



Thanks for your interest in the LiWA project! For this first newsletter, we have asked all the LiWA research partners to present their goals and summarize their achievements for the first year of the project. We welcome your feedback or questions to info@liwa-project.eu. To subscribe, and get informed, please visit <http://liwa-project.eu>

The LiWA Project in a nutshell

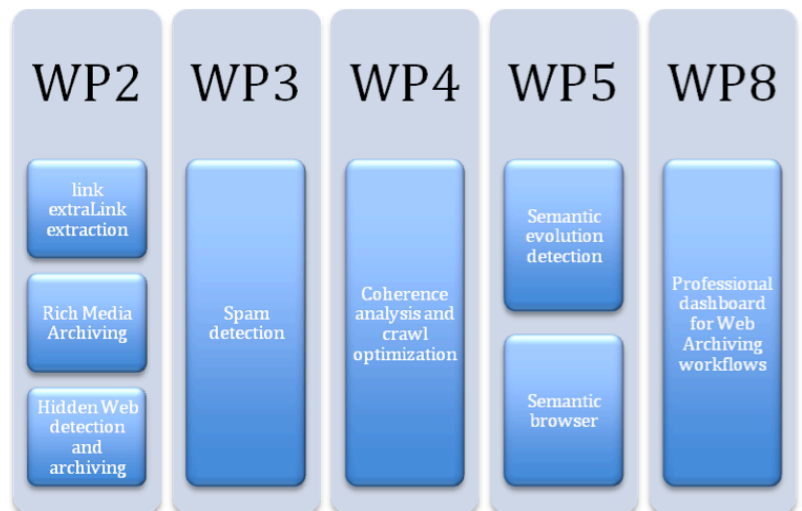
Technologies to archive the Web have been developed at a time when the Web was much more static and technically easy to capture than it is today. Spanning of new publication and authoring technologies, spreading of videos, wider participation and emergence of web spam pose serious challenges to web preservationists.

LiWA's aim is to develop new tools and approaches that can deal with these challenges and improve archive fidelity, coherence, and interpretability. The strategy is to frame them so that they are more resistant to future changes, by leveraging universal tools like the browser and by fostering community-based responses to new challenges.

LiWA research is built with scalability in mind and with the aim to be adapted to web archiving goals in institutions with varying collection policies.

Two demonstrators will be developed, one on archiving video rich websites from the TV websphere and the other on tool for helping archivists to capture the social web.

LiWA is a Seventh Framework Programme (FP7) ICT research project funded by the European Commission ■



LiWA Technologies



New LiWA videos

Want to see us, rather than read us? Look at our video section on <http://liwa-project.eu/index.php/video/>

Project partners

The project lead is the L3S Research Center in Hanover, Germany. The other partners are:

- European Archive Foundation, The Netherlands
- Max Planck Institut for Computer Science, Germany
- Computer and Automation Research Institute of the Hungarian Academy of Sciences, Hungary
- Netherlands Institute for Sound & Vision, The Netherlands
- Hanzo Archives Limited, England
- National Library of the Czech Republic, Czech Republic
- Moravian Library, Czech Republic

How to contact Liwa ?

Dr. Thomas Risse
L3S Research Center
Appelstrasse 9a
30167 Hannover - Germany
Phone: +49 (0) 511 - 762 17764
email: info@liwa-project.eu

LiWA Website
<http://liwa-project.eu>



Incoming events

We can already announce that LiWA will be presented at the following conferences this year (more to come):

AIRWeb 2009: Fifth International Workshop on Adversarial Information Retrieval on the Web 20 or 21 April 2009, in conjunction with the WWW2009 conference in Madrid, Spain.

IWAW 2009: 9th International Web Archiving Workshop in October 2009.

Capture of rich and complex Web content

The aim of this group is to improve archive completeness. We are specifically developing tools able to

- find links to web resources regardless of the encoding, using virtual browsing
- handle streaming protocols to capture rich media Web sites
- detect and capture structural hidden Web
- orchestrate various tools at crawl-time based on priorities and constraints of the crawl.

During this first year, we have been concentrating on the first two items mainly working on two promising techniques. Both are utilizing 'helper applications' to supplement the work of the crawler enabling us to reach content that would be otherwise inaccessible.

Using virtualized browsing, we are able to interpret Javascript and discover links that are only generated by execution. An interesting aspect of this approach is that it will be possible to

apply it to future technologies as well, as long as browsers can handle them.

To deal with streamed videos, we have been able to capture non-http video content using the open source media player "mplayer". A streaming archiving module has been adapted to work with the Heritrix open source crawler. Next year, we will integrate and improve both. We will also start work on the hidden web and crawl-time orchestration of tools ■

Spam cleansing

The ability to identify and prevent spam is a top-priority issue for the search-engine industry but less studied by Web archivists in spite of the quotes that estimate roughly 10% of the Web sites and 20% of the individual HTML pages constitute spam. The above figures directly translate into 10-20% waste of archive resources in storage, processing and bandwidth in

addition to the deterioration of the archive quality. Types of archives in particular sensible to spam are:

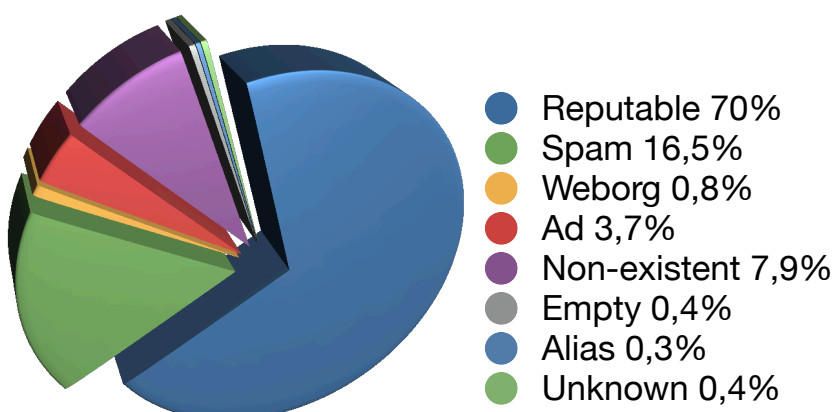
- Bulk crawls of an entire domain (all types of spam: keyword stuffing, content hiding, link farms, honey pots);
- Community content archives (comment spam: responses, posts or tags not related to the topic containing link to a target site or advertisement).

Web spam filtering know-how became widespread with the success of the Adversarial Information Retrieval Workshops since 2005 that host the Web Spam Challenges since 2007, both with active and successful participation by LiWA members. In order to tie the bonds between the archival and spam filtering communities we have provided the LiWA members time-aware Web spam benchmark data set WEBSpam-UK2006-2007 of 13 crawl snapshots simulating the operation of a real Web archive. Our deepest thanks to the Lab of Web Algorithmics in Milan for providing us with their crawl data.

The LiWA spam filtering service is designed to provide a collaboration tool for knowledge and feature sharing across participating archival institutions. Archive operators will be able to unite their effort, in particular against types of spam that span across domain and archive boundaries ■

Distribution of web pages

2004 de Crawl. Courtesy: T.Suel



Temporal Coherence

To comply with the politeness specification of a web site, archiving crawlers need to pause between subsequent http-requests in order to avoid unduly high load on the site's http-server. As a consequence, capturing a large web site may span hours or even days, and

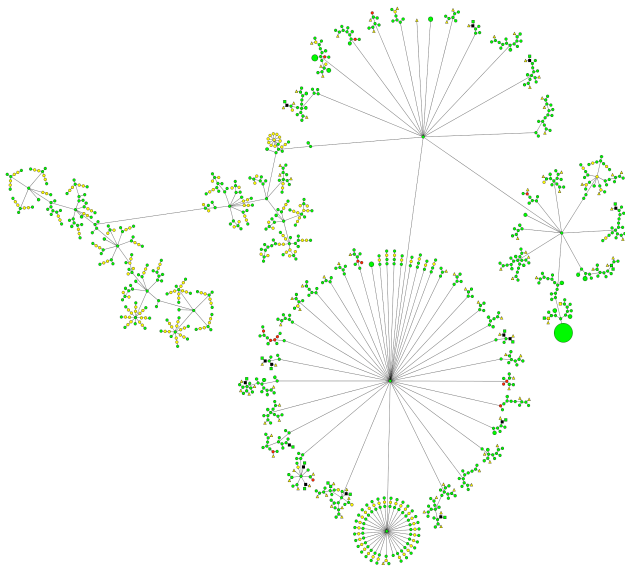
changes during this time period and temporary unavailability are the norm. Consequently, you may never be sure if the contents collected so far are still consistent with those contents the crawler needs to retrieve next. Therefore, questions arise about detecting changes and measuring their impact on the coherence and, ultimately, quality of a web archive. The LiWA research on archive coherence aims at specifying measures that identify the degree of change having occurred during a site crawl. In order to identify these changes and to archive contents as coherent as possible, recent research results focus on:

- Proper dating technologies
- Analysis of capture defects

Proper dating technologies are required to know how fresh a Web page is – that means – what the date (and time) of last modification is. The canonical way for time stamping a Web page is to use its Last-Modified HTTP header, which is unfortunately unreliable. For that reason, another dating technique is to exploit the content's semantic timestamps. This might be a global timestamp (for instance, a date

preceded by "Last modified:" in the footer of a Web page) or a set of timestamps for individual items in the page, such as news stories, blog posts, comments, etc. However, the extraction of semantic timestamps implies the application of heuristics, which imply a certain level of uncertainty. Finally, the most costly – but 100% reliable – method is to compare a page with its previously downloaded version.

The analysis of capture defects measures the quality of a capture either directly at runtime (online) or between two captures (offline). To this end, we generate per capture sophisticated statistics (e.g. number of defects occurred sorted by defect type). In addition, the capturing process is traced and enhanced with statistical data for exports in graphML. Hence, it is possible to layout a capture's spanning tree and visualize its defects (see figure). This visual metaphor is intended as an additional means to automated statistics for understanding the problems that occurred during capturing ■



Defect visualization of a sample mpi-inf.mpg.de crawl, using Visone.

Semantic Evolution

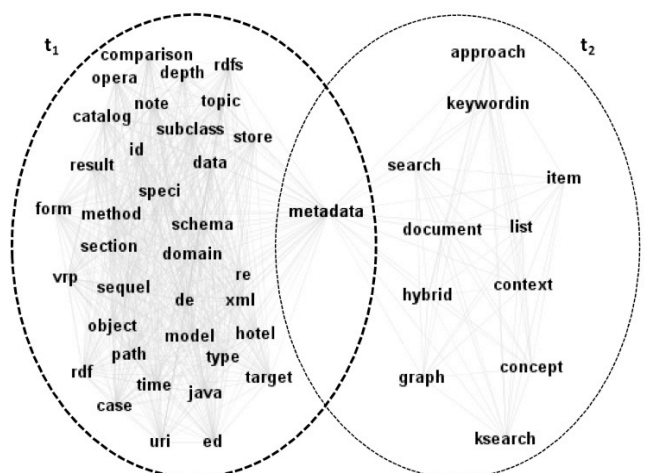
The correspondence between the terminology used for querying an archive and the one used in content objects to be retrieved, is a crucial prerequisite for effective retrieval technology. However, as terminology is evolving over time, a growing gap opens between older documents in (long-term) archives and the active language used for querying such archives. Thus, technologies for detecting and systematically handling terminology evolution are required to ensure "semantic" accessibility of archived content in the long run. The core of our approach is to derive mappings between terminologies originating from different times by the fusion of term-concept graphs.

In the context of LiWA, the semantic evolution group has developed and presented a formal model for detecting Terminology evolution. The model was presented during the 8th International Web Archiving Workshop. To verify the model we conducted some initial experiments. For these, we selected the domain of Semantic Web because the

research field was young around 2001 and has evolved to an established domain since. Therefore, we expect to identify evolution of terms or their meanings.

We compared two datasets concerning Semantic Web, the documents in the first collection originating from 2001-2002 and in the second one from 2006 and 2008. For these datasets we extracted senses for the present nouns and noun phrases. The picture shows a subset of the experiments.

We were able to detect these clusters, among others, representing senses for the term "metadata". The second cluster



ter lacks a pre-image among the clusters from the earlier collection, mainly because most of its members are new to the later collection. We can draw the conclusion that the cluster concerning ksearch, search, keyword is an emerging sense of metadata. A term that was previously used with terms like rdf, xml, schema is now also used in the context of search ■