



**Thanks for your interest in the LiWA project! This third newsletter summarizes results and findings of the LiWA project during the 36-month project runtime. A focus is done on LiWA Technologies at work.**

#### Last news:

**The final project review took place in Luxembourg on March 11th, 2011. Partners were congratulated by the European Commission for their relevant contributions on Web Archiving Research and the project outcomes.**

Web content plays an increasingly important role in the knowledge-based society, and preservation and long-term accessibility of Web history have high value. The European funded project LiWAs looked beyond the pure “freezing” of Web content snapshots for a long time, by transforming pure snapshot storage into a “Living” Web Archive. In order to create Living Web Archives, the LiWA project addressed R&D challenges in the following three areas: Archive Fidelity, Archive coherence and Archive Interpretability.

## Archive Fidelity

The first problem area is the archive's fidelity and authenticity to the original. Fidelity comprises the ability to capture all types of content. Current crawlers fail to capture all Web content, because the current Web includes much more than simple HTML pages as dynamically created pages, multimedia content that is delivered using media-specific streaming protocols; hidden Web content that resides in data repositories and content-management systems behind Web site portals.

In this context LiWA develops a new crawler based on executing pages. This method requires three specificities:

The first is to run an execution environment (HTML plus JavaScript, Flash etc.) in a controlled manner so that discoverable links can be extracted systematically.

The second challenge is to encapsulate these headless browsers in a crawler-like workflow, with the purpose of systematically exploring all the branches of the web graph.

The last but not the least of the challenges, is to optimize this process so that it can scale to the size required for archiving sites.

This three challenges have been implemented in the new crawler that one of the partner has developed (Hanzo Archives Ltd) and it is already used in production by them to archive a wide range of sites that can't be archived by pre-existing crawlers, as well as in testing by another of the LiWA partner, the Internet Memory Foundation .

Some tests on small scale have been made to compare results between 2 methods. Analyses show significant improvements in the quality of the crawls. It is worth noting that these improvements are obtained on small-scale crawls. Indeed, the link extractor increases processing time, however it is largely compensated by the fact that it saves human operator's time.

The screenshot below shows the difference between the 2 methods. A visual control shows clearly that one capture is more complete than the other one.

Figure 1: Heritrix



Figure 2 : Executing pages robot



## 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](mailto:info@liwa-project.eu)

## Archiving Rich Media Content

As part of the new technologies for Web archiving developed in the LiWA project a specific module was designed to enhance the capturing capabilities of the crawler for different multimedia content types specifically when served via streaming.

The *LiWARich Media Capture* module delegates the multimedia content retrieval to an external application (such as MPlayer or FLVStreamer) that is able to handle a larger spectrum of transfer protocols than Heritrix. The module is constructed as an external plugin for Heritrix. Using this approach, the identification and retrieval of streams is completely decoupled, allowing the use of more efficient tools to analyze video and audio content. At the same time, using

the external tools helps in reducing the burden on the crawling process. At Internet Memory, the LIWA video capture module is used on a daily basis to fetch video served with streaming server. It has significantly improved the quality of archiving for video-centric sites (like broadcasters and TV sites for instance) but also for mainstream sites that use video hosting services (like Youtube).



An example is presented below a conference video between James Cameron (UK Prime Minister) and Mark Zuckerberg (founder of Facebook) on Youtube.

The screenshot compares the archive and the online video. On left online video and on right the archive video, the only difference is the archive's video player.



## How can an existing, quality focus, web archiving workflow be improved?

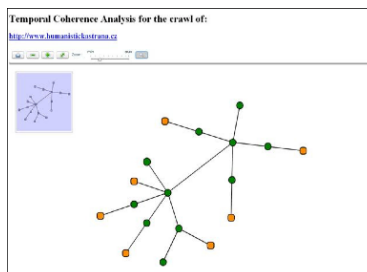
The National Library of the Czech Republic (NLP) has been building its Web archive since 2000. The archive is focused on "bohemical" web resources, i.e. websites that are related to the Czech Republic or its people.

The NLP strives to control the quality of archived websites from selective harvests. This manual, laborious and time-consuming process, often refers to quality assurance or simply QA. It basically requires visual inspection of all harvested websites by the project staff.

Archive coherence has brought a surprising and unexpected by-product in this respect: the results of a crawl's temporal coherence analysis can be used to generate a graphical representation of a website in the form of nodes representing individual pages and links between them (see Figure 4).

The colors of the nodes indicate whether pages have changed during the crawl (or, more precisely, between the initial crawl and a re-crawl).

Figure 4: Temporal coherence analysis of a website

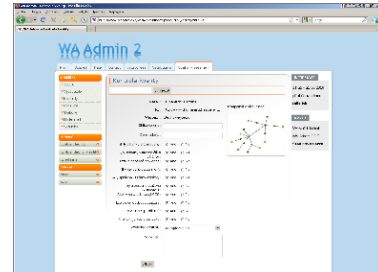


Some of these graphs are in fact little pieces of web art, but more importantly, they have a potential to reveal some irregularities or problems, such as crawler traps or missing pages. In addition, the distribution of the nodes' colour in the graph can indicate a rough estimate of the rate of change of the website.

Using the graphs for QA could bring some benefits as they can alert curators to the existence of crawler traps and other quality issues, which are easy to miss during manual QA. These Graphs are inserted into the QA workflow, thus allowing the curators a quick visual inspection.

Clicking on the graph will bring up an interactive version of it from Figure 5, which the graph can be navigated and zoomed. Hovering over a node displays the node's URL.

Figure 5: Integration of the TC module into WA admin



## LiWA Tools on Open source

LiWA released most of its developed components and tools in open-source like the RichMedia-Capture module, Spam Assessment interface or the Terminology Detection and Browsing.

<http://code.google.com/p/liwa-technologies/>

## SPAM

Internet archives are becoming more and more concerned about spam in view of the fact that, under different measurement and estimates, roughly 10% of the Websites and 20% of the individual HTML pages constitute spam. The objective is to reduce the amount of fake content the archive has to deal with. The toolkit helps prioritize crawls by automatically detecting content of value and exclude artificially generated manipulative and useless content.

With the illustration over 100,000 pages WEBSPAM-UK2007 data along with 7 previous monthly snapshots of the .uk domain, we have investigated the tradeoff between feature generation and spam classification accuracy. We proposed graph similarity based on temporal features, which aim to capture the nature of linkage change of the neighborhood of hosts. Our features achieve better performance than previously published methods; however, when combining them with the public link-based feature set we get only marginal performance gain.

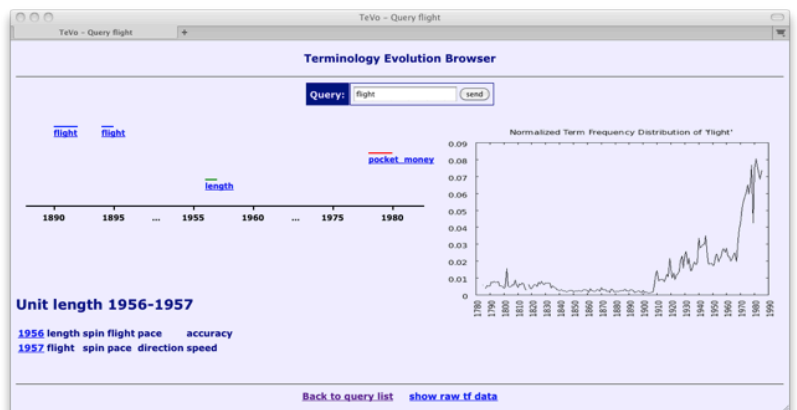
By our experiments it has turned out that the appropriate choice of the machine learning techniques is probably more important than devising new complex features. We have managed to compile a minimal feature set that can be computed incrementally very quickly to allow intercepting spam near crawl time. Our results open the possibility for spam filtering practice in web archives that are mainly concerned about their resources waste and would require fast reacting filters.

## Terminology Evolution Browser

In the longer term, Internet archives will have to deal with semantic evolution (new terms referring to the same thing or concept). LiWA has developed algorithm to analyse semantic drift through time as explained in previous issues of this newsletter.

To make the language evolution process accessible, A web service was developed, which allows exploring the evolution of a given term. As running example is presented below, based on the term "flight". After the user specifies the term of interest, we show all paths containing this term over time. As representative, we chose the term with the highest clustering coefficient (timeline on the left; Furthermore, we give the term frequency distribution of the term over time (on the right).

Figure 6: The TeVo Browser User Interface showing the paths and term frequency distribution for the term



## Dissemination results

Documentation and scientific publications have been submitted to high-level conferences, workshops and publishers in the field. One of the challenges for LiWA was that Web Archiving is a new field of research and had no established regular scientific publisher. In order to structure this, each year, a dedicated session for the LiWA project has been organized at the International Web Archiving Workshop, a workshop with a traditionally strong practical orientation (as opposed to a scientific one).

Dissemination Success Indicator	Results
Scientific outreach	
Presentations at relevant events & conferences	36
Scientific publications	24
Online Presence	
Website: # of visits per months	545
Video downloads (total)	3511
Professional outreach	
Participants in the Core User Group	32
LiWA Demonstration Workshops/Session	6
Software	
Downloads	55

# Commercial exploitation

Beyond research output and open source contributions, Liwa technology is already commercially used by two of the partners.

The Internet Memory Foundation has been launching in 2010 its new online web archiving platform ([Archivethe.net](http://Archivethe.net)) which integrates several of the modules developed in LIWA. A dozen heritage institutions in Europe already use this platform.



Furthermore, the breakthrough achieved by Hanzo Archives in the domain of Legal and Compliance-driven web archiving is impressive, with customers in the Global 2000 across United States and Europe. Because of the requirement to produce forensically sound archives, legal applications are certainly the most demanding for Web Archiving in terms of fidelity. The fact that Hanzo is leading this market worldwide can be seen as a great achievement of the LIWA



project, which came timely to provide the critical R&D support required to establish a European champion in this emerging market.

## Unlocking the value of Web Archives: a necessary evolution of the legal framework.

The work done in the LIWA project has demonstrated how it is possible to technically extend the fidelity and usability of web archives. However, several of the legal and societal challenges associated with preserving and using online society heritage remain.

Based on crawling, web archiving has made it's way following the trace of Web crawling indexers, benefiting from the fact that most web content producer find an immediate benefit to being referenced by search engine. However, when it comes to actually doing something with web archives many hurdles and uncertainty remain, especially the fact that legal frameworks lag significantly behind the evolution of web archive technology.

Two parallel developments in web archiving serve to illustrate this:

- The first, a new generation of mining and monitoring services and tools based on web data are already emerging at rapid pace, in a completely uncontrolled manner. First and foremost are a handful of large social network and search engine company that actually hold vast amount of data from the Web, thanks to their current position in the web ecology.

Due to legal risks, for example copyright laws, fair use, libel, etc., restrictions on access to archived websites and the data derived from these have resulted in a lack of open and transparent access to large Web data sets, which significantly hinders public research, even to provide society the ability to understand what can be done in this domain.

- The second, the rapidly growing commercial web archiving sector, which is driven largely by legal e-discovery, records management, and regulatory compliance needs experienced by large companies around the world.

Companies need to capture their web content (websites and social media) for these reasons and are hampered by a lag in the laws and regulations governing the ability to crawl and provide access to web content held on third party platforms, such as Facebook, Twitter,

and so on. This will lead to some interesting cases in the near future and content owners and platform providers work out their respective rights.

In addition, copyright law similarly hampers providing open access to archived material in a social context and legal precedents set by the music and film industries, where content sharing is equated to piracy. Under these conditions a social web archive, in which archived content can be stored and accessed in an open manner will struggle to prevail.

Although many web producers don't consider web archives to be directly and immediately useful for them, they clearly bring a new and valuable service by building a memory of this media. No one has evaluated the waste and frustration that this lack of memory generates partly because most people take this as a necessary limit of the media, which it is not. From this point of view, the fact that 41% of institutions responding to our survey provide on-line access without restriction is very encouraging.

However, to avoid the creation of web data monopoly, it is important that open web archives are given the leeway for their development, especially that they be allowed to give open access, facilitate research and even develop new innovative services in the future. For this to happen, governments and law makers across the major jurisdictions of the world should look to the future with much longer lens – so much is at stake, and short term gains through restrictive copyright laws will only destroy this crucial next step in the evolution of the internet.

### Contact:

Archivethenet:  
[contact@archivethe.net](mailto:contact@archivethe.net)

Hanzo  
[contact@hanzoarchives.com](mailto:contact@hanzoarchives.com)

---

## To continue

---

Although LiWA is first and foremost a research project (the first of its kind in Web Archiving in the world), the results are already very valuable to practitioners of the field.

During the LiWA project many new approaches have been developed to address major issues in Web archiving and archive accessibility. LiWA can be seen as a starting point for a number of new activities in the field of Web archiving and Web preservation. The sheer size, complexity and dynamics of the Web make high quality Web archiving still an expensive and time-consuming challenge. Therefore new crawling strategies are necessary that focus on content completeness in term of opinions, topics or entities etc..

The new Integrated Project **ARCOMEM** (From Collect-All Archives to Community Memories) leverages the Social Web for content appraisal and selection.

Beside preservation, a deeper understanding of the Internet content characteristics (size, distribution, form, structure, evolution, dynamic) is also necessary in many areas of today's science.

The European funded project **LAWA** (Longitudinal Analytics of Web Archive data) builds an experimental testbed for large-scale data analytics. Its focus is on developing a sustainable infrastructure, scalable methods, and easily usable software tools for aggregating, querying, and analysing heterogeneous data at Internet scale. Particular emphasis is given to longitudinal data analysis along the time dimension for Web data that has been crawled over extended time periods.

### More about:

**ARCOMEM:**



<http://www.arcomem.eu/>

**LAWA:**



<http://www.lawa-project.eu/>

[info@lawa-project.eu](mailto:info@lawa-project.eu)

### Project partners

**L3S Research Center**, Germany (coordinator)  
**Internet Memory 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](mailto:info@liwa-project.eu)

<http://www.liwa.project.eu>