LiWA Application
"Social Web Archiving"
Prototype Report
Deliverable No: 8.3
Version 1.0
Project Overview

Project Name: LiWA – Living Web Archives
Call Identifier: FP7-ICT-2007-1
Activity Code: ICT-1-4.1
Contract No: 216267

Partners:
1. Coordinator: Universität Hannover, Learning Lab Lower Saxony (L3S), Germany
2. European Archive Foundation (EA), Netherlands
3. Max-Planck-Institut für Informatik (MPG), Germany
4. Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI), Hungary
5. Stichting Nederlands Instituut voor Beeld en Geluid (BeG), Netherlands
6. Hanzo Archives Limited (HANZO), United Kingdom
7. National Library of the Czech Republic (NLP), CZ
8. Moravian Library (MZK), CZ

Document Control

Title: LiWA Application “Social Web Archiving”
Prototype Report

Author/Editor: Mark Middleton, Libor Coufal, Julien Masanès, Thomas Risse

Document History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Author/Editor</th>
<th>Description/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>V 0.1</td>
<td>11/2/2011</td>
<td>Julien Masanès, France Lasfargues</td>
<td>First draft with comparative test results</td>
</tr>
<tr>
<td>V 0.2</td>
<td>17/2/2011</td>
<td>Libor Coufal, Thomas Risse</td>
<td>Introduction, Revision</td>
</tr>
<tr>
<td>V 0.3</td>
<td>20/2/2011</td>
<td>Mark Middleton, Julien Masanès</td>
<td>Section 3. Revision</td>
</tr>
<tr>
<td>V 1.0</td>
<td>21/2/2011</td>
<td>Thomas Risse</td>
<td>Final Version</td>
</tr>
</tbody>
</table>
Legal Notices

The information in this document is subject to change without notice.

The LiWA partners make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The LiWa Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.
Table of Contents

1 INTRODUCTION 5

2 TEST 6
   2.1 Presentation 6
   2.2 Traditional crawl 7
   2.3 LiWA Web social application crawl 8

3 INTEGRATION IN AN ARCHIVING WORKFLOW 10

4 CONCLUSIONS 14
1 Introduction

Social media and the social web are becoming more and more pervasive in all types of activities. In contrast to the “classic” Web, users are generating huge amounts of content in different kinds which range from simple tagging to complex blog entries or videos. Sites like Facebook, Flickr or Youtube are actively supporting the content generation and social interactions. As a result the web pages generated by social media sites are in no way comparable to traditional HTML pages. Such pages consist of complex web objects with many properties and functionalities. In addition popular pages have a high change frequency and also the network of pages they are involved in are constantly growing. For standard web crawlers, and therefore for Web Archivists, these pages represent a huge challenge.

The LiWA technologies developed in WP2 and WP4 are helping (1) to ease the crawling of single complex pages and (2) better crawl scheduling and analysis of social media sites.

The aim of WP8 is to demonstrate the applicability of the LiWA technology in two ways. The dramatic quality increase when using LiWA technology is demonstrated by a comparative social web site crawl done with LiWA technology and a standard Heritrix web crawler. In addition, it will be demonstrated how Social Media can be integrated in a professional archiving workflow.
2 Test

2.1 Presentation

The goal of WP8 is to test LIWA technologies in the context of social Web archiving. The application developed is expected to cope with complex inter-linking of objects, widgets, tools, and the high degree of personalization typical of the social Web.

The goal of this test is to compare two captures of the same social web page crawled, one with Heritrix and the other with Hanzo’s crawler, based on the execution approach developed in LIWA. The target was chosen to be representative of a wide range of difficulties that can typically be found on the social Web.

A very good candidate for this is Facebook, which is known for the complexity of its page rendering scripts and that combine all types of content (text, photos, video, games). It has been decided to collect the Vitaminwater Facebook public profile.

On the centre of the Facebook profile, several tabs allow user to access the Wall, Info, home, xgames, photos, event and more (“>>” button)

The left part contains list of people who ‘like this’, favourite pages, photos and videos.

To establish a comparison, NKP and Hanzo have launched two crawls on the same day (4th of February). NKP used Heritrix and Hanzo used Hanzo's crawler and Hanzo's web archive access tools, which provides the foundation of the LIWA social web application.

Access is made through a local proxy of the archive (to ensure all elements are rendered).
2.2 Traditional crawl

Heritrix results can be accessed at: http://har.webarchiv.cz:8080/AP2

Proxy settings:
host: har.webarchiv.cz
port: 8092

The capture presents the home page of the Facebook Vitaminwater profile. The homepage is displayed in Czech and not in English as Facebook uses geo-localization to serve content and this crawl was operated from Czech Republic.

Even if the images are presented, content in the center of the page is missing. User can click on thumbnails but nothing happens.

A lot of contents items are missing. Overall the archive gives only a general idea of what the page looks like but a large portion of the relevant information is missing.
2.3 LiWA Web social application crawl

#http://cc-qa.hanzoarchives.com/archive/20110204103047/vitaminwater?v=wall

Proxy settings :host: cc-qa.hanzoarchives.com
port: 8091

The homepage is displayed as the original. All the images and content items are archived and rendered.

User can click on thumbnails and browse through the different content items. In the Photos section, user can browse the catalogue photos, click on next, choose one and access to the photos and their comments. In the Event section, user can access all the events; navigate to the previous and next one.
On the wall, publications are complete with photos and comments. By clicking on the older one user can reach previous comments.

On the left part, profile and image are clickable and give access all the contents which are public.

The archival crawl made with the social web application offers a good capture of the Vitaminiwater Facebook Profile. Users can browse the archive as if they were online.
3 Integration in an archiving workflow

HZ and NLP have integrated the LIWA component in their existing workflow. This has been done with success to the point that during the course of the project, HZ has been proposed to become a specialist partner for integrating Web and Social Web archiving in Symantec, the leading provider in Enterprise Archiving and E-Discovery platform (Enterprise Vault).

The following screenshots present the successful result of this integration work. One of the challenges was to fully capture the conversational aspect of Social Web. Symantec needed to ensure that not only complex pages are fully captured, but also that the senders, receivers, viewers of the conversation are identified and documented. This has been possible in a relatively short period of time (6 months) thanks to the progress made with LIWA technologies and their integration. This is the first time that Social Media content can be fully integrated in such a platform. It will certainly extend to more platforms in the future.

The Social Media Archiving application is first presented in Hanzo’s environment, then in the Symantec Enterprise Vault.
Figure 1: Navigation in the Hanzo archive by archive policy and date

Figure 2: View of the Vitamin web site in Symantec Enterprise Vault. Note the From, To and CC fields, created to understand the conversation actors. All these fields (as well as the content of text) are searchable.
Figure 3: Same conversational view of the archive, but this time for Twitter.

Figure 4: Metadata documenting the archived Twitter pages.
Figure 5: PDF generated from the archive, with all elements of the original Twitter page.

Figure 6: Web view of the archived twitter pages in Symantec Enterprise Vault.
4 Conclusions

The social web is an ever-growing information space, which reflects the daily activity of many people. Social web pages are as complex and dynamic as the lives of their users. Preserving this content for the future results in many challenges. The major ones are addressed in the LiWA project.

In WP8 it has been successfully demonstrated that the LiWA technology can help to dramatically increase the quality of social media site archive. This has reached the point where some technological results of LiWA can already be implemented and used, such as in the legal discovery example shown.

The positive results indicate that the development is on the right track. They also show that the technology is capable of supporting a product level deployment within a commercial setting, specifically in Hanzo Archives.