Thanks for your interest in the LiWA project! As the first Newsletter was presenting the LiWA Research Project in a nutshell, this second newsletter will introduce the LiWA Application work packages. Your feedback or questions are welcome to info@liwa-project.eu. To subscribe and get informed, please visit http://liwa-project.eu

The LiWA Applications in a nutshell

LiWA is concentrating a large part of its research effort to improve the ability to capture in an authentic and extensive manner today's web, regardless of the publishing technology used. This encompass new capture methods (based on execution of pages), streaming archiving, hidden web and more.

Archiving at scale today's web also requires the ability to automatically filter out spams and traps that represent a large portion of the web (up to 20%).

To achieve this, LiWA develops a serie of modules (capture enhancement, spam detection, temporal coherence and semantic evolution) that were presented in LIWAnews#1.

Technologies are developped to be used either at crawl-time or after completion of the crawl, integrated with existing web archiving workflow.

In order to test and apply these new methods and results, an integration platform of the modules is being built both by the European Archive Foundation (using open source tools) and Hanzo Archives.

Why two platforms? We are convinced that by using different environments (crawlers etc.), we can better demonstrate the versatility of the LiWA results as well as their applicability in different contexts: heritage archiving for EA using Heritrix large scale crawler and commercial corporate compliance archiving for Hanzo Archive using its high-quality focussed crawler.

This will also be illustrated in two applications scenarios, one aiming at integrating web material in a large audio-visual archive, the other focussing on archiving the social web.

Both applications are built upon LiWA technologies, integrated in real world scenario whose scope is wider than what LiWA specifically adresses.

The first application is dealing with the integration of web rich material (specifically audio and video) into one of the largest european audio-visual archive (Beeld en Geluid).

The second one, on archiving the social web, is developed by both an experienced web archiving library and a commercial company proposing its archiving services to corporates using the social web for their communication.

Together, these two applications cover a rich range of type of use and type of content. From migration of tradtional media online (TV, radio etc.) to the emerging social web, applicability and effectiveness of LiWA technologies will be tested on the most challenging areas of today's web.

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

– L3S Research Center, Germany (coordinator)
– 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 ?

info@liwa-project.eu

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

Streaming application

LiWA technology for content and context in Sound and Vision archive

The Netherlands Institute for Sound and Vision is one of the largest audiovisual archives in Europe. Its archive consists of around 700,000 hours of Dutch television, radio, music and film and its collection grows every day. The cultural heritage preservation policy of the Institute implies that the AV archive should preserve the Dutch *audiovisual cultural heritage*. As the Internet is increasingly becoming an important source for (user generated) audiovisual cultural heritage content, Sound and Vision has a strong commitment to capture information available on the Web. More specifically, the institute is eager to capture broadcast related websites, including streaming content.

However, as capturing streaming content from the web is difficult, until now only a limited selection of user generated video content is downloaded manually from the internet. This content is used in the exhibition space of the institute.
With the streaming content capturing technology developed in the LiWA

project, Sound and Vision is able to address the capturing of Dutch cultural heritage content in a much more efficient way.

Besides being a potential provider of audiovisual content (i.e. video's available on countless websites such as YouTube, Vimeo, Blip.tv and many others), the Web is regarded as a valuable source for gathering *contextual information* that relates to the audiovisual collections. Context information is relevant for both archivists, and other users interested in a specific broadcast or a broadcasting related topic, such as broadcast professionals, journalists, teachers, citizens, researchers and so on. Typically, these users have

to collect the information and the audio or video broadcasts from several sources and need to use different interfaces to search these sources. Ideally, Sound and Vision provides these users with a single interface that allows searching both the digital asset management system of the AV archive (iMMix) and related web content. Figures 1 and 2 show mock-ups of such an interface that will be developed in LiWA ■
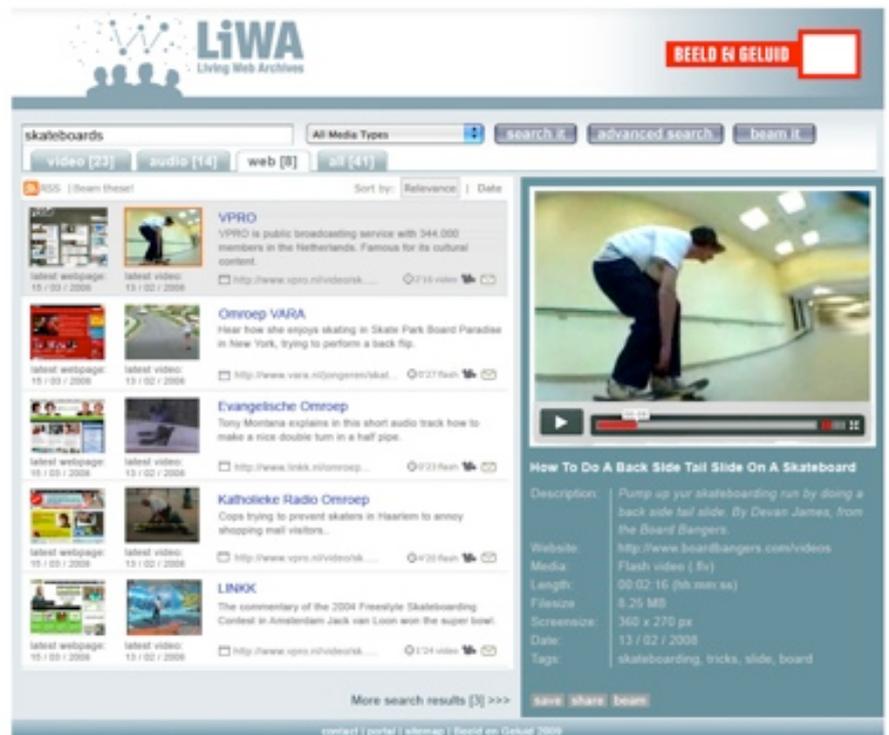

Figure 1 Mock-up of a user interface for browsing various types of archived web content: video and audio (AV content), and web pages (context).

Figure 2 Mock-up of a user interface for browsing the broadcast related web archive

Create and enrich metadata

In order to maximize the potential use of available context information sources, Sound and Vision is currently developing an infrastructure for the *automatic* deployment of context data. The aim here is to build a platform for context data aggregation, analysis and access. On the aggregation level, various types of context documents are collected. Here LiWA comes in by providing technology and maintenance tools for the capturing of web content.

In the next level of the context data platform, raw context data such as web crawls are analyzed and ultimately linked to items in the Sound and Vision archive. How to automatically create links between archived content and other information sources in an efficient, effective and robust manner will be addressed in the recently started Dutch research project BRIDGE that is part of the CATCH (Continuous Access to Cultural Heritage) program.

Finally, the context data platform provides various access services so that both (i) end users and (ii) automatic processes can make use of the added value of having AV content connected with context. Examples of automatic processes are keyword recommendation and speech recognition ∎

Recommendation and recognition

Keyword recommendation is part of a Documentalist Support System under development at Sound and Vision. On the basis of automatically aggregated context information and semantic analysis, this system suggests relevant keywords from the thesaurus (Common Thesaurus Audiovisual Archives, GTAA so that a documentalist who is annotating the data can quickly pick terms from a relatively small list instead of going through a list of more than 150.000 terms.

Speech recognition is currently used at the Institute to automatically generate metadata based on the speech that is present in the video content. Speech is converted to text with time-codes that can be indexed to allow fine-grained, fragment level access to audiovisual documents. Speech recognition systems need to be adapted to specific task domains for example to learn which words they should be able to recognize. When a word is 'out-of-vocabulary' (OOV) it can never be recognized. Context data can help to solve this OOV problem in speech recognition so that adequate speech recognition performance levels can be obtained ∎

LiWA

The LiWA application Streaming demonstrates how broadcast related web content could be accessed by potential end users. How users within Sound and Vision value this presentation of web content will be evaluated. The archived content will be used as test data for the development of the Sound and Vision context data platform that specifically addresses the linking of web context to the digital asset management system of Sound and Vision, iMMix ∎

Social web application

Social web sites typically contain highly inter-linked content and use dynamic linking, widgets and tools as well as high degree of personalisation. Capturing social web sites is extremely challenging and cannot be fully achieved using current methods and tools. Social web thus represents one of the greatest challenges in web archiving. The tools and technologies developed within LiWA can be an answer to such a challenge.

With the Social web application, LiWA intends to demonstrate a dramatic improvement in both archive structure and content completeness so that the rapidly evolving and increasingly diverse content of the social Web is captured more accurately and evenly.

The aim of the application is to show how the LiWA technology fits in the workflow of an active Web archiving institution, by considering a real-life scenario of the National Library of the Czech Republic. The application is designed as a set of independent mo-

dules developed in LiWA work packages 2 (capture of rich content), 3 (data cleansing and noise filtering), 4 (archive coherence) and 5 (semantic evolution). The modules can be readily integrated with existing Web archiving workflow management tools. A Web

archiving institution can choose to deploy all of the modules or just some of them, depending on its needs and particular workflow. The application is designed as generic and can be used to enhance archiving of any type of web content, not just social web ■


The next step will involve finding temporal incoherencies with the coherence graph. The graph indicates parts of a site where temporal incoherence patterns can be found and allows the Quality Assurance (QA) operator to click on nodes to see the problematic sections of the site. Clicking in the graph will bring up the archived version of the website ■
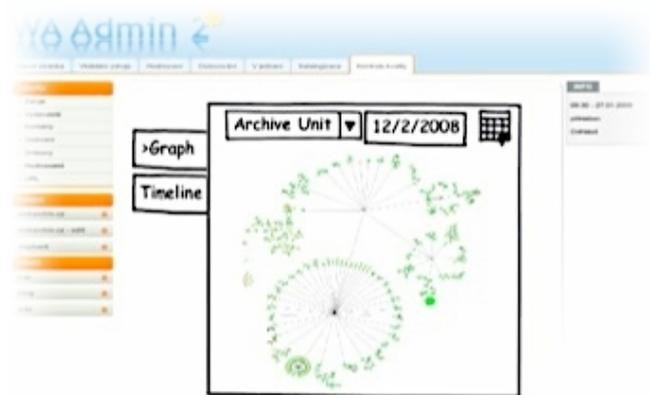
An exciting tool is the advanced **link extractor**. Whenever sections of a website were not harvested due to JavaScript etc., the curator can run a virtual browser that will generate a list of all links on the page. Missing links will be added to the seed list for a re-crawl or next harvest. The link extractor is shown in picture bellow – the top half of the screen shows a challenging site to archive while the bottom half displays a list of discovered links.

Quality assurance also relies on the **spam detection engine**. Whenever curators discover spam or other irrelevant content during this stage, they can use the spam module to mark it as such and document the relevant feature set for enhancing detection accuracy. The curators may also manually accept or reject the decision of the automatic filter ■
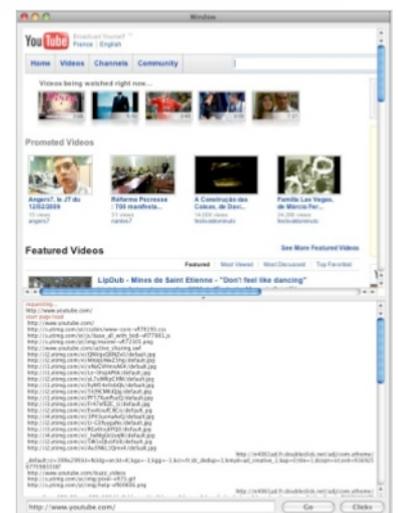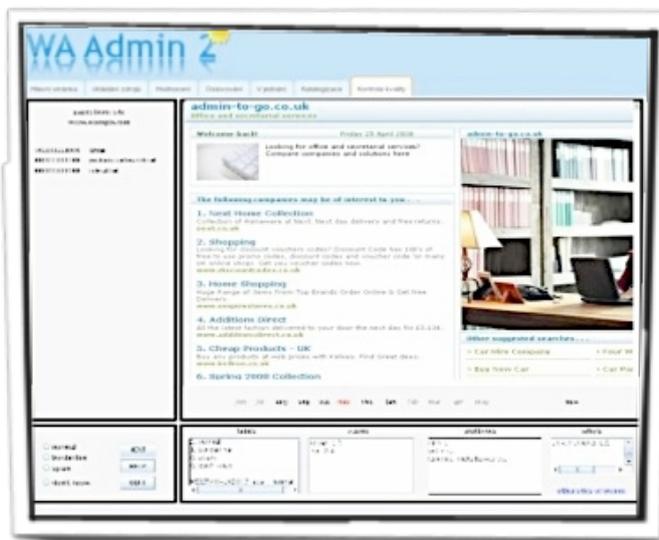
The National Library of the Czech Republic provides unlimited online access to a part of its web archive covered by permissions from authors and publishers. If we want to provide the best possible experience and a high-quality service for our users we need to be very concerned about the quality and completeness of the archived and displayed content. It is therefore crucial that we put a lot of effort into managing quality assurance processes. We believe that the new LiWA tools will help us significantly to improve the quality of our archive and its completeness while reducing the amount of time and resources spent! ■