



**LiWA**  
Living Web Archives

# From Web Page Storage to Living Web Archives

Thomas Risse

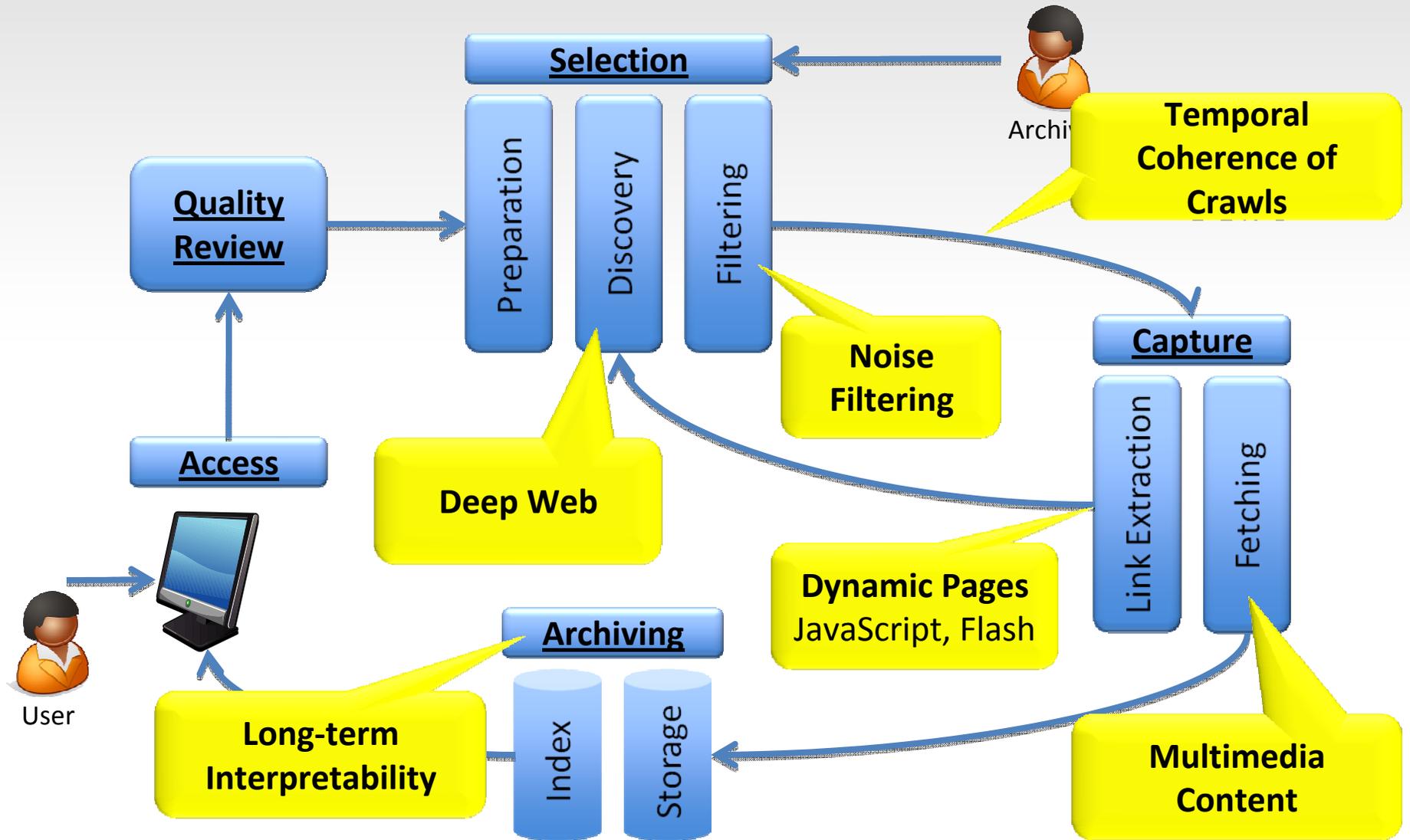
JISC, the DPC and the UK Web Archiving Consortium  
Workshop

British Library, London, 21.7.2009

# Agenda

- Web Crawling today & Open Issues
- LiWA – Living Web Archives Project
- Selected working areas of LiWA
  - Dynamic Pages
  - Handling of Spam
  - Temporal coherence of crawls
  - Archive Interpretability
- Conclusions and Expected Project Results

# Current Web Archiving at a Glance



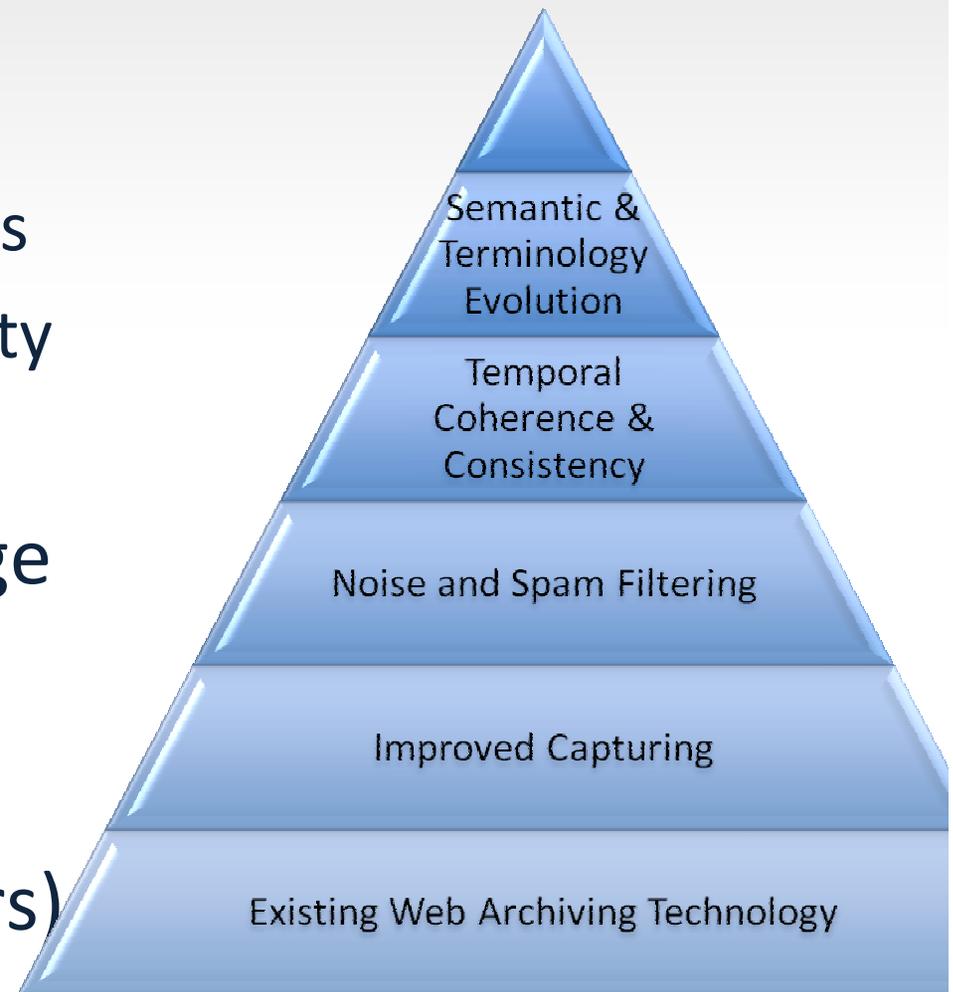
# LiWA – Living Web Archives (EU-IST 216267)

Next generation Web Archiving technology for:

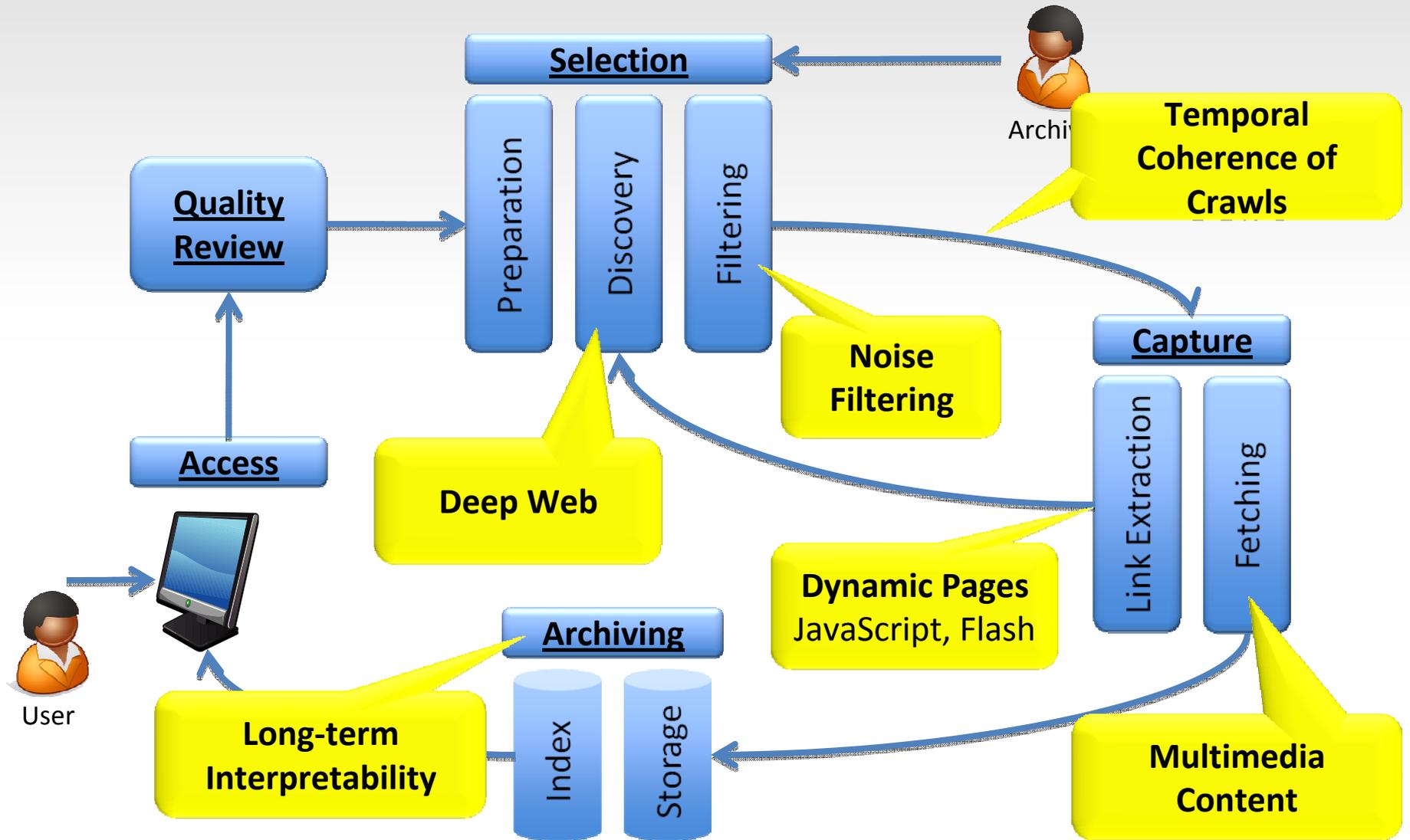
- High Quality Web Archives
- Long-term Archive usability

➔ From Web page storage to “Living Web Archives”

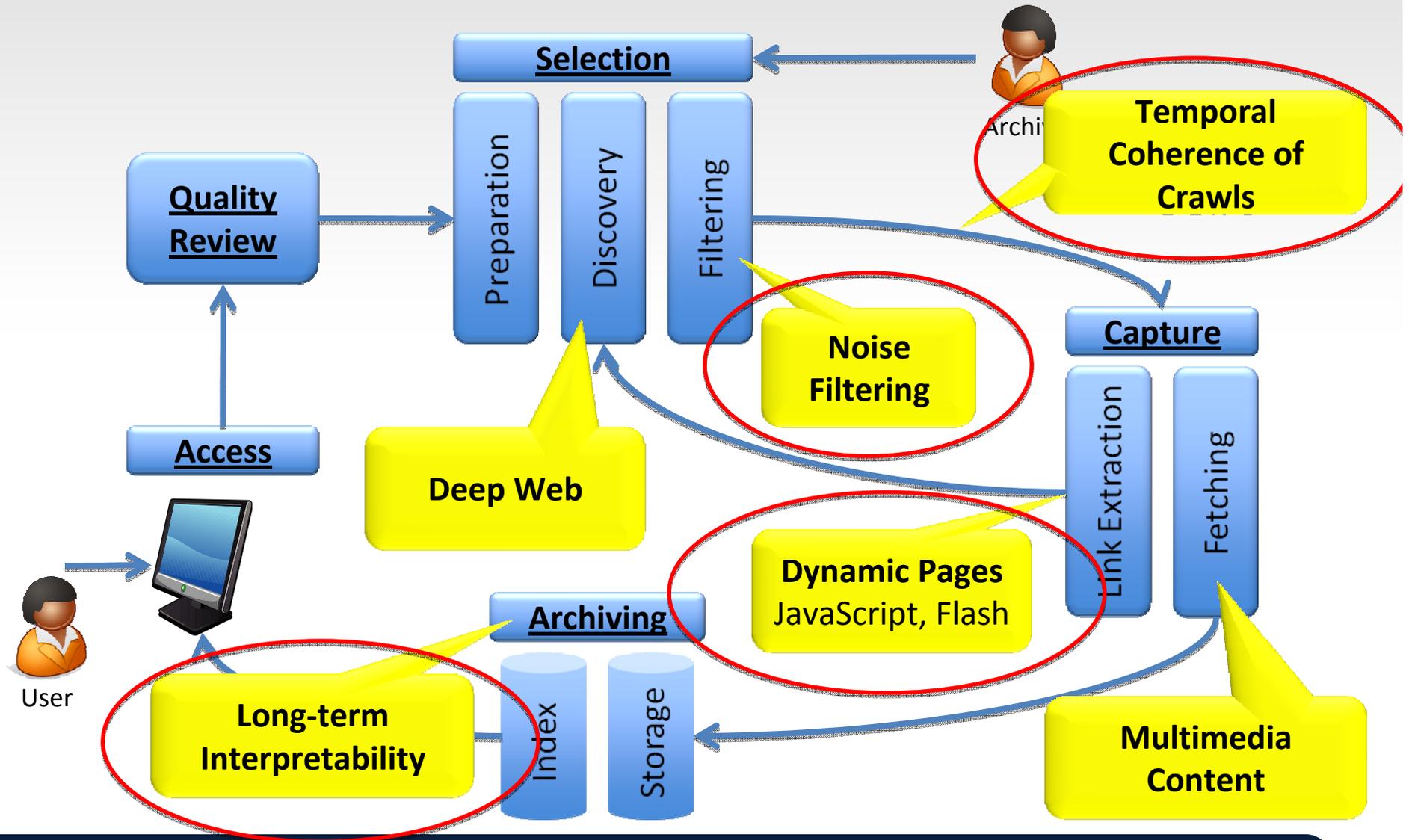
Started Feb. 2008 (3 Years)



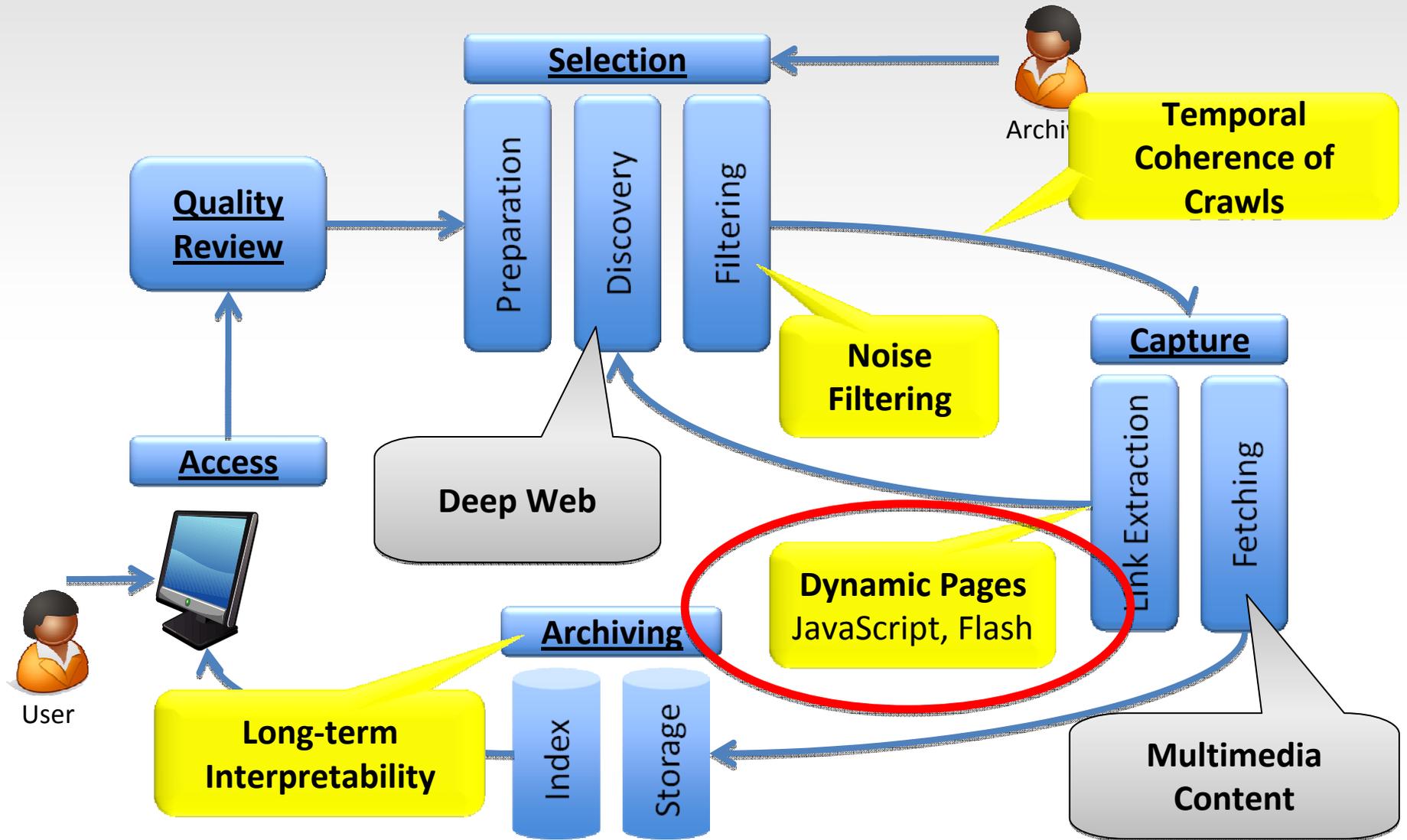
# Some LiWA Objectives in more Detail



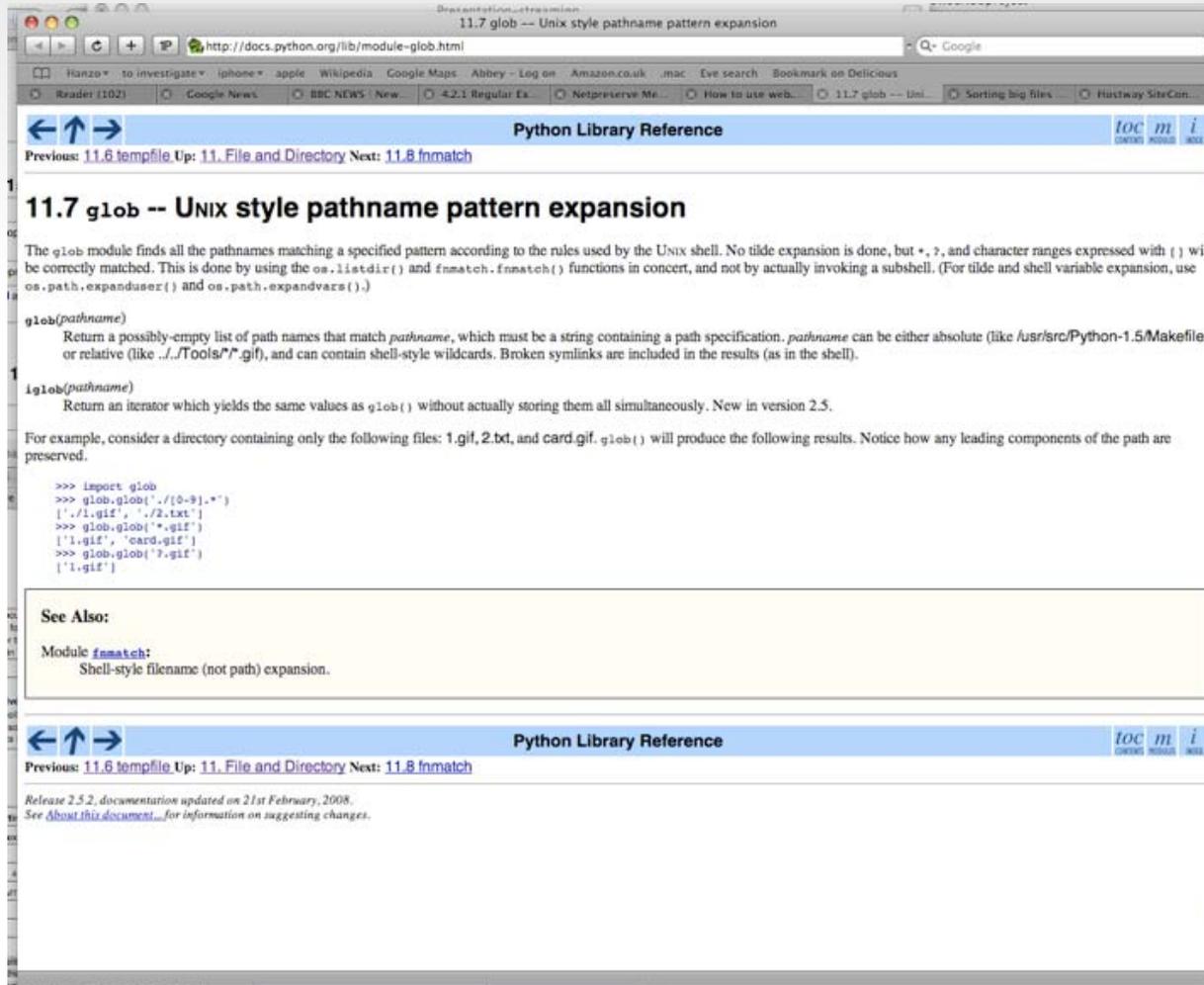
# Some LiWA Objectives in more Detail



# Link Extraction of Dynamic Pages



# Follow the links....

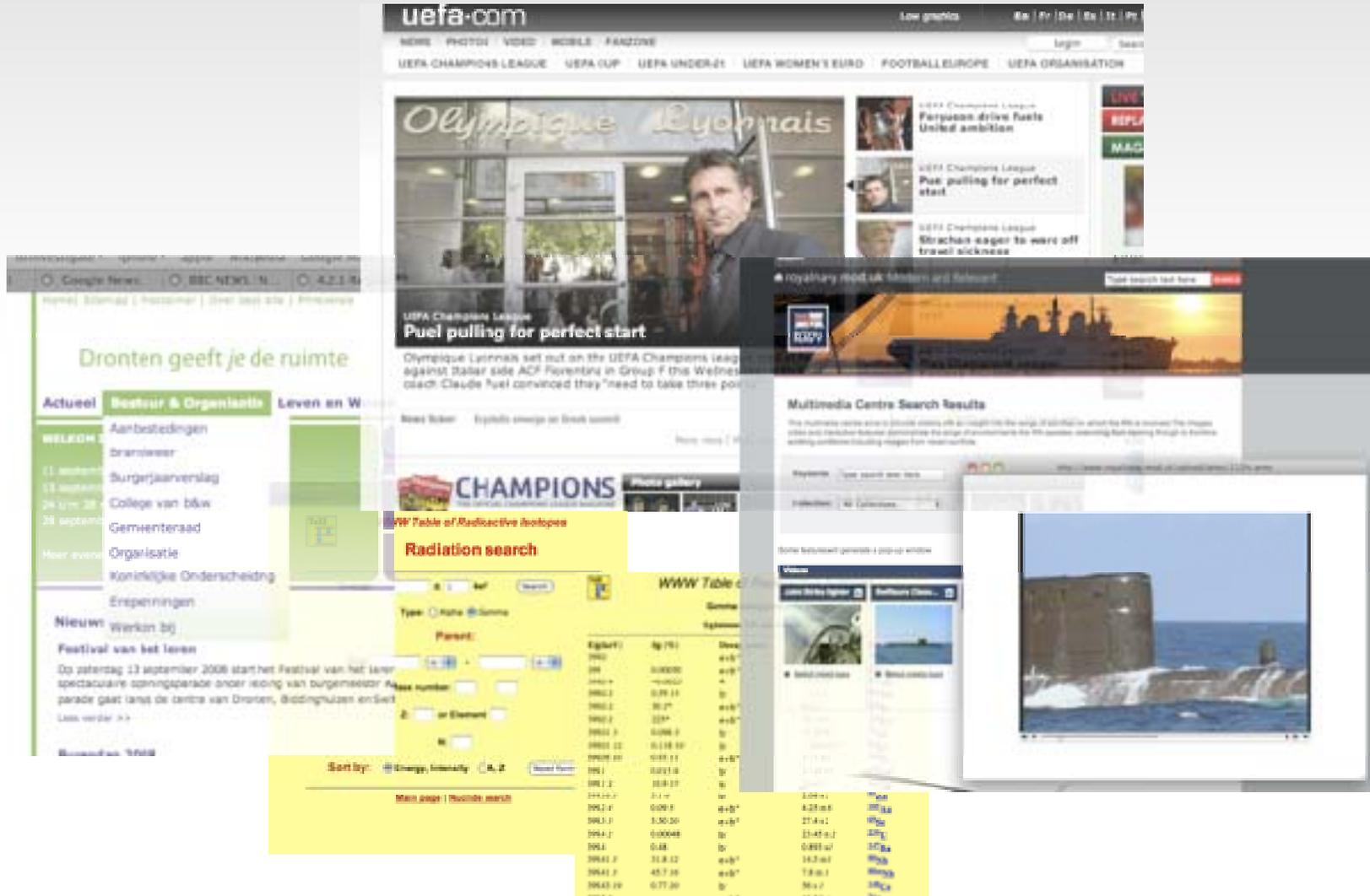


The screenshot shows a web browser window displaying the Python Library Reference page for the `glob` module. The browser's address bar shows the URL `http://docs.python.org/lib/module-glob.html`. The page title is "11.7 glob -- UNIX style pathname pattern expansion". The page content includes a description of the `glob` module, its functions (`glob` and `iglob`), and a code example demonstrating its usage. The code example shows the following output:

```
>>> import glob
>>> glob.glob('*/[0-9].*')
['./1.gif', './2.txt']
>>> glob.glob('*.*gif')
['1.gif', 'card.gif']
>>> glob.glob('?.gif')
['1.gif']
```

The page also includes a "See Also" section with a link to the `fnmatch` module, which is used for shell-style filename expansion. The page footer indicates it is Release 2.5.2, documentation updated on 21st February, 2008, and provides a link to `about:thisdocument` for information on suggesting changes.

# Hmm...

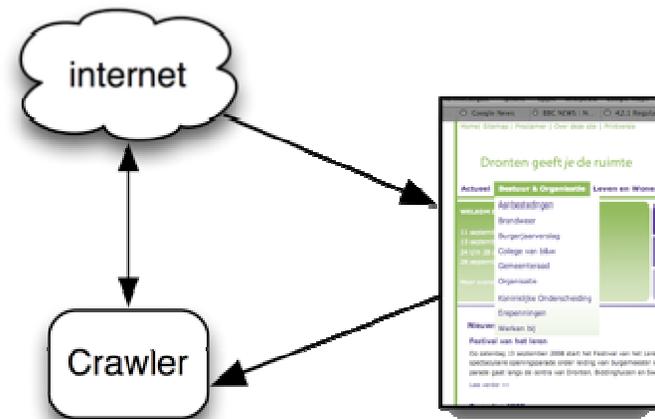


# Link Extraction of Dynamic Pages

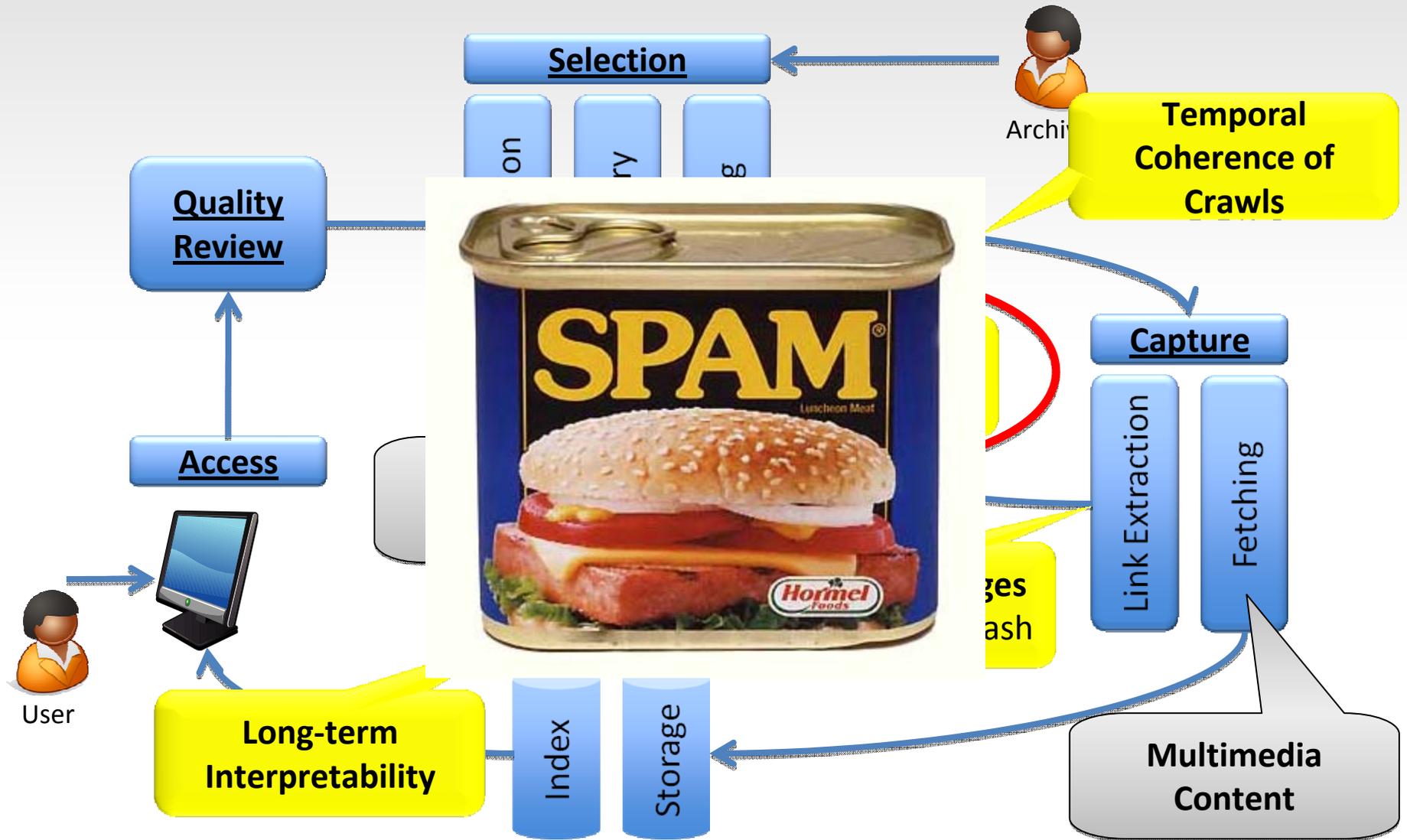
- The Problem:
  - links don't exist as raw text lying around
  - user interaction and code assemble them
- Current Approach:
  - “guessing” by assembling any fragments that look like links into URLs and trying them out
  - Can be very noisy - lots of wrong URL's

# Link Extraction of Dynamic Pages

- Approach
  - “pressing” the links and see what comes out
  - Execute code in a Javascript engine
  - Extract links from resulting DOM tree
  - Implementation based on WebKit



# Noise Filtering



# Web spam: for (or against) search engines

http://4485.1poap7.info/

The Mozilla Organiza... Latest Builds

## Compute the out degree

[On the Feasibility of Low-rank Approximation for Personalized PageRank](#)

File Format: PDF/Adobe Acrobat - View as HTMLtransition matrix of the Web graph for computing personal- ized PageRank. ...  
out-degree. Hence the base of links ...

[http://www.ilab.sztaki.hu/~stamas/publications/benczur05low\\_rank\\_ppr.pdf](http://www.ilab.sztaki.hu/~stamas/publications/benczur05low_rank_ppr.pdf) [Cached](#) - [Similar pages](#)

---

[schools for pharmacy phh mortgage cendant songs ring tones community credit union houston philadelphia penn s](#)  
[settlement hawaii insurance commissioner debt coverage ratios auto loan refinance classic video games online wha](#)  
[health insurance long beach schools financial credit union insurance umbrella policy disaster unemployment insurar](#)  
[mag mutual insurance company debit & credit chevron gas credit card money affiliate car loan application paradis](#)  
[casino photos progressive insurance claims office halloween bingo sheet binion world poker open pharmacy mass](#)  
[services credit union mortgage rates outlook cover insurance arts administration degree credit counseling governm](#)  
[lose weight casino star odds against 7 even party poker ipo](#)

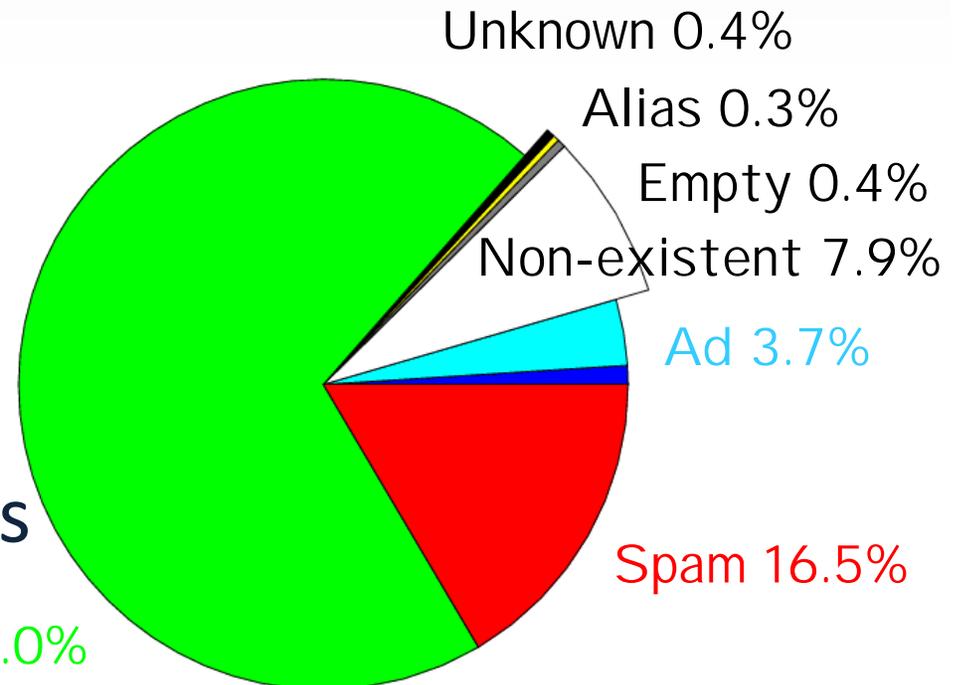
---

Compute the out d4egree compute tne out degree compute the out degree compute the Out degree compute the  
the out degree compute thye out degree compute the out degrwee comppute the out degree comute the out degre  
dree comopute the out degree compu5te the out degree compute t5he out degree ocmpute the out degree comp  
compute the oujt degree compute the outt degree vompute the out degree compute hte out degree compute the o  
ourt degree compute the out debree compute the out dergee compute the out degree compute the out degree co

# Web Spam: indexing vs. archiving

- Primary target: search engines to manipulate ranking
- As side effect, we also archive spam
- But very costly if not fought against:
  - traps crawler
  - 10+% sites
  - near 20% HTML pages

Reputable 70.0%



2004 .de crawl courtesy: T. Suel

# What can we do?

Ideal solution

- Automatic identification of spam pages

Requires

- Right selection of features to identify spam
- Development of new features e.g. creation and disappearance of new sites, pages
- Good training sets

Problem: Spam is constantly changing

→ Features need to be adapted

→ Updated training sets are necessary

Training set need to be prepared manually

# Spam Assessment Interface

pages from site  
www.example.com

- 001111110000 1.html
- 001111111100 products/software.html
- 001111111100 index.html

LiWA - Living Web Archives

Home | Abstract | Events | Partners | Contact

**- Abstract -**

Web content plays an increasingly important role in the knowledge-based society, and the preservation and long-term accessibility of Web history has high value (e.g., for scholarly studies, market analyses, intellectual property disputes, etc.). There is strongly growing interest in its preservation by library and archival organizations as well as emerging industrial services. Web content characteristics (high dynamics, volatility, contributor and format variety) make adequate Web archiving a challenge.

LiWA will look beyond the pure "freezing" of Web content snapshots for a long time, transforming pure snapshot storage into a "Living" Web Archive. "Living" refers to a) long term interpretability as archives evolve, b) improved archive fidelity by filtering out irrelevant noise and c) considering a wide variety of content.

LiWA will extend the current state of the art and develop the next generation of Web content capture, preservation, analysis, and enrichment services: to improve fidelity, coherence, and interpretability of web archives. By developing methods which improve archive fidelity, the project will contribute to adequate preservation of complete and high-quality content. By developing methods for improved archive coherence and interpretability, the project contributes to ensuring its long-term usability.

LiWA RTD will focus on innovative methods for content capturing, filtering out spam and other noise, improving temporal archive coherence, and dealing with semantic and terminology evolution. Two exemplary LiWA applications - focusing on individual streams and social web content, respectively - will show the benefits of advanced Web archiving to interested stakeholders.

To ensure demand-driven RTD development and broad, sustained project impact, the LiWA consortium will closely work with the International Internet Preservation Consortium (IIPC) as well as important library and archiving organizations, two of which are members of LiWA.

Home | Abstract | Events | Partners | Contact

jun jul **aug** sep oct **nov** dec jan feb mar apr may **now**

normal

borderline

spam

don't know

NEXT

HELP

QUIT

**labels**

2: normal  
1: borderline  
0: spam  
0: don't know

WEbspam-UK2007 label : normal

**scores**

sonar: 0.3  
lda: 0.4

**attributes**

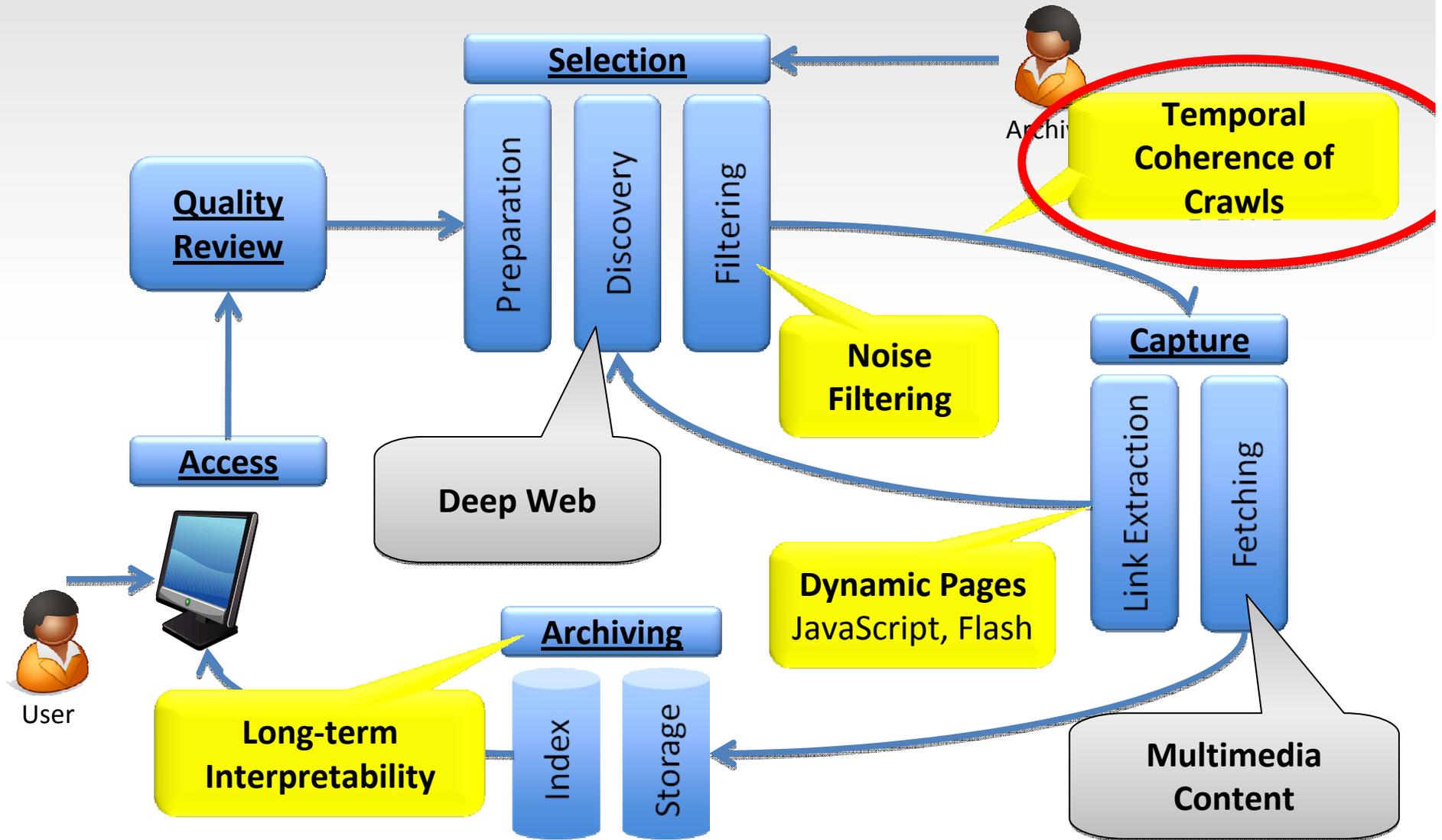
inlinks:  
outlinks:  
length of meta keywords:

**whois**

Univ. Hannover L3S

other sites of owner

# Temporal Coherence of Crawls

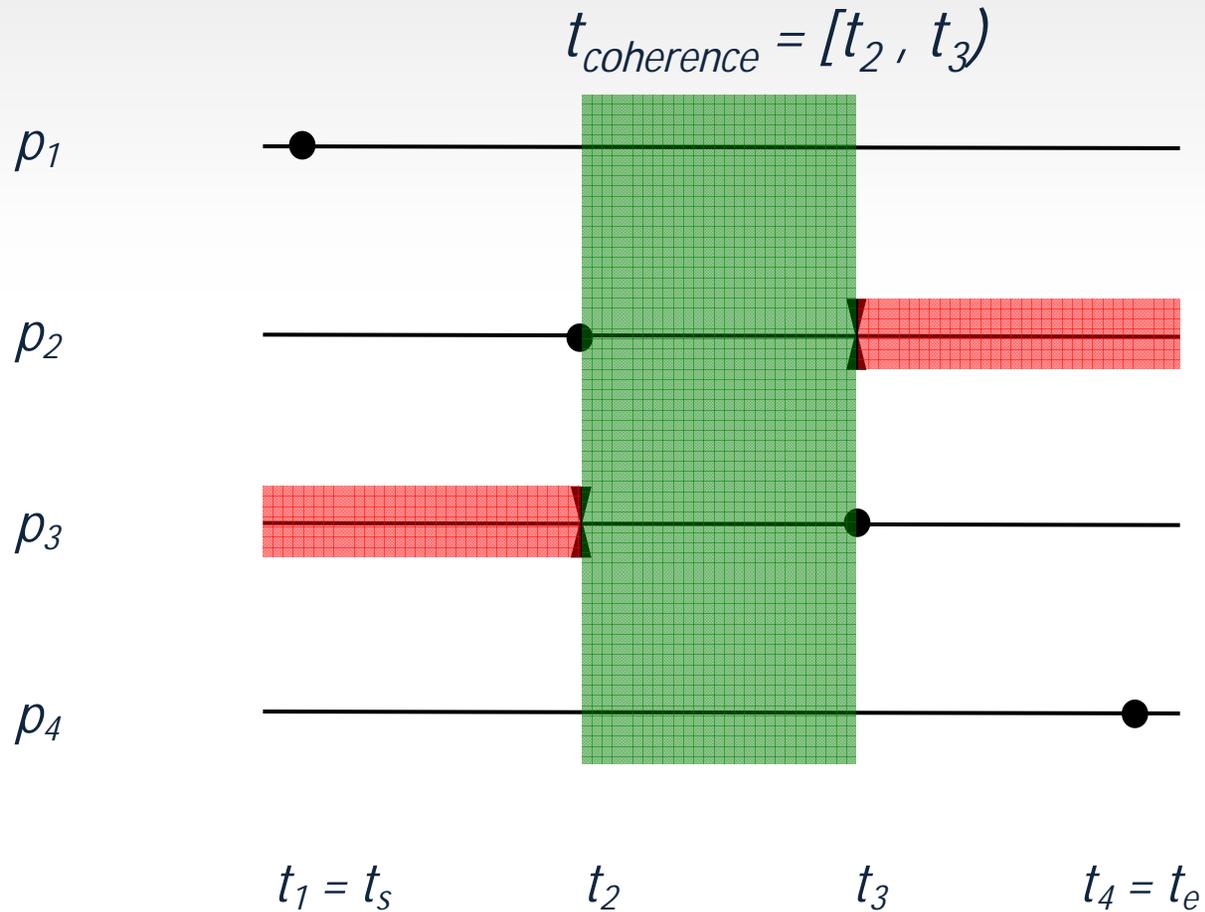


# Temporal Coherence

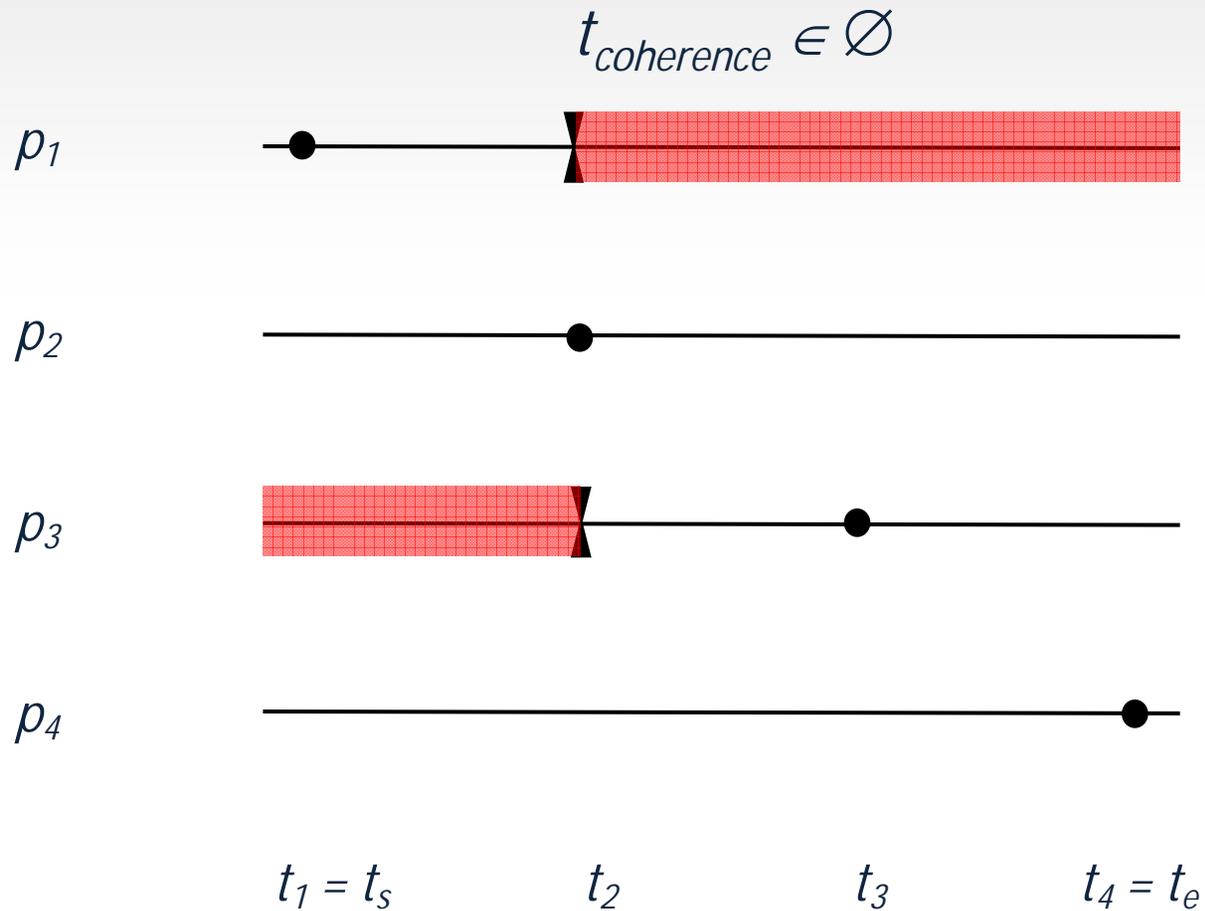
- Capturing Web sites as “authentic” as possible
- Make a site snapshot at once is not possible
- Crawlers need to be polite to web sites
  - Slow crawling, maybe with delays
  - Pages are changing during site crawl

→ When Do we have a coherent crawl?

# Coherence by Example



# Coherence by Example



# Coherence Analysis Technology



## Temporal Coherence Report

### Overview

Seed:	<a href="http://www.mpi-inf.mpg.de/">http://www.mpi-inf.mpg.de/</a>
Total count:	65046
Revisited pages:	65046
Crawl Duration:	2h 52m
Revisit Duration:	1h 19m

### Changes

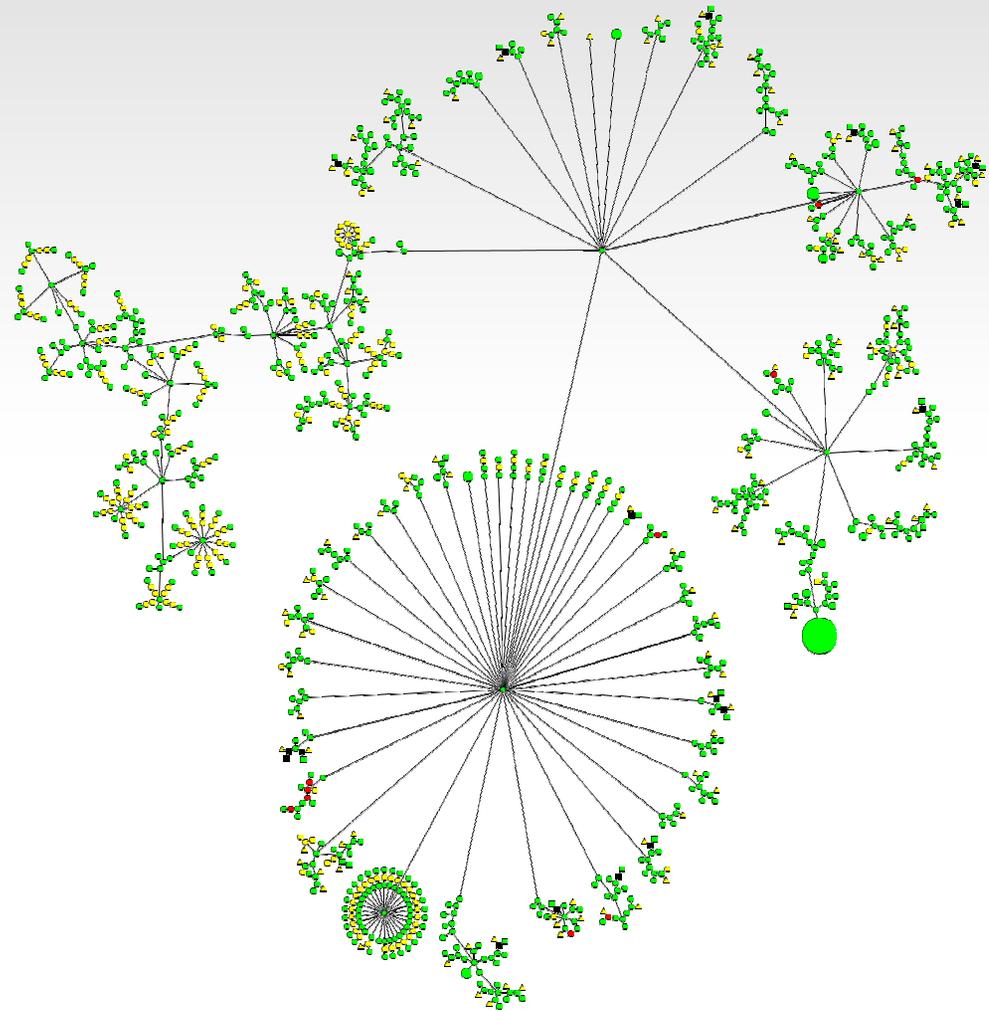
Pages with changed links:	16
Revisited pages:	65343
Revisit Duration:	1h 58m

### Details

#### Pages with changed links:

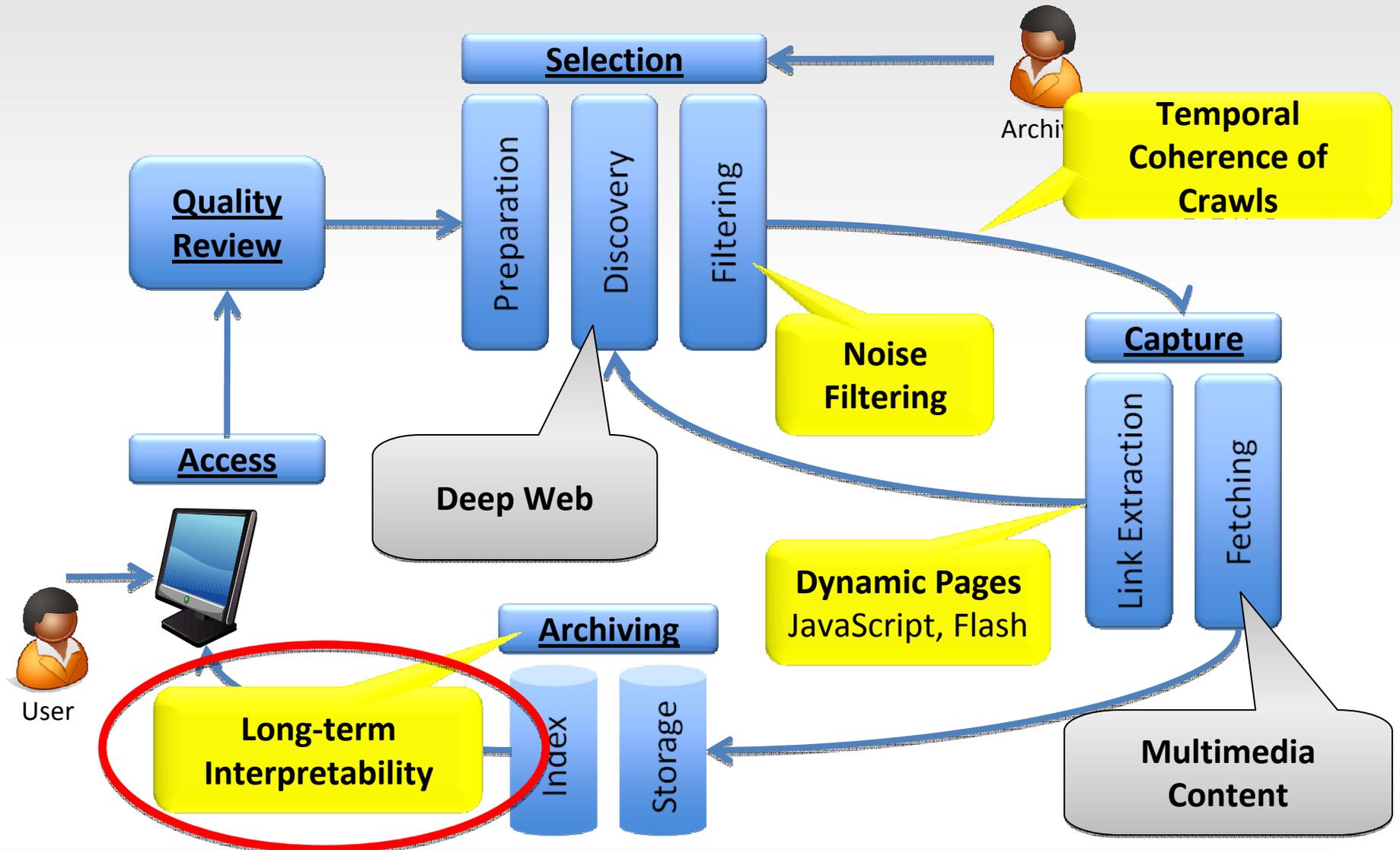
<a href="http://www.mpi-inf.mpg.de/departments/d4/teaching.html">http://www.mpi-inf.mpg.de/departments/d4/teaching.html</a>
<a href="http://www.mpi-inf.mpg.de/~ansips/">http://www.mpi-inf.mpg.de/~ansips/</a>
<a href="http://www.mpi-inf.mpg.de/~fried/index.html">http://www.mpi-inf.mpg.de/~fried/index.html</a>

Creation of automatically generated reports



Visualization of coherence defects

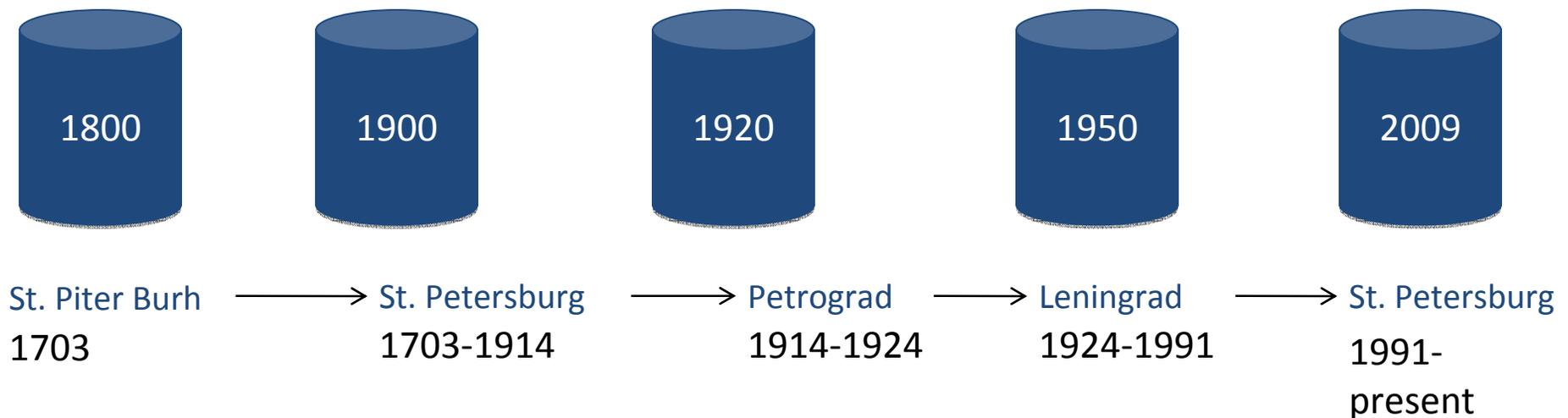
# Long-term Interpretability



# Motivation

Archives store content over long time ranges

- Content is created latest in the year of archiving
- Content typically creators use the language of that time



# Overview Process



**Step 1:** Word Sense Discrimination

**Step 2:** Tracking Evolution

# Data Sets for Evaluation (1/2)

## Data Set Requirements

- Large corpus
- Fully digitized
- Long time range – Increase probability of temporal domain evaluation
- Not too domain specific (like the Mesh corpus)
- Homogeneous language
- Time annotated

## Using Web Archives

- Large digital corpus
- At most 10 Years old
- Inhomogeneous with all the "noise" of the web
- Not suitable for initial evaluations



the national archives

UK GOVERNMENT WEB ARCHIVE

Search:

the URL's

Search >

Department for Constitutional Affairs [ live site <http://www.dca.gov.uk/pubs/archive.htm> ]

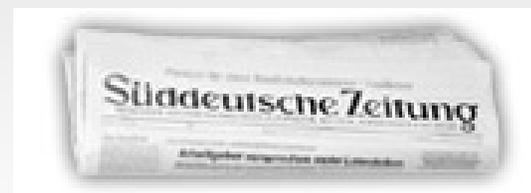
Search results for Jul 22 2004 - Sep 01 2008 [12 results]

2004 [ 1 instance]	2005 [ 1 instance]	2006 [ 3 instances]	2007 [ 5 instances]	2008 [ 2 instances]
<a href="#">Jul 22 2004</a>	<a href="#">Mar 01 2005</a>	<a href="#">Feb 13 2006</a> <a href="#">Oct 10 2006</a> <a href="#">Dec 09 2006</a>	<a href="#">Jan 01 2007</a> <a href="#">Feb 05 2007</a> <a href="#">Mar 05 2007</a> <a href="#">Apr 02 2007</a> <a href="#">May 06 2007</a>	<a href="#">Jun 09 2008</a> <a href="#">Sep 01 2008</a>

Powered by: european archive About | Contact | Terms, Privacy & Copyright

# Data Sets for Evaluation (2/2)

- Newspaper Archives
  - Fully digitized corpora
  - Controlled language
  - Clear time annotations
- Süddeutsche Zeitung (ger.)
  - Spans from year 1994 - 2006
  - ~ 1.3 Million articles
- London Times Archive (engl.)
  - Spans from year 1785-1985
  - ~ 20 Million articles



## Strategy

- Year 2: Initial evaluations on well known corpora
- Year 3: Apply technology to web archives, .gov.uk crawls provided by EA

A screenshot of the Times Online website. The top navigation bar includes 'NEWS', 'COMMENT', 'BUSINESS', 'SPORT', 'LIFE &amp; STYLE', 'ARTS &amp; ENTERTAINMENT', 'LUXE', 'ARCHIVE', 'OUR PAPERS', and 'JOBS &amp; CLASSIFIEDS'. The main content area features a search bar for the 'TIMESARCHIVE' with a date range selector (FROM: 1 Jan 1785 TO: 31 Dec 1985) and a 'SEARCH' button. Below the search bar is a 'WELCOME' section with a 'Welcome' heading and a paragraph about exploring 200 years of history. There are also 'FEATURED SEARCHES' and 'REVIEW OUR TOPICS' sections. The bottom right corner shows a 'Capgemini' advertisement with the text 'Together. Free your energies.'

# Term. Evol - Conclusions and Future Work

## Terminology Extraction

Find methods that are  
time independent for

- Extraction
- Stop word removal
- Lemmatization
- Correcting OCR errors

## Word Sense Discrimination

- Other types of clustering
- Metrics for evaluation

## Detecting Evolution

- Methods for comparing clusters and detecting evolution
- Methods for evaluation

# Conclusions and Expected Project Results

## Improving Web Archiving Technology

- Rich Media Capturing
- Spam Processing
- Archive Coherence
- More general scope: Improving Archive Interpretability

## Selected results will be integrated in

- Heritrix Crawler (our test-bed)
- Hanzo Archives Crawler

## Evaluation in two test cases:

- Streaming Media WebArchive by Sound & Vision
- „WebArchivists Workbench“ by European Archive and Nat. Lib. Czech Republic



**LiWA**  
Living Web Archives

**Thank you!**



**More information on  
<http://www.liwa-project.eu/>**

# The LiWA Consortium

## Archiving Companies

European Archive  
Hanzo Archives



**Archiving Users**  
Stichting Nederlands Instituut  
voor Beeld en Geluid,  
National Library of  
the Czech Republic,  
Moravian Library



## Technical & Scientific

Leibniz University  
L3S Research Center,  
Max-Planck-Institut für Informatik,  
Computer and Automation Research Institute  
Hungarian Academy of Sciences

