LiWA
Living Web Archives

From Web Page Storage to Living Web Archives

Thomas Risse

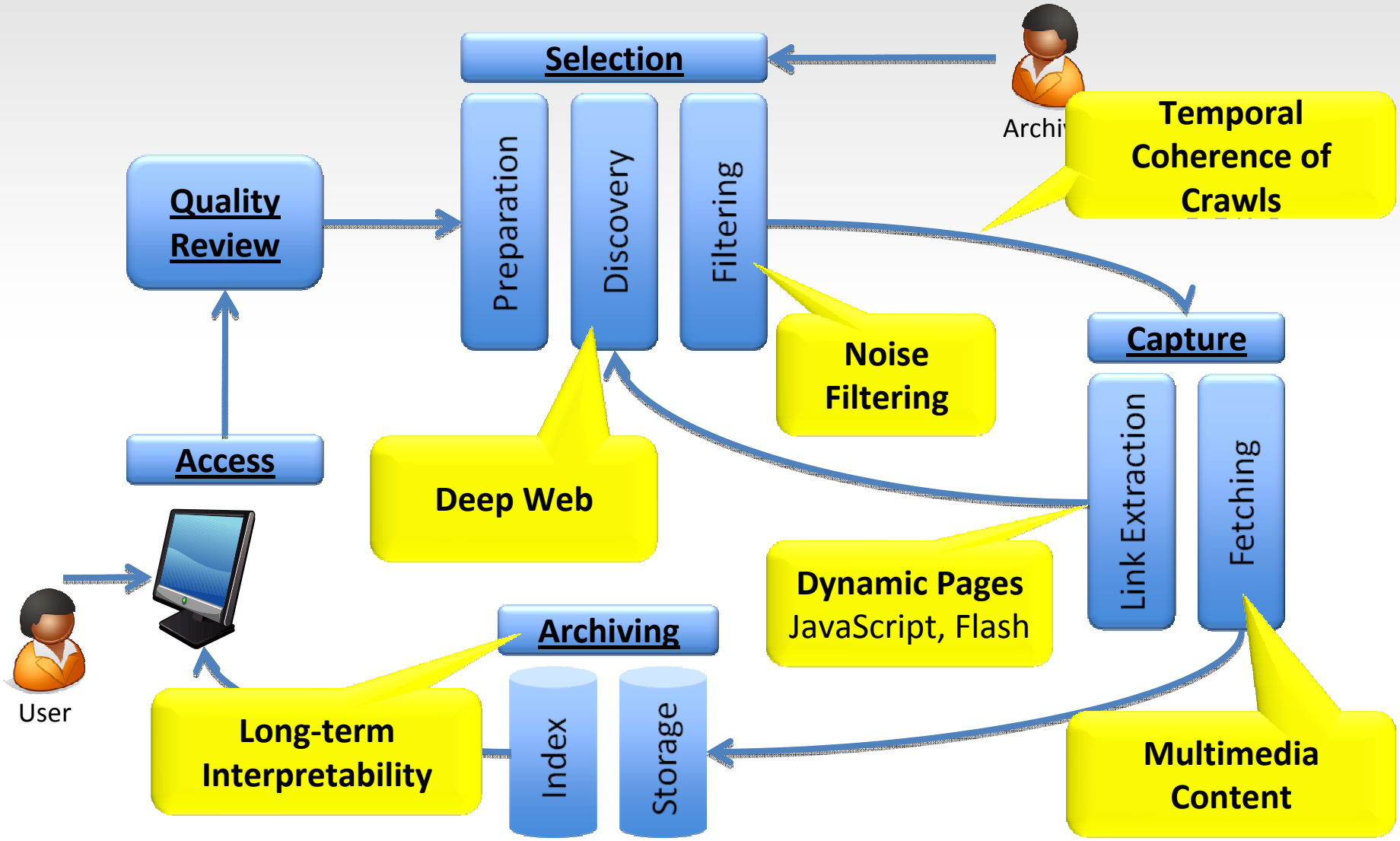
JISC, the DPC and the UK Web Archiving Consortium Workshop

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



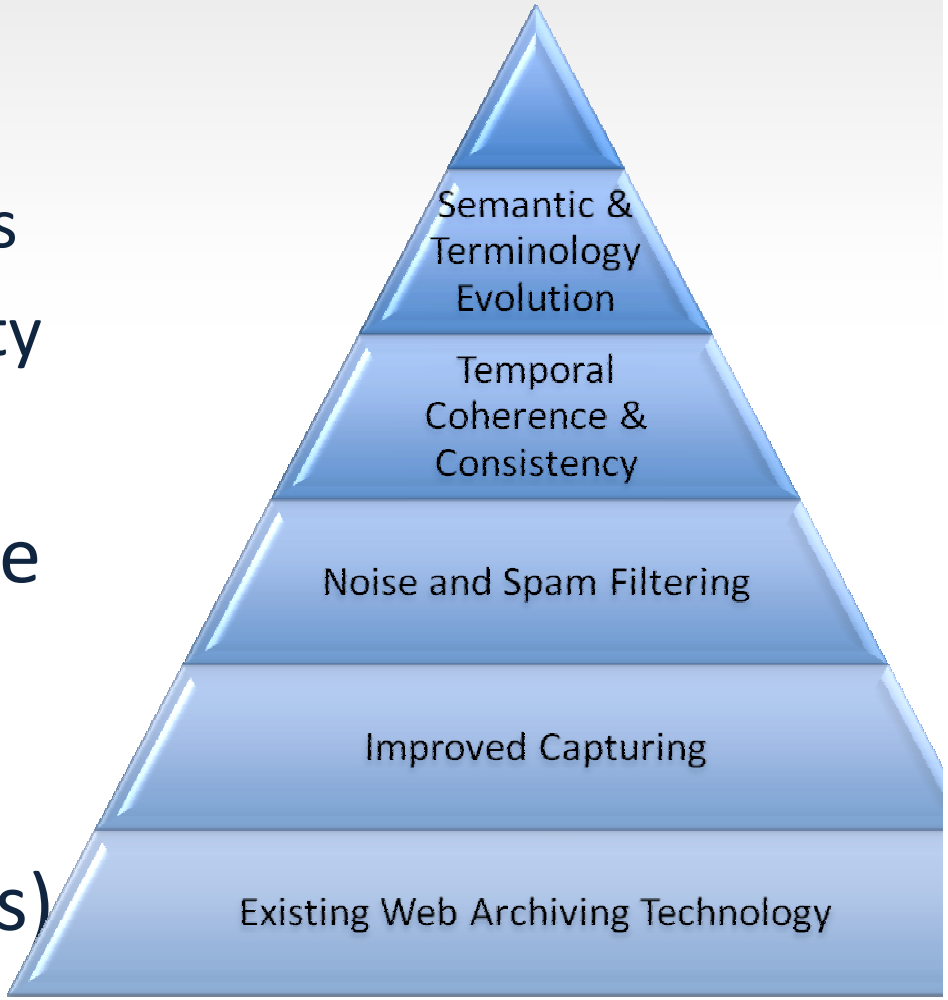
LiWA – Living Web Archives (EU-IST 216267)

Next generation Web Archiving technology for:

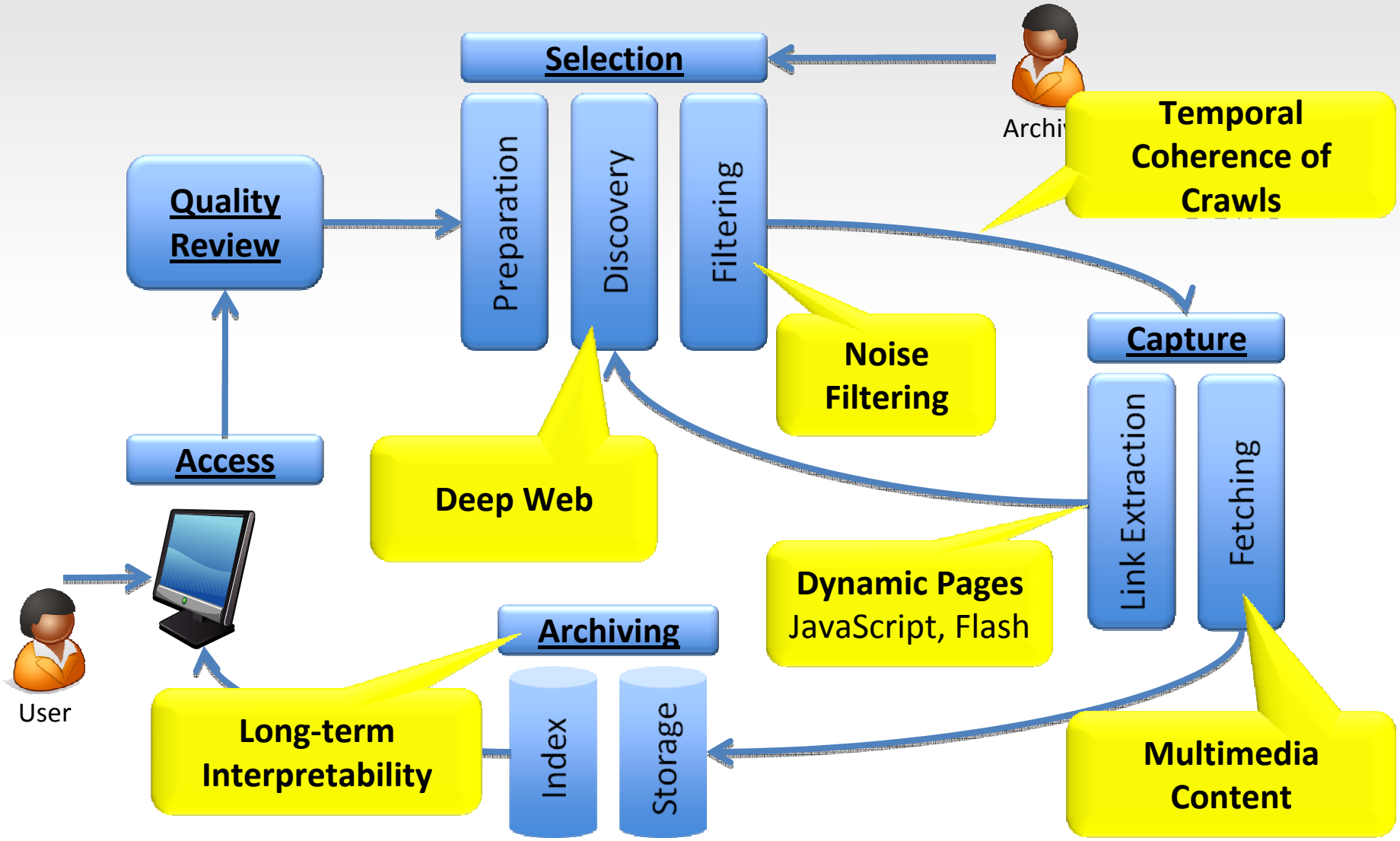
- High Quality Web Archives
- Long-term Archive usability

➔From Web page storage to “Living Web Archives”

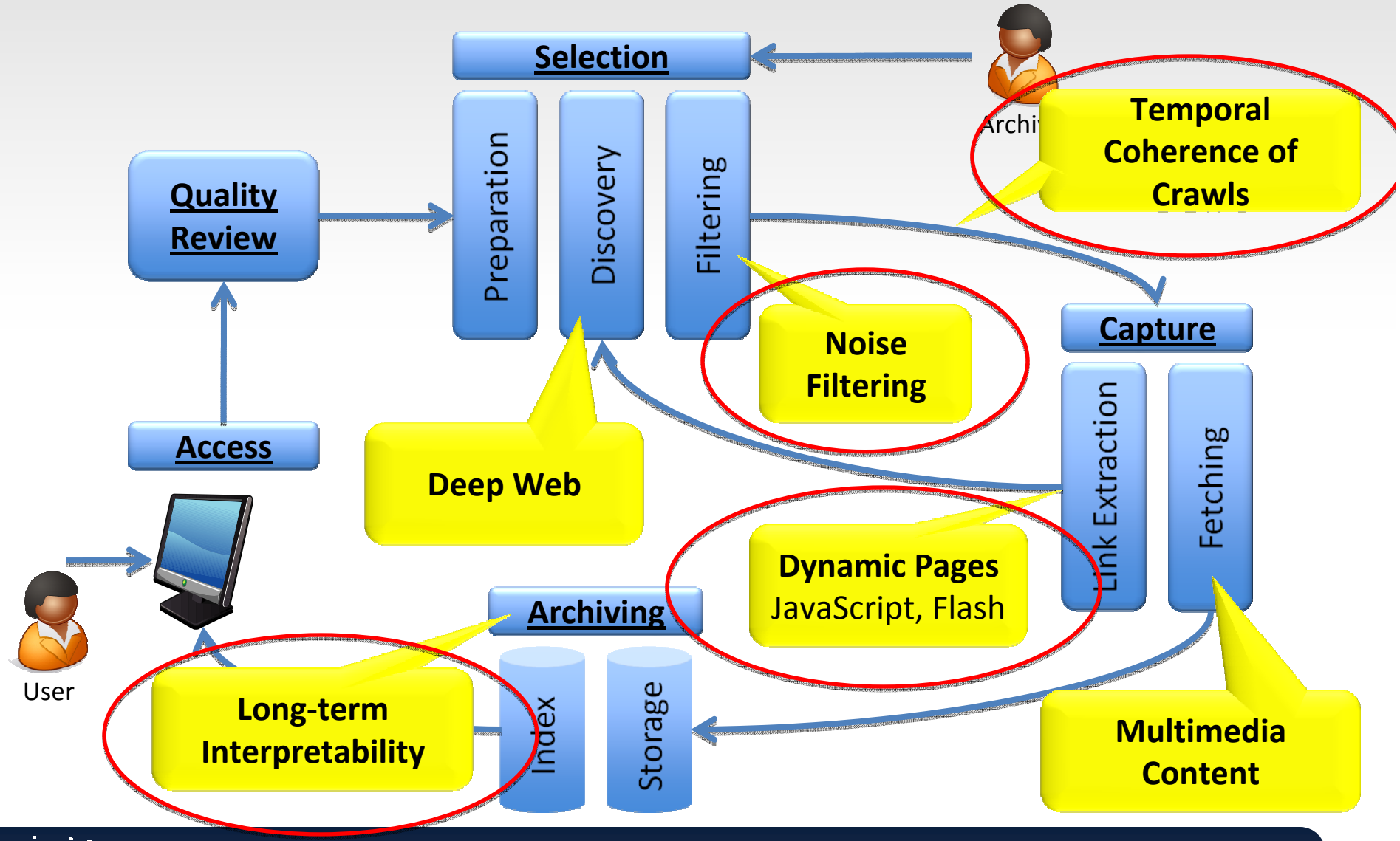
Started Feb. 2008 (3 Years)



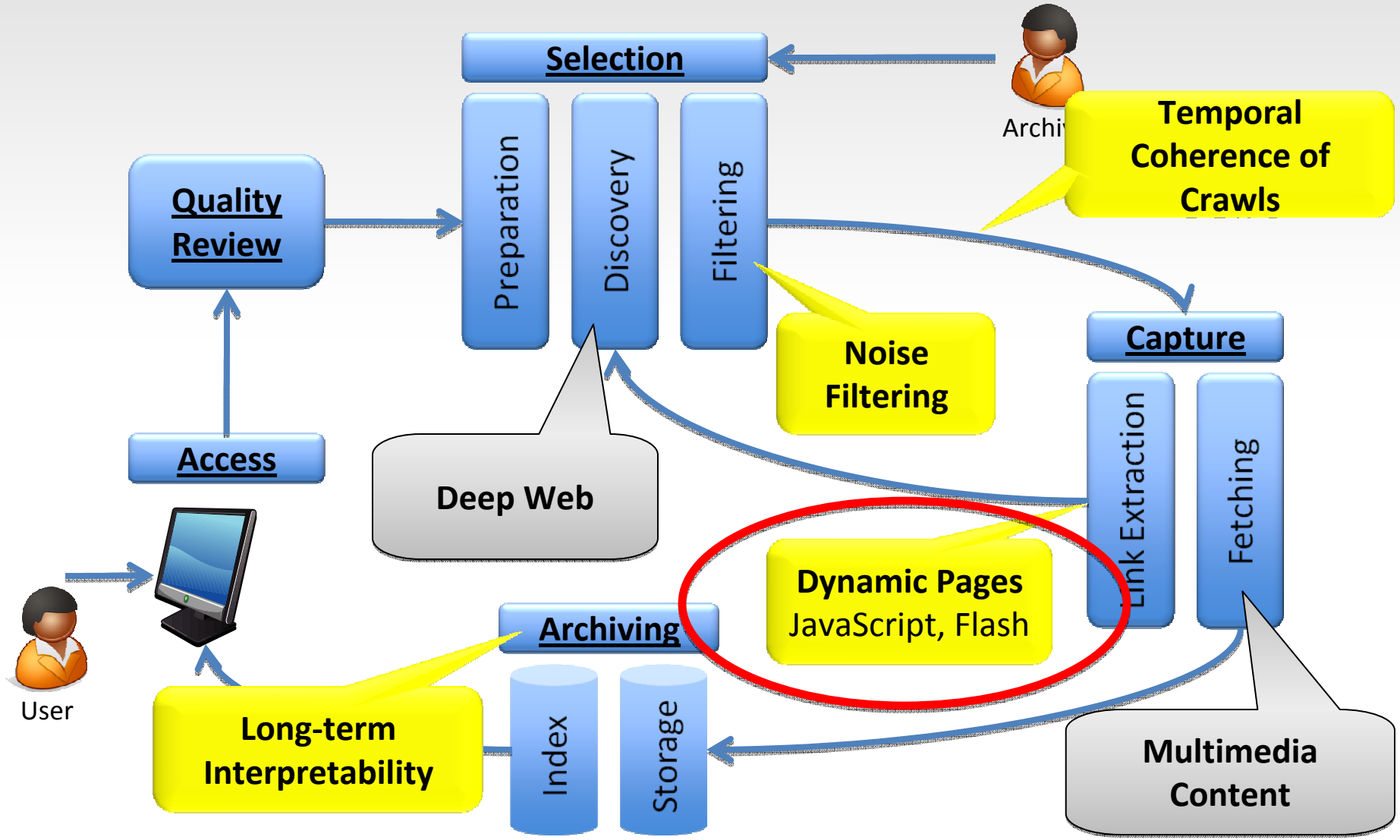
Some LiWA Objectives in more Detail



Some LiWA Objectives in more Detail

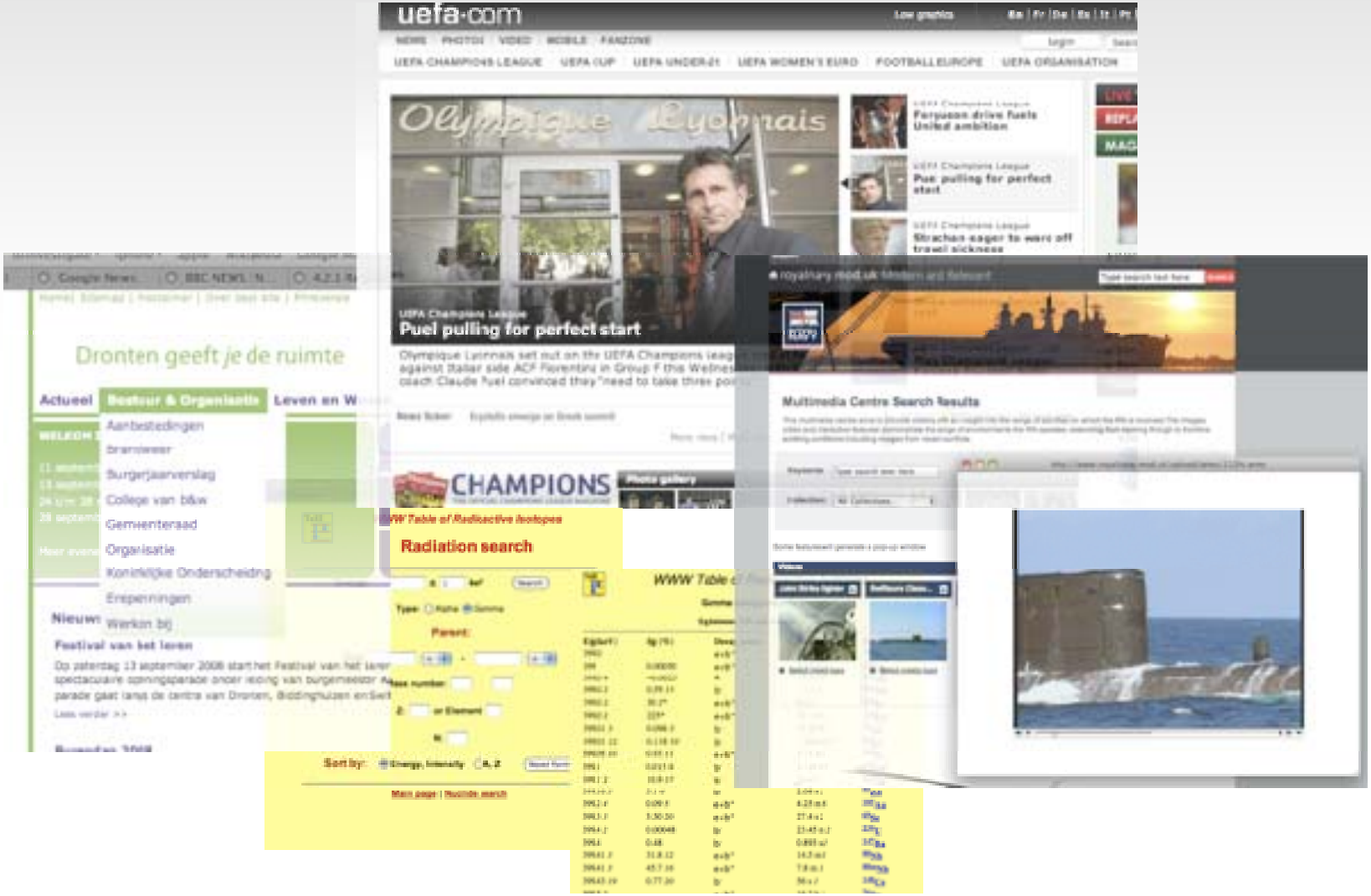


Link Extraction of Dynamic Pages



Follow the links....

Hmm…

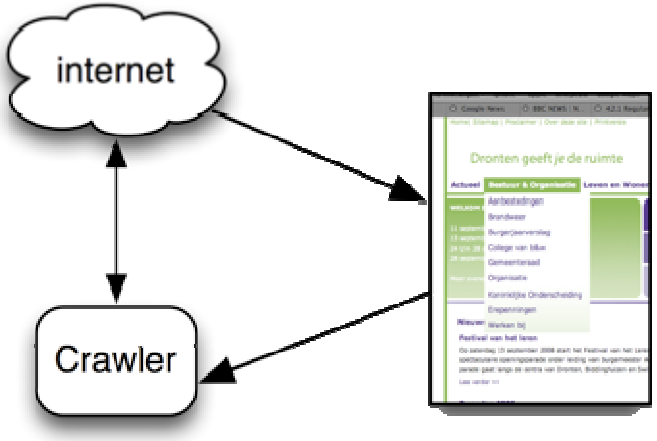

uefa.com
Olympique Lyonnais
UEFA Champions League
Puel pulling for perfect start
Ferguson drive fuels United ambition
Strachan eager to ward off travel sickness
Dronten geeft je de ruimte
Actueel
Bestuur & Organisatie
Aanbestedingen
Brandweer
Burgerjaarverslag
College van b&w
Gemeenteraad
Organisatie
Koninklijke Onderscheiding
Ereperningen
Werken bij
Nieuws
Festival van het leven
CHAMPIONS
Radiation search
Multimedia Centre Search Results

Link Extraction of Dynamic Pages

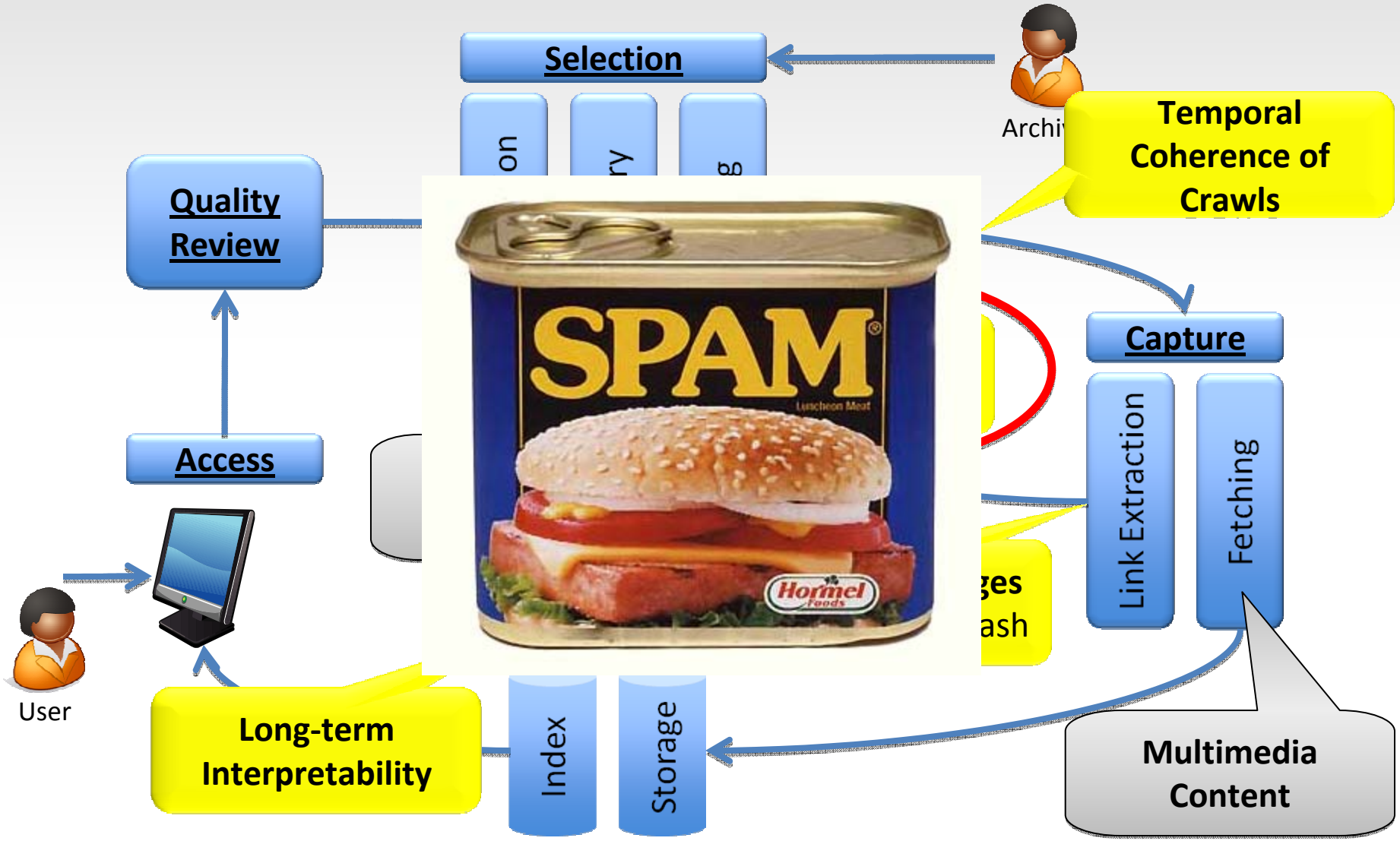
- The Problem:
  - links don't exist as raw text lying around
  - user interaction and code assemble them
- Current Approach:
  - "guessing" by assembling any fragments that look like links into URLs and trying them out
  - Can be very noisy - lots of wrong URL's

Link Extraction of Dynamic Pages

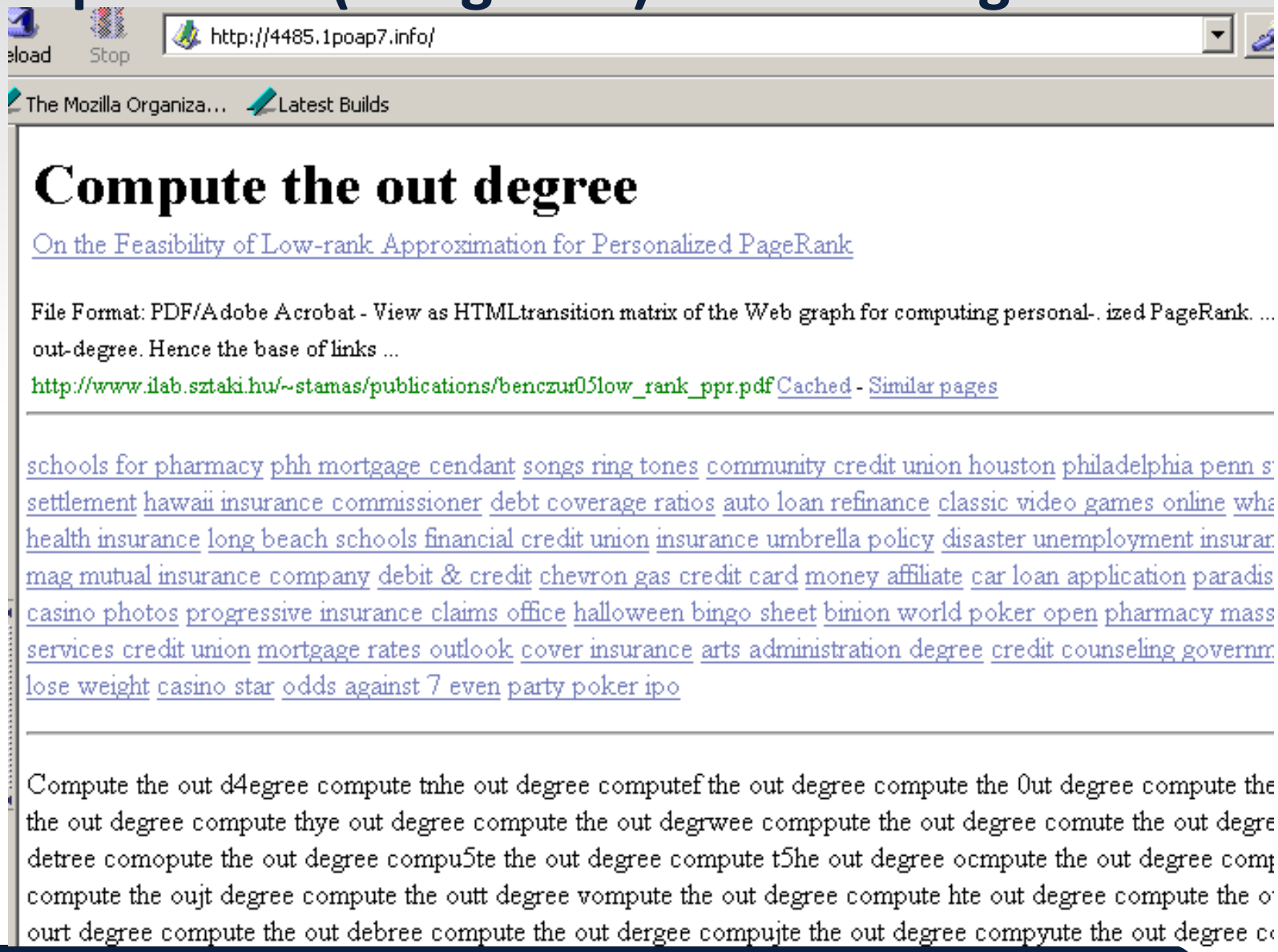
- Approach
  - "pressing" the links and see what comes out
  - Execute code in a Javascript engine
  - Extract links from resulting DOM tree
  - Implementation based on WebKit



Noise Filtering

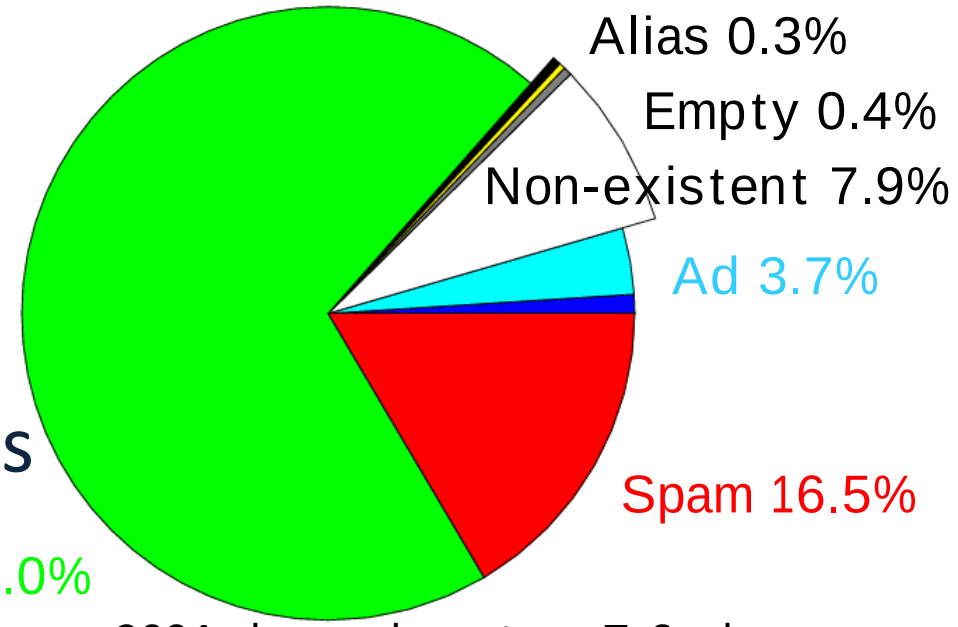


Web spam: for (or against) search engines



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



2004 .de crawl courtesy: T. Suel

What can we do?

Ideal solution

- Automatic identification of spam pages

Requires

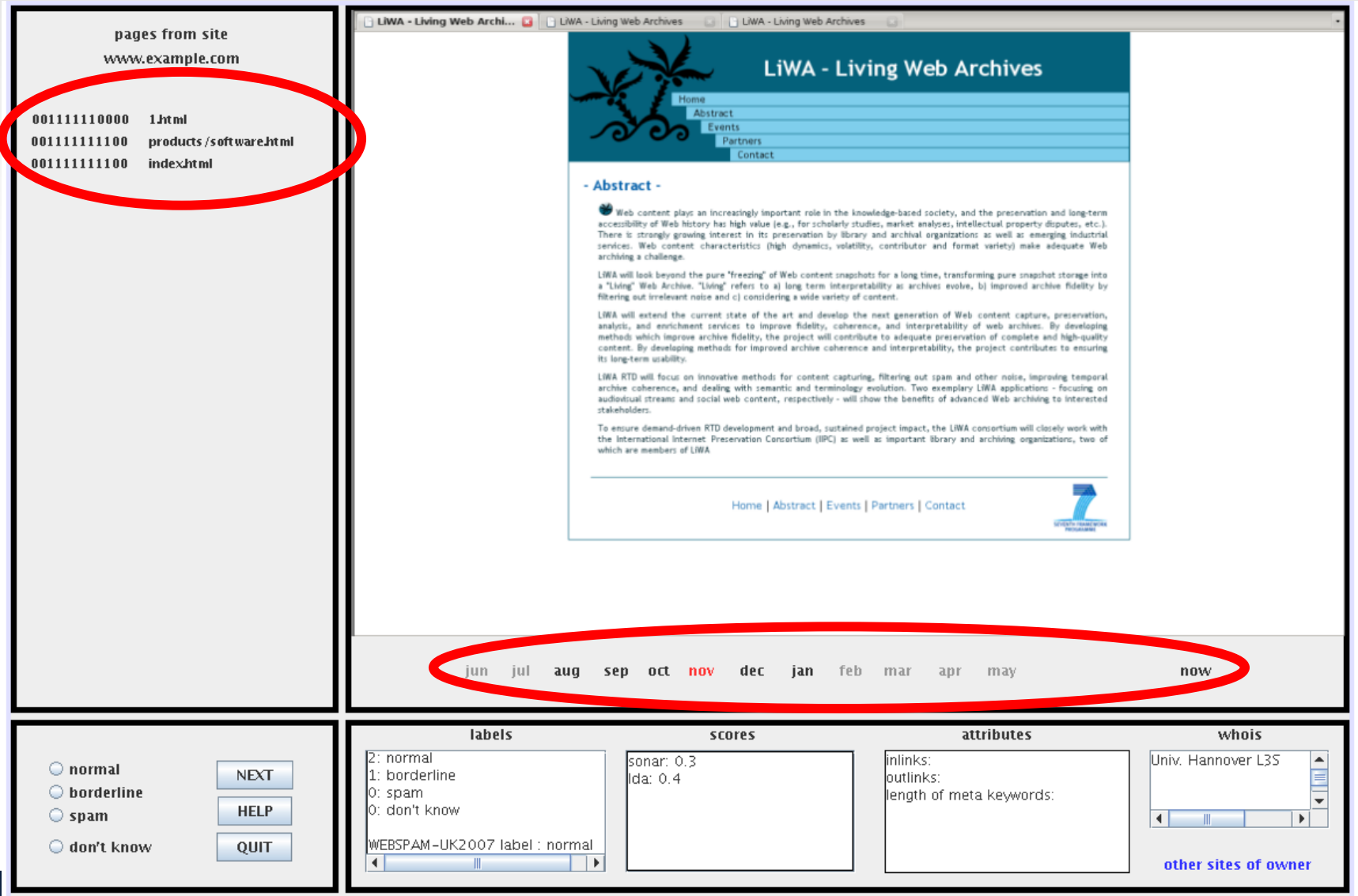
- Right selection of features to identify spam
- Development of new features e.g. creation and disappearance of new sites, pages
- Good training sets

Problem: Spam is constantly changing

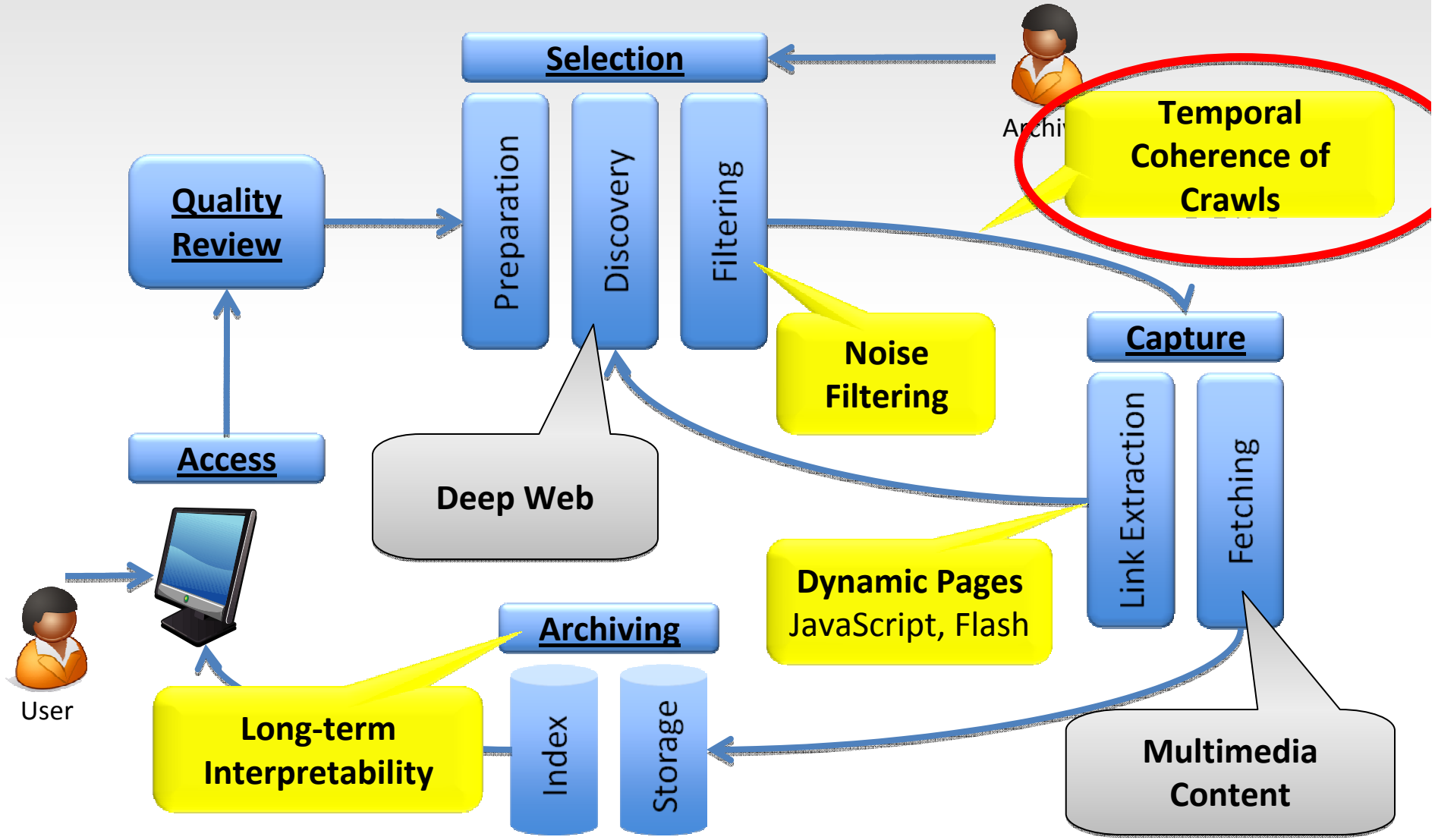
→ Features need to be adapted

→ Updated training sets are necessary
Training set need to be prepared manually

Spam Assessment Interface



Temporal Coherence of Crawls

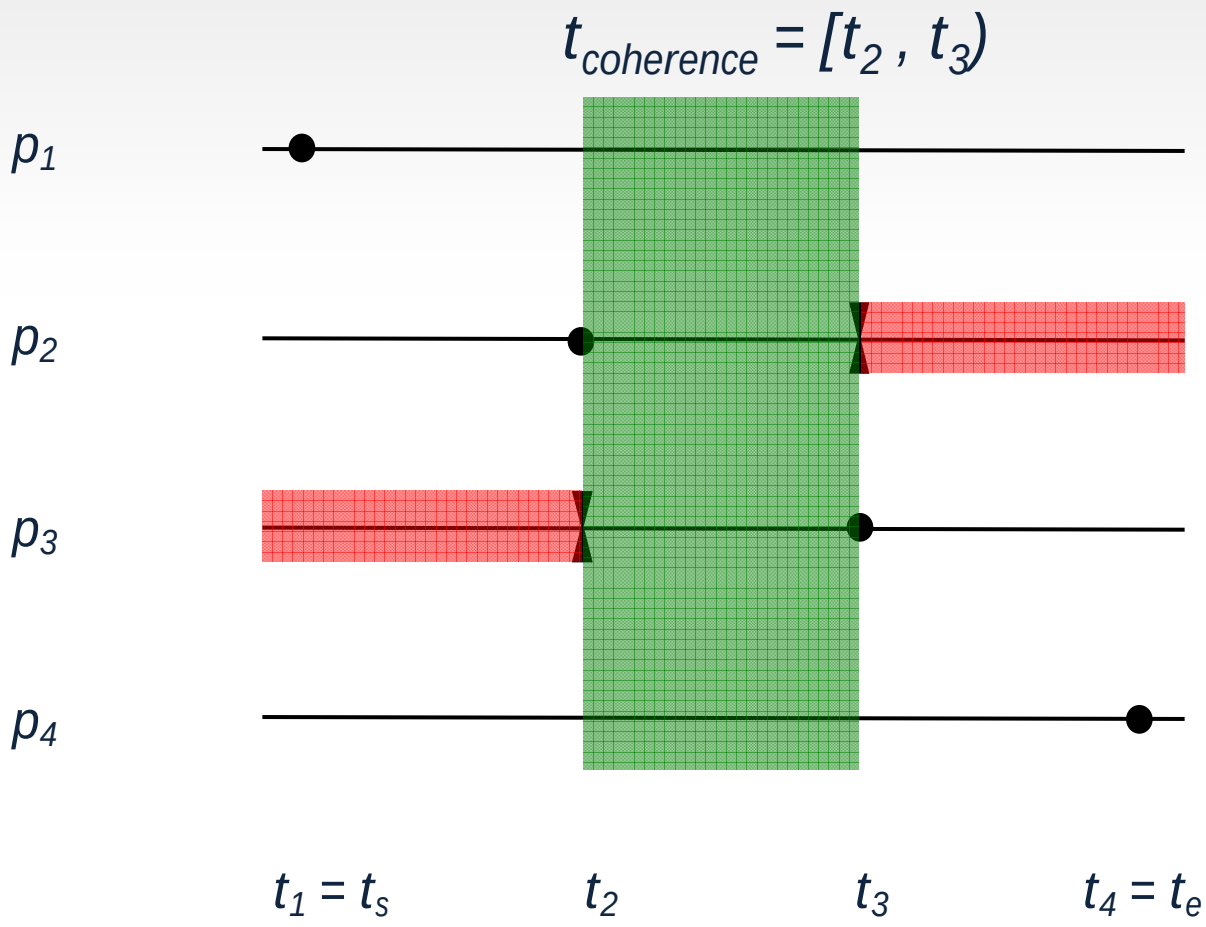


Temporal Coherence

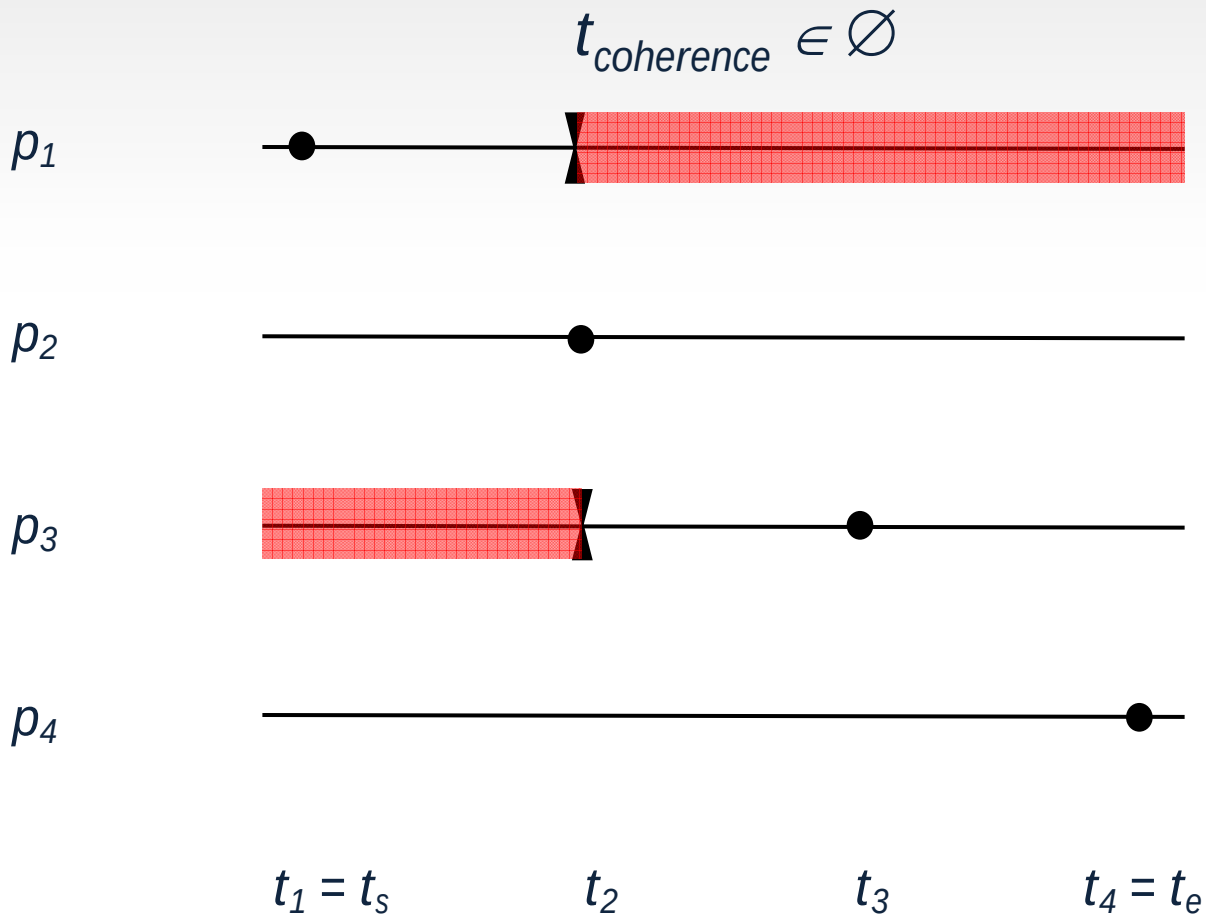
- Capturing Web sites as “authentic” as possible
- Make a site snapshot at once is not possible
- Crawlers need to be polite to web sites
  - Slow crawling, maybe with delays
  - Pages are changing during site crawl

→When Do we have a coherent crawl?

Coherence by Example



Coherence by Example



Coherence Analysis Technology

Temporal Coherence Report

Overview

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

Changes

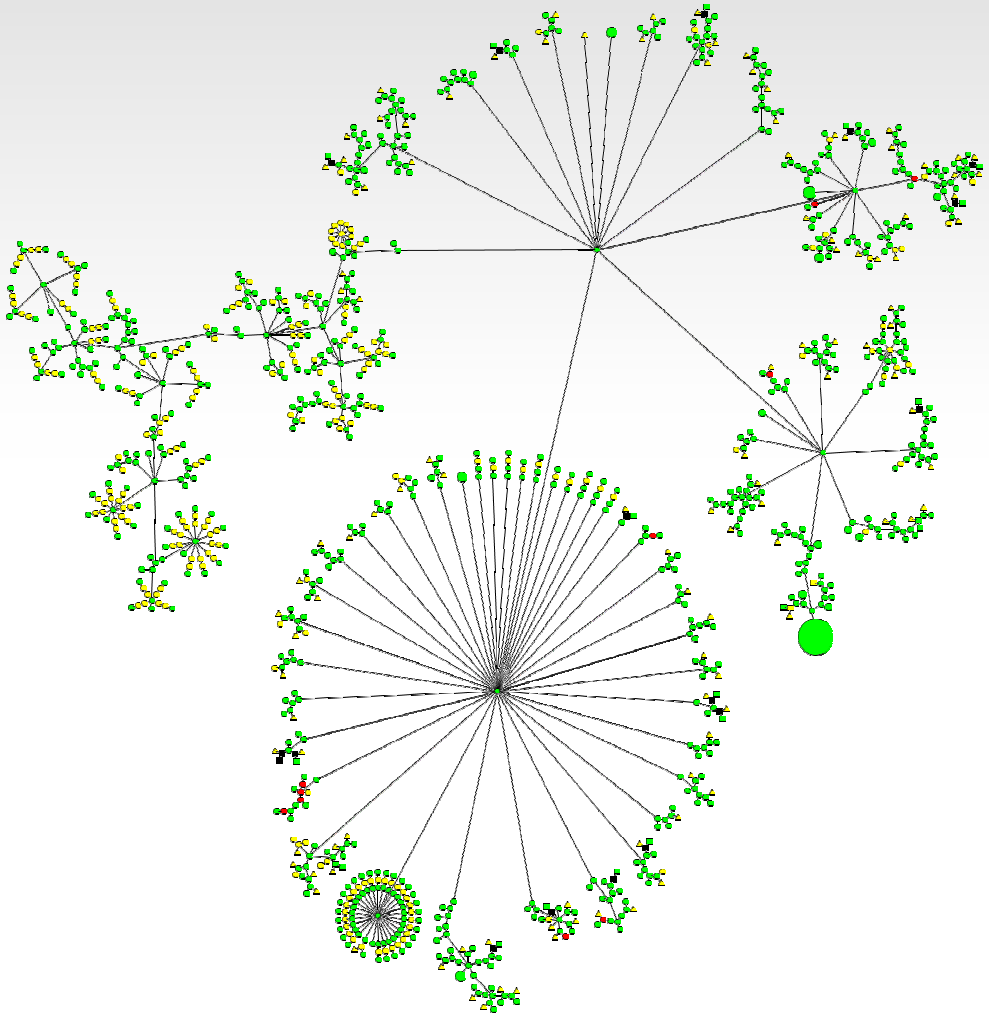
Pages with changed links:	16
Revisited pages:	65343
Revisit Duration:	1h 58m

Details

Pages with changed links:

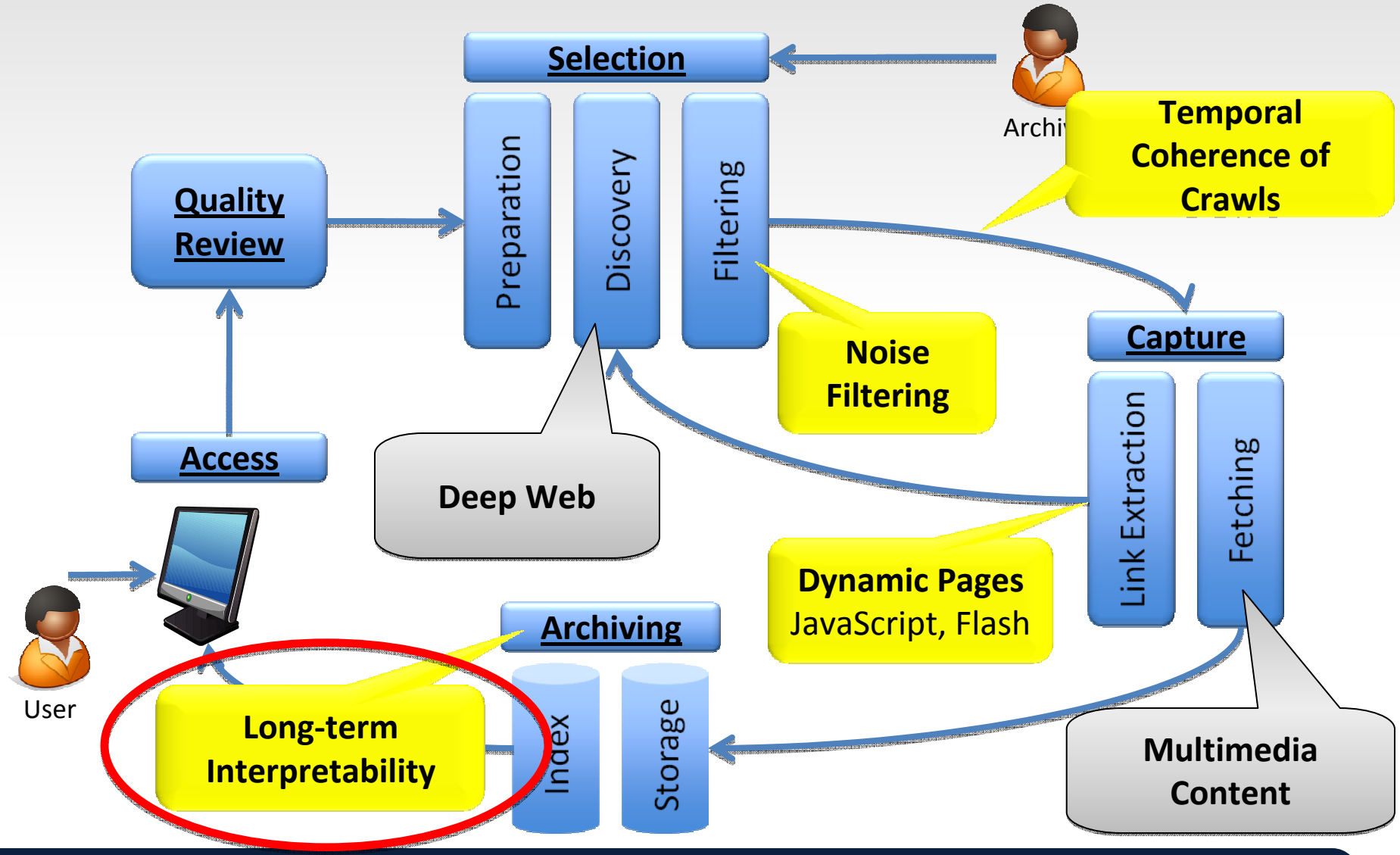
http://www.mpi-inf.mpg.de/departments/d4/teaching.html
http://www.mpi-inf.mpg.de/~ensips/
http://www.mpi-inf.mpg.de/~sfries/index.html

Creation of automatically generated reports



Visualization of coherence defects

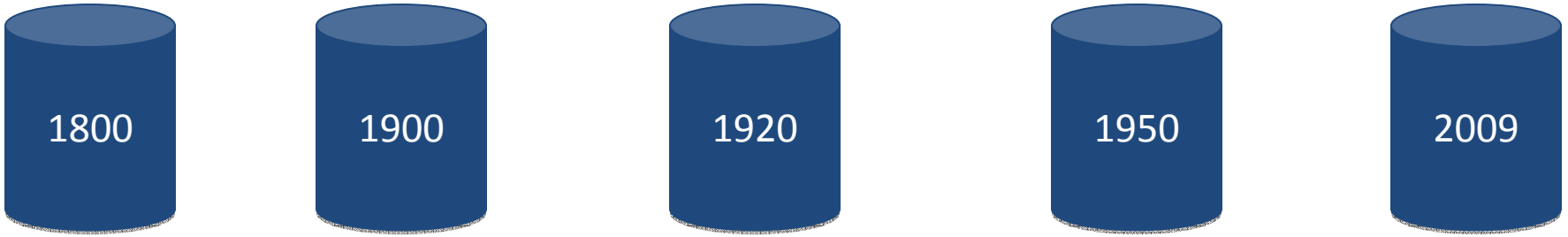
Long-term Interpretability



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



St. Piter Burh → St. Petersburg → Petrograd → Leningrad → St. Petersburg

1703 | 1703-1914 | 1914-1924 | 1924-1991 | 1991-present

Overview Process

Step 1: Word Sense Discrimination

Step 2: Tracking Evolution

Data Sets for Evaluation (1/2)

Data Set Requirements

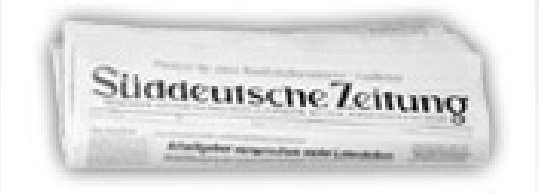
- Large corpus
- Fully digitized
- Long time range – Increase probability of terminology evolution
- Not too domain specific (like the Mesh c
- Homogeneous language
- Time annotated

Using Web Archives

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

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

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