrats, ...

Keywords
US Election, CommitToMitt, Teaparty, Budget deficit, ...

Social Media Seedlist

Seedlist
http://news.bbc.co.uk/, http://telegraph.co.uk/, ...

Internet Crawling

Crawling Online Processing Offline Processing

Offline Processing

ARCOMEM Storage

SARA for Broadcaster, Parliaments

Cross Crawl Processing

Appraisal Selection

Archive

Obama, Romney, Biden, Ryan, Republicans, Democrats, ...

US Election, CommitToMitt, Teaparty, Budget deficit, ...

http://news.bbc.co.uk/, http://telegraph.co.uk/, ...


Crawling

Online Processing

Entities
Keywords
Social Media Seedlist
Seedlist

Internet

ARCOMEM Crawling Phases

Congress of Catalan Archivists 2015

Thomas Risse

25/11/15
Architecture Overview

Application

SARA
- Broadcaster
- Parliament

WARC Files + SOLR Index

WARC Export

Crawler Cockpit

Cross Crawl Analysis


Twitter Dynamics

Social Web Analysis

Image/Video Analysis

GATE Offline Analysis

Consolidation Enrichment

Online Processing

GATE Online Analysis

Social Web Analysis

Relevance Analysis & Prioritization

Resource Selection & Prioritization

Intelligent Crawl Definition

Offline Processing

Extracted SocialWeb Information

URLs

Relevance Analysis & Prioritization

Resource Selection & Prioritization

Intelligent Crawl Definition

Queue Management

Resource Fetching

Application-Aware Helper

Intelligent Crawl Definition
Crawler Cockpit

CRAWL DESCRIPTION:

Title: US elections 2012

Description: The United States presidential election of 2012 was the 57th quadrennial presidential election. It was held on Tuesday, November 6, 2012.

Comments:

SETUP YOUR CRAWL:

- Person: Barack Obama, Last name: Obama, Gender: Male, Age: 51-60 Years
- Event: US elections 2012, Date: 2013-04-17
- Place: City: Washington, Country: United States
- Keywords: democrats, white house
- URL: http://www.barackobama.com/, Path
- URL: http://www.mittromney.com/, Path
Relevance Distribution Time

![Graph: Focused vs. Unfocused Crawling]

- X-axis: Crawled Docs x 1000
- Y-axis: Relative Relevance
- Blue line: Focused Crawling
- Red line: Unfocused Crawling

The graph illustrates the distribution of relevance over time for focused and unfocused crawling methods. The trends show that focused crawling generally maintains a higher relative relevance compared to unfocused crawling across the documented timeline.
New Strategies lead to …

Example: German Elections Crawl 2013

- Completely user defined crawl (German Broadcasters)
- Many focused terms and entities in German
- Many full names e.g. “Angela Merkel”
- Few less focused keywords in English
- 1st Phase
  - Many English pages with no relation to German Elections
- 2nd Phase
  - Refinement of Crawl Specification
    - Focused English terms
    - Last names instead of full names
  - Finding a good cut off point of archive threshold
  - Smaller result set with higher focus
... new Experiences ...

A good specification of the crawl intention becomes critical

- Crawl specifications are rather complex compared to seed lists
- More experiences, observations and analysis are necessary from Users and Developers
- Room for more sophisticated guidance in the 1st phase

Finding the right cut-off point

- Depends on the user requirements
- Highly focused $\rightarrow$ Smaller Archives but higher risk of losing interesting information
- Less focused $\rightarrow$ Lower risk of missing information but larger archives
- Translation into scoring of individual crawls
... and new Limitations

Social Media and Web Crawling are separate systems

• Process
  1. Crawling of Social Media content
  2. Extraction of Links
  3. Crawling of Web Pages

• Static integration of Social Media
• Uni-directional Path: Social Media → Web Content
• Missing Path: Web Content → Social Media

In Addition

• Complex system
• Required many changes in the Heritrix queue handling
Integrated crawling with the L3S iCrawl System

Apache Nutch based Web Archive crawler (under development)

- Learning the intention of the crawl
- Integration of Web and Social Media Crawling
- Content based monitoring of the crawl process
Web Science – Investigating the Future of Information and Communication

iCrawl Wizard

Ukraine

Search sources: [ ] Web [ ] Twitter

Web

Ukraine im Themenpecial: Lesen Sie hier alle Nachrichten und Hintergründe zur Ukraine-Krise und den Konflikt um die Abspaltung des russischen Teils des ...

Crisis in Ukraine: In depth news, commentary and analysis ...

Crisis in Ukraine: Get in深度 coverage of the country's political crisis and Russia's de facto annexation of Crimea, from FT.com

Ukraine News + Nachrichten / Ukraine aktuell
http://news.feed-reader.net/886-ukraine.html


Welcome to Ukraine
http://ukraine.org/

Includes press, maps, business contacts, tourism and links to sites in and about Ukraine.

Ukraine News Today - Ukraine

Twitter

The Interpreter auf Twitter: "Ukraine Live: NATO Secretary General Tells Russian Foreign Minister To End Support for Ukrainian Rebels http://t.co/c8EX5TkrP6"
https://twitter.com/Interpreter_Mag/status/600661018666769409

Keywords: Retweets, Medien, Schießen, Schießen, Übersetzung, Liste, Land, Folgen, Folge, Vodafone, Verifizierter, Liste

Entities: NATO, Secretary General, Tells, Russian, Foreign Minister, To End Support for Ukrainian Rebels, Bing, Haiti, Magyar, English, TIM, Fran, air, Michael, Russian, volunteer, Russian, army, Embed this Video

RT @MillerMENA: NATO is wasting its breath – neither Russia nor its proxies want peace in Ukraine https://t.co/MV3OaVcD

Civilian killed in intense Ukrainian army shelling of...
Example for Integrated Crawling

### Twitter #Ukraine Feed

- **Lars Olberg** @MissileMonitor - 5h  
  Media: #Ukraine may buy French #Exocet anti-ship missiles, missilethreat.com/media-ukraine

- **Michael Baskin** @michael3000 - 5h  
  #Ukraine accuses #Russia of sending in dozens of tanks, weapons into Russian separatists held areas fn.ws/1smw1Lp

**Fox News**

Ukraine accuses Russia of sending in dozens of tanks, weapons into...

Ukraine on Friday accused Russia of sending dozens of tanks and other heavy weapons into its rebel-controlled easternmost regions.

View on web

### Crawler-Queue

<table>
<thead>
<tr>
<th>ID</th>
<th>Batch</th>
<th>URL</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK1</td>
<td>1</td>
<td><a href="http://www.foxnews.com/world/2014/11/07/ukraine-accuses-russia-sending-in-dozens-tanks-other-heavy-weapons-into-rebel/">http://www.foxnews.com/world/2014/11/07/ukraine-accuses-russia-sending-in-dozens-tanks-other-heavy-weapons-into-rebel/</a></td>
<td>1.00</td>
</tr>
<tr>
<td>UK2</td>
<td>1</td>
<td><a href="http://missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles/">http://missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles/</a></td>
<td>1.00</td>
</tr>
<tr>
<td>UK3</td>
<td>x</td>
<td><a href="http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/">http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/</a></td>
<td>0.40</td>
</tr>
<tr>
<td>UK4</td>
<td>y</td>
<td><a href="http://missilethreat.com/turkey-missile-talks-france-china-disagreements-erdogan/">http://missilethreat.com/turkey-missile-talks-france-china-disagreements-erdogan/</a></td>
<td>0.05</td>
</tr>
</tbody>
</table>

### Extracted URL

<table>
<thead>
<tr>
<th>Web Link</th>
<th>Extracted URL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Agenda

• Motivation
• The Crawl Process and its Challenges
  • Process Overview
  • Dynamic Pages
  • Basic Crawl Strategies
  • Termination of Crawls
• Next Generation Web Archiving
  • Requirements
  • Topical Crawling
  • The ARCOMEM Approach
  • iCrawl - Integrated Crawling
• Web Archive Access and Usage Examples
  • Current methods
  • The ARCOMEM Approach
  • Time and Topic aware diversification
  • Language Evolution
• Conclusions
WayBackMachine

- Basic tool to access Web Archives
- Also used in combination with other tools
- Only URL based access
- Provides
  - Capture overviews
  - Details about crawl dates
Web Science – Investigating the Future of Information and Communication

Searching the Web Archive

Hostal Central Guesthouse in Triple Room Barcelona, Spain
25 December, 2009 - other dates
Hostal Central Guesthouse in Triple Room Barcelona, Spain Catalonia Mikado 3 hotel, stars ... star caters for Disabled. Single: From €65! Mikado Catalonia hotel, Stars, 3 Barcelona, Triple ... and Breakfast (B&B) in Barcelona, Catalonia, Spain. Between 25 Esteleda由于.com - 1. - Cornel Kart ...
http://www.hostaldelendaer.france.com/triple-room-barcelona.html

http://www.catalonia.com.br/catalunha_cultura5.asp

Versions of the archived the Web pages
Other Ways to Explore Web Archives

The Web Archive of Catalonia

Monographics
Monographic collections, or PADICAT monographs, are collections of digital resources around a specific event in the Catalan public life, or on a particular topic of political, social, cultural, etc., of Catalonia.

The selected resources are classified or ordered according to different patterns depending on the nature of each monograph.

Catalan elections (2006)
On the 1st November of 2006 the elections to Parliament were held in Catalonia. To understanding this key event for the Catalan society, we selected and harvested several versions of 61 online digital resources (political parties with and without parliamentary representation that presented candidature, personal blogs, web pages of the candidates, mass media, institutional digital resources, political foundations and research centers, etc) that have been incorporated to the Digital Heritage of Catalonia, to follow the evolution of the election campaign up in the Catalan Web.

Catalan elections (2010)
Since 2006, the Biblioteca de Catalunya has been working on the internet monographics about the elections carried out in Catalonia. The last 29th of November were held parliamentary elections, and consequently were harvested over 22.000 captures in different versions from 957 websites (blogs and websites of 245 candidates from 39 political parties with and without parliamentary representation, institutions and 56 media).

Catalan elections (2012)
According to the trend initiated on 2006 with the first monographic compilation of web resources for the elections to the Catalan Parliament that year, PADICAT has been carrying out a monographic for each electoral campaign involving the Catalan citizenship. As well for the elections to the Catalan Parliament that took place on November 25, 2012.

European Elections (2009)
The 7th elections to the European Parliament since 1979 were held in 27 member states of the EU between the 4th and the 7th June 2009. Among 491 millions European citizens, 375 millions were called up to the ballot box to elect 736 MEP. In Spain the elections took place on the 7th June.

Folk-rock (2008)
Internet has become a new mean of music distribution where the bands broadcast their sonorous proposals. The proliferation of the music in the net involves, some times, the absence of this same music in a physical format as, for example, a CD, and therefore, it has been edited by any Record company.

Local elections (2007)
Along the same lines as the special issue about the elections to the Catalan Parliament 1st November 2006, PADICAT has intensively monitored the electoral campaign for the local elections of 27th May Information was taken from the blogs and web pages of principal candidates for the mayor's office of 48 Catalan municipalities (i.e.
Web Archive Access

Different user groups with different needs
• Not the typical Web Search Engine user

Explorative Search
• Web Archive Search ≠ Web Search
• Time Dimension
• Many versions of the same page
• Comparing versions

Large scale analysis across time
• Information extraction (Entities, Events, Topics, Opinions)
• Interlinking
• Visualizations
The ARCOMEM Approach to Web Archive Access
Historical Search on News Archives*

Consider a journalist interested in the history of ....... Rudolph Giuliani

- News articles encode history as it happens.
- Aspects are diverse across time.
- Time windows can be diverse in aspects.

*Jaspreet Singh, Avishek Anand; Historical Search on News Archives; L3S Report 2015
Large Scale Data Analytics Example

Named Entity Evolution Recognition
Language changes over time!

Our language is DYNAMIC and changes with our culture, politics, technology, social media, etc.

- Different spellings over time
- New words are introduced
- Words change their meanings

In long-term digital archives → Problems!

<table>
<thead>
<tr>
<th>St. Petersburg</th>
<th>Petrograd</th>
<th>Leningrad</th>
<th>St. Petersburg</th>
<th>St. Petersburg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1914</td>
<td></td>
<td>1924</td>
<td>1991</td>
<td>today</td>
</tr>
</tbody>
</table>
Problem: Finding Documents

A scholar writing a thesis about "Pope Benedikt":

→ "I want to know more about Pope Benedikt"
Named Entity Evolution

**Joseph Ratzinger**

**Pope Benedict**

- **Named Entities (NE):** people, places, companies...
- **Characteristics of Named Entity Evolution (NEE):**
  - Same thing but different terms over time
  - Change occurs over short periods of time
  - Small or no concept shift
  - Announced to the public repeatedly
- **Goal:** Find method for named entity evolution recognition independent from external knowledge sources

**Change Period**
In his latest address to American bishops visiting Rome, Pope Benedict XVI stressed that Catholic educators should remain true to the faith — a reminder issued just in time for another tense season of commencement addresses. No, the pope did not mention Georgetown University by name when discussing the Catholic campus culture wars.

Evaluation Results
- Burst detection found total 73% of all change periods
- High recall for unsupervised method
- Machine Learning boosts precision
- Data Set: http://www.l3s.de/neer-dataset/
Motivation

• Optimal access to Web archives taking into account
  • the temporal dimension of Web archives
  • structured semantic information available on the Web
  • social media and network information

Objectives

• Evolution-Aware Entity-Based Enrichment and Indexing
• Aggregating Social Networks and Streams
• Temporal Retrieval and Ranking
• Collaborative Exploration and Analytics

Testbeds

• Temporal Wikipedia
• Academic Web Archive
• Politics on the Web

More information: http://alexandria-project.eu/
Conclusions

Web Crawling
• Mature standard Web crawlers but still many challenges (e.g. Social Media, Dynamics)
• Manual interventions will always be necessary
• New user communities with different interests and requirements
• Additional crawling strategies are necessary

Web Archive Access
• Current approaches far behind the State of the Art
• Increasing Web Archive interest will lead to increasing user expectations
• Different usage categories: Explorative search vs. Big Data analysis
• Legal Aspects are still a big issue
Thank You!

Dr. Thomas Risse
Forschungszentrum L3S
Leibniz Universität Hannover
Appelstrasse 9a
30167 Hannover, Germany

E-Mail: risse@L3S.de
Telefon: +49-511-762 17764
Telefax: +49-511-762 17779