Grant Agreement number: 216267
Project acronym: LiWA
Project title: Living Web Archives
Funding Scheme: FP7, Theme 3, Information and Communication Technologies
Date of latest version of Annex I against which the assessment will be made:
08 December 2008

Periodic report: 1st □ 2nd □ 3rd □ 4th □
Period covered: from 01 February 2008 to 31 January 2009

Name, title and organisation of the scientific representative of the project's coordinator¹:
Dr. Thomas Risse, Senior Researcher, Leibniz Universität Hannover, L3S Research Center
Tel: +49 511 762 17764
Fax: +49 511 762 17779
E-mail: risse@l3s.de

Project website² address: http://www.liwa-project.eu

¹ Usually the contact person of the coordinator as specified in Art. 8.1. of the grant agreement
² The home page of the website should contain the generic European flag and the FP7 logo which are available in electronic format at the Europa website (logo of the European flag: http://europa.eu/abc/symbols/emblem/index_en.htm
; logo of the 7th FP: http://ec.europa.eu/research/fp7/index_en.cfm?pg=logos). The area of activity of the project should also be mentioned.
Declaration by the scientific representative of the project coordinator

I, as scientific representative of the coordinator of this project and in line with the obligations as stated in Article II.2.3 of the Grant Agreement declare that:

- The attached periodic report represents an accurate description of the work carried out in this project for this reporting period;
- The project (tick as appropriate):
  - X has fully achieved its objectives and technical goals for the period;
  - □ has achieved most of its objectives and technical goals for the period with relatively minor deviations;
  - □ has failed to achieve critical objectives and/or is not at all on schedule.
- The public website is up to date, if applicable.
- To my best knowledge, the financial statements which are being submitted as part of this report are in line with the actual work carried out and are consistent with the report on the resources used for the project (section 6) and if applicable with the certificate on financial statement.
- All beneficiaries, in particular non-profit public bodies, secondary and higher education establishments, research organisations and SMEs, have declared to have verified their legal status. Any changes have been reported under section 5 (Project Management) in accordance with Article II.3.f of the Grant Agreement.

Name of scientific representative of the Coordinator: .................................................................

Date: ........../ ............/ ............

Signature of scientific representative of the Coordinator: ............................................................

---

3 If either of these boxes is ticked, the report should reflect these and any remedial actions taken.
4 If either of these boxes is ticked, the report should reflect these and any remedial actions taken.
1. Publishable summary

LiWA – Living Web Archives

The Web today plays a crucial role in our information society: it provides information and services for seemingly all domains, it reflects all types of events, opinions, and developments within society, science, politics, environment, business, etc. Due to the central role the World Wide Web plays in today's life, its continuous growth, and its change rate, adequate Web archiving has become a cultural necessity in preserving knowledge. Consequently a strong growing interest in Web archiving library and archival organizations as well as emerging industrial services can be observed.

However, web preservation is a very challenging task. In addition to the “usual” challenges of digital preservation (media decay, technological obsolescence, authenticity and integrity issues, etc.), web preservation has its own unique difficulties:

- distribution and temporal properties of online content, with unpredictable aspects such as transient unavailability,
- rapidly evolving publishing and encoding technologies, which challenge the ability to capture web content in an authentic and meaningful way that guarantees long-term preservation and interpretability,
- the huge number of actors (organizations and individuals) contributing to the web, and the wide variety of needs that web content preservation will have to serve.

A first generation of Web archiving technology has been built by pioneers in the domain like the Royal Library of Sweden and the Internet Archive based on existing search technology. It is now time to develop the next generation of Web archiving technology, which is able to create high-quality Web archives overcoming the limitations of the previous generation. The aim of the European funded project LiWA (IST FP7 216267) is to create innovative methods and services for Web content capture, preservation, analysis and enrichment.

Project Objectives

The LiWA project, started in February 2008 and led by the L3S Research Center at the Gottfried Wilhelm Leibniz University Hannover, brings together a consortium of highly qualified researchers (LUH, Max Planck Society, Hungary Academy of Science), industrial users (European Archive, Hanzo Archives), and archiving organizations (Sound and Vision, National Library of the Czech Republic, Moravian Library). It is the intention of the project partners to turn Web archives from pure Web page storages into “living Web archives” within the next three years. Such living archives, will be capable of: handling a variety of content types, dealing with evolution as well as improving long-term content usability.

In order to create Living Web Archives, the LiWA project will address R&D challenges in the three areas: Archive Fidelity, Archive coherence and Archive interpretability:

- **Archive Fidelity**: development of effective approaches and methods for capturing all types of Web content including the Hidden and Social Web content, for detecting capturing traps as well as for filtering out Web spam and other types of noise in the Web capturing process.

- **Archive Coherence**: development of methods for dealing with issues of temporal Web archive construction, for identifying, analysing and repairing temporal gaps as well as methods for enabling consistent Web archive federation for fostering synergies between Web archiving stakeholders;
• **Archive Interpretability**: development of methods for ensuring the accessibility, and long-term usability of Web archives, especially taking into account evolution in terminology and conceptualization of a domain;

The results of the project will be demonstrated within two application scenarios namely “Streaming Archive” and “Social Web Archive”. The Streaming Archive application will showcase the building of an audio-visual Web archive and how audio and video broadcast related web information can be preserved. The Social Web application will demonstrate how web archives can capture the dynamics and the different types of user interaction of the social web.

**Work completed in Year 1**

The first project year focused on the detailed understanding of the requirements, the system architecture and the development of initial approaches to address the identified challenges. The requirements analysis collected the requirements from three different angles. The user angle describes the desirable usage of web archives by libraries and archives. The technical angle collects functional requirements necessary to meet the user requirements of libraries and archives and the intention to extend the current state-of-the-art in/of web archiving. Finally the architecture angle defines functional requirements necessary to integrate LiWA services into one advanced web archiving infrastructure.

Based in the requirements analysis initial approaches and prototypes has been developed in the four technology areas of LiWA. In the following the major achievements are presented in more detail.

*Capture of Rich and Complex Web content*

The aim of this working area is to improve dramatically the fidelity of Web archives by enabling capture of content defeating current Web capture tools. This comprises the ability to find links to resources regardless of the encoding using virtual browsing, the detection and capture of structural hidden Web and the capacity to handle streaming protocols to capture rich media Web sites.

In order to develop an interpretation/execution-based link extractor for complex and dynamic objects, potential Javascript rendering engines for tasks were identified and tested. The comparison lead to select "WebKit" for implementation as it offers a huge number of features like JavaScript getters and setters, DOM class prototypes, significant JavaScript speed improvements, support of new CSS3 properties. DOM manipulation issues were analysed in depth to develop better links extraction. Various strategies to manipulate DOM from Webkit were tested. The result is a customized version of WebKit for the special use of link extraction.

For capturing rich media open source modules and helper application to support AV applications were tested. The Mplayer was selected as the basis for the helper tool implementation. In order to develop an improved rich media capture module, the crawlers were de-coupled from the identification and retrieval of streams and then moved to a distributed architecture where crawlers communicated with stream harvesters through messages.

*Data Cleansing and noise filtering*

The main objective of the first year was the reduction of the amount of fake content the archive will have to deal with. The envisioned toolkit will help prioritize crawls by automatically detecting content of value and exclude artificially generated manipulative and useless content.

Spam classification and page-quality assessment is a difficult issue for search engines; for archival systems it is even more challenging as they lack information about usage patterns (e.g., click profiles) at capture time. We survey methods that fit best the needs of an archive that are capable
of filtering spam during the crawl process or in a bootstrap sequence of crawls. Our methods combine classifiers based on terms over the page and on features built from content, linkage and site structure.

Archive Coherence

A common notion of “coherence” refers to the explanations given in the Oxford English Dictionary (cf. http://dictionary.oed.com) describing coherence as “the action or fact of cleaving or sticking together”, which - in terms of a Web site - results in a “harmonious connexion of the several parts, so that the whole ‘hangs together’”. From an archiving point of view, the ideal case to ensure highest possible data quality of an archive would be to “freeze” the complete contents of an entire Web site during the time span of capturing the site. Of course, this is illusion and practically infeasible. Consequently, one may never be sure if the contents collected so far are still consistent with those contents to be crawled next. However, temporal coherence in Web archiving is a key issue in order to capture digital contents in a reproducible and, thus, later on interpretable manner. To this end, we are developing strategies that help to overcome (or at least identify) the temporal diffusion of Web crawls that last from only a few hours up to several days. Therefore, we have developed a coherence framework that is capable of dealing with correctly as well as incorrectly dated contents. Depending on the data quality provided by the Web server, we have developed different coherence optimizing crawling strategies, which outperform existing approaches and have been tested under real life conditions. Even more, due to the development of a smart revisit strategy for crawlers we are also capable of discovering and (as a consequence) of ensuring coherence for contents, which are incorrectly dated and thus not interpretable with conventional archiving technologies. Summarizing, Coherence Analysis Technology V1 makes temporal coherence of Web archiving traceable under real life applications and provides strategies to improve the quality of Web Archives, regardless of how unreliable Web servers are.

Archive Interpretability

The correspondence between the terminology used for querying 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. Language changes are triggered by various factors including new insights, political and cultural trends, new legal requirements, high-impact events, etc.

An abstract model has been developed that allows the representation of terminology snapshots at different times (term-concept-graphs). In addition the model allows merging of term-concept-graphs while retaining time information. Beside the model for representing terminology evolution, necessary main functions have also been specified: representation of term concept relations, fusion of terminology graphs and mapping of concepts to terms. The analysis of documents for detecting terminology evolution is a complex process composed from many individual processing steps. Many tools in the areas of Natural Language Processing, Extraction, Sense Disambiguation/Discrimination, Clustering, etc. with different properties already existing. In LIWA the Apache UIMA framework is used as a basis for the integration of the existing and newly developed tools, which allow a flexible testing of approaches as well as an efficient processing of web archive documents in a later stage. In the first project year the framework has been configured and installed and allows the extracts of terms from a variety of formats like HTML or PDF.

Initial work has been done in the detection of terminology evolution. The current approach is based on co-occurrence analysis, clustering of terms and statistics. The assumption is that terms that often co-occur within a window have a semantic relationship. However many approaches for Sense Detection / Sense Discrimination already exist and the best approach for our scenario still needs to be identified or, if necessary, developed.
Applications

To showcase the improvements in Web archiving by LiWa and for the internal user evaluation of the developments, two applications will be implemented within the project. The “Streaming” application brings two spaces together, the Dutch broadcast Web sites and the digital asset management system from the Sound and Vision archive, and enriches them with streams and documents found in the Web archive. A mockup of the planned functionality has been developed in the first year of the project.

During the course of collation of the user requirements, it became clear that a need for Web archiving workflow user interface existed across the user partners. Therefore a generic Web archiving UI framework will be developed that will be usable to display information provided by the LiWA modules. This includes, for instance, the information provided on temporal coherence of archived sites and spam detection for quality checking purposes. Such UI widgets will also be usable for the spam assessment interface and the semantic evolution browser. To demonstrate the potential functionalities a mockup has been designed.

Anticipated impact of results

The most direct impact of the LIWA project results and of the capture, preservation, analysis, and enrichment services developed in this project is an improvement of the quality, comprehensiveness and accessibility of Web archives. These project results contribute to unlocking organizations’ and people’s ability to really master Web content - including its temporal perspective and dynamics. Furthermore, advanced Web archiving technologies, as they are developed in LiWA, clearly contribute to preserving Web content over time. The developed methods and technologies prepare Web content for long-term preservation by ensuring fidelity of the archive to the original content, improving coherence of the archived content and by taking provisions for its long term interpretability by developing methods for dealing with semantic evolution.

In general Web archiving technologies and their further evolution are important for many memory institutions. For memory institutions such as museums, which themselves create Web content as part of their service portfolios (e.g. virtual exhibitions), the use of technologies developed in LiWA eases the archiving of such content and opens up new ways of using the archived content (e.g. in later exhibitions, research, education, etc.). Other memory institutions, such as National Libraries, have an explicit archiving mandate which also includes Web content. This results in an even greater impact of improved Web archiving technologies on such memory institutions.

As a further impact of the project, adoption of the LiWA services will enable institutions and companies involved in Web archiving to provide better services and richer service portfolios to their clients. This may also provide additional motivation for institutions with an archiving mandate, like National Libraries and Archives, to invest more eagerly and successfully in Web archiving activities, since a clearer added value can be achieved.

More information about the project can be found on the LiWA homepage:
http://www.liwa-project.eu/
2. Project objectives for the period

To create Living Web Archives, the LiWA project will address R&D challenges in the following three areas: Archive Fidelity, Archive coherence and Archive interpretability:

- **To enhance archive fidelity and authenticity, LiWA will develop and test new methods based on content interpretation (automatic surfer) and intelligent pattern detection of traps and Web spam.** The objective is to reduce the amount of fake content the archive will have to deal with and help prioritize crawls by automatically detecting content of value, independent of the techniques used to generate it and exclude artificially generated content, again independent of the cause, i.e. crawler trap or Web spam. LiWA will also test various approaches developed to harvest the hidden Web for indexing purposes and see how they can be applied in the domain of Web archiving. Increased archive fidelity will have an impact on both archiving institutions and archive users. The former will be able to automatically archive higher volumes of dynamic and volatile digital content, resulting in a significant increase of preserved digital content. The latter will benefit from the higher quality of archive content. The higher content quality enables improved search services, which –in turn- allows archive users to acquire relevant knowledge more efficiently from the Web archive.

Objectives for the first year include the implementation of an interpretation/execution-based link extractor for complex and dynamic objects and the work on the rich media capture module. For the detection and processing of spam the objectives in the first year were to build the baseline technology for distinguishing useful, high-quality Web content from spam content.

- **To improve the integrity and temporal, structural and semantic coherence of Web archives and to enable archive users and archivists to assert control in this area, LiWA will develop a wide range of facilities for archive content analysis and enrichment.** For instance, LiWA will - for the first time in Web archiving - deal with the issue of temporal Web archive construction, including the constraints that crawling entails. This will serve the objective to significantly improve content positioning in time and (topic) space. In turn, this will improve access to and analysis capabilities for the often-complex objects contained in the archive. The achieved coherence will safeguard archive integrity over time, and will lay the foundations for fast and effective access to evolving Web content.

Objectives for the first year are the development of methods for the proper dating of web contents, ensure coherence of captures with respect to a timepoint or interval and the identification of contents violating coherence.

- **To facilitate archive interpretability, LiWA will apply methods for semantic and terminology extraction and will develop methods for detecting and handling evolving semantics and interpretation of domain concepts and terminology in the archive, thereby contributing to the objective of preserving – over the long-term - the usefulness, quality, and accessibility of Web archives.** This will allow the archive to keep track of evolving terminology semantics within its collection, thus significantly lowering the barrier to accessing and mastering of Web archive content.

The objectives of the first project year includes clear understanding of the terminology evolution problem, development of a terminology evolution representation format, implementation of the terminology extraction technology, testing of co-occurrence and clustering methods for the sense disambiguation and initial work on the Semantic Terminology Browser.

To validate the LiWA approach and illustrate the benefits of the developed services for the LiWA target groups, namely library and archiving organizations as well as other stakeholders interested in getting involved in Web archiving, two demonstrator applications will be built on top of the LiWA services. These applications will focus on the social Web and on the special challenge of archiving audio visual content, respectively.
3. Work progress and achievements during the period

WP1 Requirement Analysis

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 1</th>
<th>Start date:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Requirement Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH EA MPG MTA BeG HANZO NLP MZK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3)</td>
<td>2 4 7 2 4 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expended PM per beneficiary (Year 1)</td>
<td>2 4 8.46 2 4 3.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
The objective of WP1 is to gather requirements for Living Web Archives, namely:

- Requirements from the intended target groups of the LiWA services, namely library and archiving organizations towards the LiWA services and the LiWA applications
- Technical requirements for component functionality and component interaction as well as for representation formats
- Requirements for the integration of the LiWA services with each other and into an overall advanced Web archiving infrastructure;

The collected requirements will be documented in a Requirement Analysis report and serve as an input for steering the technology development as well as the dissemination activities in the LiWA project.

Task 1.1 Preparation of requirement analysis (BeG, EA) → D1.1

Activities:
- Define Methodology for User Centered Design in internal deliverable
- Creation of the LiWA user group

Data gathering techniques included desk research, a quantitative questionnaire, focus groups and interviews with experts. As part of this activity, LiWA created a user community comprised of three different types of institutions: Active Web Archiving Institutions, Short-Term Engaging Institutions, and Long-Term Engaging Institutions. These users will be contacted in consecutive stages throughout the project to ensure LiWA will meet their requirements. 19 institutions decided to join the core user group (including the three LiWA partners).

Task 1.2 Analysis of target group requirements for the LiWA applications (BeG, EA, NLP, MTA)

---

3 To improve readability, MTA is used as shortcut for MTA SZTAKI throughout workpackage descriptions
The activities have been structured in three areas. They represent the different angles that WP1 has applied while collecting the requirements, namely:

- The user angle. The desirable usage of web archives by libraries and archives.
- The technical angle. The functional requirements necessary to meet:
  - the user requirements of libraries and archives and
  - the intention to extend the current state-of-the-art in/of web archiving.
- The architecture angle. The functional requirements necessary to integrate both LiWA services into one advanced web archiving infrastructure.

**Task 1.3 Collection of technical requirements (BeG, EA, LUH, HANZO) [⇒ D1.1]**

Significant progress has been made by collaboratively developing tools and standards to enable organizations to engage with web archiving in the last 5 years. However, four main limits have been reached using current state of the art technology. The deliverable lists the level of urgency of these issues: [1] Completeness of capture [2] Web spam. [3] Archive temporal coherence. [4] Archive Interpretability.

As a further step in defining the work, the deliverable presents the use cases for the LiWA research areas and the two LiWA applications. Deriving requirements from real-life use-cases enables us to position the research activity in the context of practitioner’s needs and steer the entire project in this direction.

The deliverable concludes with a description of the system's functionalities using a bottom-up approach. Considering the modular architecture that we expect for the LiWA system, the global set of functional requirements could has been organised in two distinct categories describing:

- The main functionalities provided by the LiWA components and
- The integration of these components into an integrated web archiving infrastructure.

**Deliverables**
- D1.1 Requirement Analysis Report “Living Web archive” delivered in Month 5.

**Deviations from Annex I and impact**
There were no major deviations from Annex I in WP1.

**Use of resources**
BeG’s overall person months reported for this work package are slightly higher than the planned person-months (planned PM: 7, actual PM: 8.5). This was due to some additional efforts required in setting up the user group, interviewing experts instead of using online questionnaires as originally envisioned and because the technical architecture (section 4 of the deliverable) required a substantial efforts on telephone conferences and dialogue between partners. As a result, the overall number of person months expended for this work package was slightly higher than that budgeted in Annex I, but entailed no significant deviation from the global budget.
All other resources were used in accordance with the budget. The person months incurred per partner are shown in the above table.

**Corrective action**
No corrective action necessary.
WP 2: Capture of rich and complex web content

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 2</th>
<th>Start date:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Capture of rich and complex web content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH EA MPG MTA BeG HANZO NLP MZK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planned PM per beneficiary (Year 1-3)</td>
<td>25 12 16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expended person-months per beneficiary (Year 1)</td>
<td>2.4 0 10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
The objective of this WP is to improve dramatically the fidelity of Web archives by enabling capture of content defeating current Web capture tools. This comprises the ability to find links to resources regardless of the encoding using virtual browsing, the detection and capture of structural hidden Web and the capacity to handle streaming protocols to capture rich media Web sites.

Task 2.1 Implement a interpretation/execution-based link extractor for complex and dynamic objects (EA, HANZO) ➔ D2.1, D6.5, D6.6, D6.9, D6.10

Potential Javascript rendering engines for tasks were identified and tested and three of these engines were compared. (SpiderMonkey, SEE and Webkit). The comparison lead to select "WebKit" to implement WP2 as WebKit offers a huge number of features:

- Active community
- JavaScript getters and setters
- DOM class prototypes
- Improved support for HTML editing
- Significant JavaScript speed improvements
- Support of new CSS3 properties

Extensive comparison of various open source JavaScript engines, using a baseline, was conducted. The First test with the WebKit, Javascript engine was conducted. DOM manipulation issues were analysed in depth to develop better links extraction (WebKit, Qt4). The team tested and performed various strategies to manipulate DOM from Webkit. Extensive technical meetings on how to enhance links extraction were held with the aim of building a customized version of WebKit for the special use of link extraction. The result was an improved Demonstrator tool showing click based link extraction.

Task 2.2 Hidden Web gather prototype (EA, HANZO) ➔ D2.1, D6.5, D6.6, D6.9, D6.10

Preliminary work by the team to allow crawler to work with html forms that are pre-requisite to the research of the WP on hidden web has commenced in the fourth quarter.

Task 2.3 Rich media capture module (EA, BeG) ➔ D6.5, D6.6, D6.9, D6.10

We identified and tested possible open source modules and helper application to support AV application. The Mplayer was selected as the basis of the helper tool. We proceeded to integrate
external software to Heritrix to analyse and download stream. In order to develop an improved rich media capture module we de-coupled the crawler from identification and retrieval of streams and then moved to a distributed architecture where crawlers communicated with stream harvesters through messages. The extended and enhanced capture technology for Web archives was demonstrated.

**Task 2.4 Adversarial gather management module (EA, BeG, HANZO) [➔ D6.5, D6.9, D6.10]**
Not yet started; not due in Year 1.

**Deliverables**
- **D2.1: Enhanced Capture Technology V1** delivery of the extended capturing technology for Web archives developed in the first year of the project. Delivered in Month 12.

**Contributions to deliverables**
- **D6.2**: The components and services developed in this WP will be integrated into the Integration testing test-bed. Delivered in Month 10.

**Deviations from Annex I and impact**
There were no major deviations from Annex I in WP2.

**Use of resources**
All resources were used in accordance with the budget for Year 1. The person months incurred per partner are shown in the above table.

**Corrective action**
No corrective action necessary.
WP 3: Data Cleansing and Noise Filtering

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 3</th>
<th>Start date:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Data Cleansing and Noise Filtering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH EA MPG MTA BeG HANZO NLP MZK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planned PM per Beneficiary (Year 1-3)</td>
<td>15 4 36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expended PM per Beneficiary (Year 1)</td>
<td>4.6 1.5 21.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
The main objective of Year 1 is to build the baseline technology for distinguishing useful, high-quality Web content from spam content. The special needs for Web archives need to be analyzed in terms of maintaining the quality of periodic recrawls as well as collaboration and filtering model transformation across different archives and domains. Year 1 provides baseline measures in terms of spam classification accuracy (AUC and F measures).

Task 3.1: Evaluation methodology and test bed contribution (MTA, LUH, MPG) ➔ D6.2
An invaluable data set of 13 UK snapshots provided by the Laboratory for Web Algorithmics of the Università degli studi di Milano supported from DSI-DELIS project was processed. The LiWA test bed consists of 500GB data in the latest WARC version with an additional more than 10,000 UKWEBSPAM2007 labels that proved to be useful. The data set took 2 weeks to recompile over a server in Milan and download over the Internet. Data together with the Task 3.2 and 3.3 V1 features reside on a disk provided for WP3 partners as well as the WP6 test bed. As part of the test bed, the mockup for the assessment interface that allows expert judging to be entered was completed.

Task 3.2: Content analysis (MTA, LUH) [ ➔ D3.1, D6.5, D6.6, D6.9, D6.10]
The V1 filtering implementation was completed and Web Spam Challenge state-of-the-art was reached. In Year 1, we improved on the state of the art in two areas. First we used dimensionality reduction by Latent Dirichlet Allocation, a method that improves the AUC error measure of spam classification by 2-5%. Secondly, we defined new features based on the time series of host content properties and changes over the 13.uk snapshots, a completely new area as, to our best knowledge, this type of a data has not been available for spam researchers before. Progress on cloaking detection were started and the results will appear in the LiWA Integrated Prototype V2.
Work has begun on filtering out non-spam noise. The Results will appear in LiWA Integrated Prototype V2 in the following areas:
- Duplicate detection
- Identification of real changes
- Site structure, template analysis
Task 3.3: Hyperlink analysis (MTA, MPG, LUH) [→ D3.1, D6.5, D6.6, D6.9, D6.10]
The V1 filtering implementation was completed and Web Spam Challenge state-of-the-art reached. As in Task 3.2, the use of the time series of the link features over the 13 .uk snapshots and our new results on stacked graphical classification significantly improve the state-of-the-art in this area.

Task 3.4: Blogosphere (MTA, LUH) [→ D6.5, D6.9, D6.10]
Progress has started on this task and the results will appear in LiWA Integrated Prototype V2.

Task 3.5: Crawl time plug-in development (MTA, LUH) [→ D6.6, D6.9]
The Interface design was presented in D6.1 Integration Strategy. Task 3.2 and 3.3 codes were implemented to meet the interface requirements. The integration will be presented as part of LiWA Integrated Prototype in Year 2.

Deliverables
- D3.1: Archive Filtering Technology V1 delivered in Month 12.

Contributions to deliverables
- D6.2, D6.6, D6.9: The components and services developed in this WP were integrated into the Integration testing test-bed. D6.2 was delivered in Month 10.

Deviations from Annex I and impact
There were no major deviations from Annex I in WP2.

Use of resources
All costs were used in accordance with the budget for Year 1. There is a minor deviation in the person months expended by MTA STZAKI which are higher than originally anticipated. This is due to the fact that several junior researchers were working on the project in the first year. This resulted in the need for additional efforts, however at no additional expense as the salary rate for junior researchers is lower. All other efforts were incurred as planned. The person months incurred per partner are shown in the above table.

Corrective action
No corrective action necessary.
WP 4: Archive Coherence

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 4</th>
<th>Start date or starting event:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Archive Coherence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH EA MPG MTA BeG HANZO NLP MZK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3)</td>
<td>6 6 24 9 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expended PM per beneficiary (Year 1)</td>
<td>1 1 8.5 1.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
In order 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, one 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. Hence, the aim of temporal coherence is to – given a certain timepoint or interval – capture Web sites as “authentically” as possible. In order to ensure an “as of timepoint x (or interval [x; y])” capture of a Web site(s), research on temporal coherent capturing of web contents therefore pursues the following (sub-)goals:

- **Proper dating** of web contents (Task 4.1)
- Ensure **coherence of captures** with respect to a **timepoint or interval** (Task 4.2 & 4.3)
- **Identification of contents violating coherence** (Task 4.2 & 4.3)

In order to be able to test our development with real life data we have set up a measuring environment, which is used to validate findings discovered in a simulation environment and subsequently visualize them in a user-friendly analysis environment. Research within these environments takes places along the aforementioned three intertwined threads, which will be described in the following.

**Task 4.1: Dating Methods (MPG, LUH, EA) ➔ D4.1, D6.5, D6.6, D6.9, D6.10**

**Proper dating technologies** are required to know how fresh a Web page is – that means – what is the date (and time) of last modification. 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 requires 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. Due to cost and efficiency reasons we pursue a potentially multistage change measurement procedure:
1) Check HTTP timestamp. If it is present and is trustworthy, stop here.
2) Check content timestamp. If it is present and is trustworthy, stop here.
3) Compare a hash of the page with previously downloaded hash.
4) Elimination of non-significant differences (ads, fortunes, request timestamp):
   - only hash text content, or “useful” text content
   - compare distribution of n-grams (shingling)
   - or even compute edit distance with previous version.

On the basis on these dating technologies we have been able to develop coherence improving capturing strategies that allow us to reconcile temporal information across multiple captures and/or multiple archives.

Task 4.2: Cross-Crawl Time Reconciliation (MPG, EA, MTA, LUH) ➔ D4.1, D6.5, D6.6, D6.9, D6.10

Coherence of captures with respect to a timepoint or interval

The development of coherence improving capturing strategies starts with the development of proper coherence definitions. These definitions become necessary as depending on the applied dating technology different measures come into play. From a technical point of view, an in-depth analysis of state-of-the-art crawler technology (online scenario) and existing WARC files (weekly UK-gov-captures) provided by EA served as a starting point to identify existing (in-)coherences in current archives (offline scenario). For this purpose, a measuring environment and a simulation environment have been developed that serve as a means for validation of the underlying theoretical model.

In the online scenario, first, the current implementation of existing crawlers was investigated and employed for experimental tests on proper dating. Within the measuring environment experiments with real life data are undertaken. Here, techniques for (meta-)data extraction of web contents have been implemented and the correctness of these methods has been tested. As mentioned in the previous section, particularly the reliability of Last-Modified turned out to be poor. Hence, we separated the capturing process into an initial capture in order to download the contents and to validate them in a subsequent revisit. To this end, we developed an efficient revisit strategy that allows testing for content changes right after the capture has completed. To this end, we apply conditional GETs that make use of the contents’ etags. As a result, the subsequent validation phase becomes faster by simultaneously reducing bandwidth as well as server load. Technically, this environment is subdivided into a modified version of the Heritrix crawler (including a LiWA coherence processor V1) and its associated (Oracle) database. Here, (meta-)data extracted within the modified Heritrix crawler are stored and made accessible as distinct capture-revisit tuples. In addition, arbitrary captures can be combined as artificial capture-revisit tuples of “virtually” decelerated captures. In parallel, we created a simulation environment that employs the same algorithms we have developed in the measuring environment, but gives us full control over the content (changes) and allows us to perform extreme tests (in terms of change frequency, crawling speed and/or crawling strategy). Thus, experiments employing our coherence ensuring crawling algorithms can be carried out with different expectations about the status of web contents and can be compared against ground truth.

In the offline scenario, existing WARC files (weekly UK-gov-captures) provided by EA have been investigated. In order to better understand the amount of change between two consecutive captures visualisations of change ratios have been undertaken. The so discovered change ratios are also an important parameter for our simulation environment, as it helps us to resemble real life crawling conditions. In addition, we have studied shingling techniques, which reduce the impact
of minimal changes in web pages (such as change of date only) on the overall comparison process.

In both scenarios, research takes place in cooperation with LUH and EA (integration) as well as LUH and MTA (coherence analysis for web spam detection).

**Identification of contents violating coherence**

The **analysis of coherence defects** measures the quality of a capture either directly at runtime (**online scenario**) or between two captures (**offline scenario**). To this end, we have developed methods for automatically generating sophisticated statistics per capture (e.g. number of defects occurred sorted by defect type) as part of our **analysis environment**. In addition, the capturing process is traced and enhanced with statistical data for exports in graphML. Hence, it is also possible to layout a capture’s spanning tree and visualize its coherence defects by applying graphML compliant software. This visual metaphor is intended as an additional means to automated statistics for understanding the problems that occurred during capturing. Figure 1 depicts a sample visualization of an mpi-inf.mpg.de domain capture (about 65.000 web contents) with the visone software (cf. [http://visone.info/](http://visone.info/) for details). Depending on the nodes’ size, shape, and color the user gets an immediate overview on the success or failure of the capturing process. In particular, a node’s size is proportional to the amount of coherent web contents contained in its sub-tree. In the same sense, a node’s color highlights its “coherence status”. While green stands for coherence, the signal colors yellow and red indicated (content modifications and/or link structure changes). The most serious defect class of missing contents is colored in black. Finally, a node’s shape indicates its MIME type ranging from circles (HTML contents), hexagons (multimedia contents), rounded rectangles (Flash or similar), squares (PDF contents and other binaries) to triangles (DNS lookups). Altogether, the **analysis environment** aims at helping the crawl engineer to better understand the nature of change(s) within or between web sites and – consequently – to adapt the crawling strategy/frequency for future captures. As a result, this will also help increase the overall archive’s coherence.
Figure 1: Coherence defect visualization of a sample domain capture (mpi-inf.mpg.de) by visone

Task 4.3: Crawl Cost Optimization for Temporal Coherence (MPG, EA, MTA) ➞ D6.5, D6.6, D6.9, D6.10
See description of task 4.2.

Task 4.4: Temporal References and Mappings (MPG, BeG) ➞ D6.5, D6.6, D6.9, D6.10
Not started; not due in Year 1.

Task 4.5: Cross-Archive Coherence (MPG, EA) ➞ D6.5, D6.9, D6.10
Not started; not due in Year 1

Deliverables Year 1
- D4.1 Coherence Analysis Technology V1 delivered in Month 12.
Contributions to deliverables
- **D6.2, D6.6, D6.9**: The components and services developed in this WP will be integrated into the Integration testing test-bed. D6.2 delivered in Month 10.

Deviations from Annex I and impact
There were no major deviations from Annex I in WP4.

Use of resources
All resources were used in accordance with the budget for Year 1. The person months incurred per partner are shown in the above table.

Corrective action
No corrective action necessary.
WP 5: Handling Semantic Evolution in Web Archiving

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 5</th>
<th>Start date or starting event:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Handling Semantic Evolution in Web Archiving</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH</td>
<td>EA</td>
<td>MPG</td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3):</td>
<td>36</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>Expended PM per beneficiary (Year 1)</td>
<td>18.6</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
It is the goal of this workpackage to develop methods and services for the detection of semantic and terminology evolution within Web archives, for the adequate representation of this information, and for preparing archives for dealing with the effects of semantic evolution. The methods and services aim at improved long-term archive interpretability by making semantic evolution, which manifests in changes in (domain) terminology and conceptualization, explicit.

The objectives of the first project year include the following goals:
- Clear understanding of the terminology evolution problem
- Development of a terminology evolution representation format
- Identification of related work in the related research areas
- Implementation of the terminology extraction technology
- Testing of co-occurrence and clustering methods for the sense disambiguation
- Initial work on the Semantic Terminology Browser
- Identification of suitable test collection

Task 5.1 Semantic Evolution Representation Formats (LUH, EA, MPG) ➔ D5.1
Within this task the problem of terminology evolution has been analyzed in detail. Several cases and reasons for evolutions like neologisms (words borrowed from other languages), political correctness, brandings, or political and geographical name changes have been identified. Afterwards an abstract model has been developed that allows the representation of terminology snapshots at different times (term-concept-graphs). In addition the model allows merging of term-concept-graphs while keeping time information. Besides the model for representing terminology evolution, necessary main functions have also been specified: representation of term concept relations, fusion of terminology graphs and mapping of concepts to terms. For all these main functionalities the related work has been studied. The model has been published and presented during the International Web Archiving Workshop in Aarhus, September 2009 [1].

The result forms the basis for the detection of terminology evolution in task 5.4.

For the seamless communication between the different processing steps several data formats for the term dictionary, co-occurrence matrices, term-concept-graphs, etc. has been specified in an initial version and contributes to the deliverable D5.1.

Task 5.2 Contribution to test bed (LUH, EA) ➔ D6.2
A critical aspect for the development of terminology evolution technologies is the availability of
suitable test cases and test collections. An important requirement for test collection is that they cover a longer time range in which terminology evolution occurred. As no reference collections with documented results exist, several digital collections like newspaper archives, blogs, web archives have been analysed. For the internal test beds at LUH and MPG collections from the London Times Archive and the Süddeutsche Zeitung will be used as a starting point. The London Times Archive is a unique test collection as it covers all its publications from 1785 till 1985. The Süddeutsche Zeitung covers the years 1994 through 2006.

For the integration test bed (D6.2) the requirements for suitable web pages has been specified and a first set of web pages are integrated for testing the UIMA framework and extraction components. This test collection will be extended as soon as more experiences with the terminology evolution analysis technology exist.

In addition to the integration test bed test data also the REST interfaces for the integration into the Heritrix and other Web crawlers has been specified and implemented.

Task 5.3 Methods for Terminology Extraction (LUH, MPG)   D5.1, D6.5, D6.6, D6.10
The analysis of documents for detecting terminology evolution is a complex process composed from many individual processing steps. Many tools in the areas of Natural Language Processing, Extraction, Sense Disambiguation/Discrimination, Clustering, etc. with different properties already exist. The Apache UIMA framework is the basis for the integration of the existing and newly developed tools, which allow a flexible testing of approaches as well as an efficient processing of web archive documents in a later stage. The framework as configured and installed in this task allows the extracts of terms from a variety of formats like HTML or PDF.

The UIMA based framework with all components for language processing and extraction are contributing to the deliverable D5.1. A detailed documentation will follow in D6.5.

Task 5.4 Methods for Detecting Semantic Evolution (LUH, MPG)   D6.5, D6.6, D6.9, D6.10
Initial work has been done in the detection of terminology evolution. The current approach is based on co-occurrence analysis, clustering of terms and statistics. The assumption is that terms that co-occur within a window very often have a semantic relationship. However many approaches for Sense Detection / Sense Discrimination already exists and the best approach for our problem still needs to be identified or, if necessary, be developed.

Task 5.5 Refinement of the semantic evolution detection methods (LUH, MPG)   D6.9, D6.10
Not started; not due in Year 1.

Task 5.6 Semantic Terminology Evolution Browser (LUH)   D6.5, D6.6, D6.9, D6.10
For visualizing the results in task 5.4 an initial visualization tools for term co-occurrences and clusters has been developed and can be used within a web browser. This component will later be integrated into the Web applications (see WP 7 and WP 8).

Task 5.7 Experiments with Terminology based archive search (LUH, EA)   D6.10
Not started; not due in Year 1.

Deliverables
- **D5.1 Terminology Extraction Technology V1** delivered in Month 12. Delivery of the terminology extraction technology for Web archives developed in the first year of the project. The deliverable D5.1 includes the UIMA framework as the common baseline for the development of terminology evolution technology. Within UIMA several components for accessing different document formats, annotating documents with Part of Speech and extracting nouns and noun phrases. The code has been included into the test bed (s. D6.2) together with initial JUnit tests. A detailed documentation of the technology will follow in D6.5.

Contributions to deliverables
- **D6.2, D6.6, D6.9**: The components and services developed in this WP will be integrated into the Integration testing test-bed. D6.2 was delivered in Month 10.
- In the first project year WP 5 contributed to D6.2 by defining the public REST API for the integration of the semantic analysis components into the Heritrix workflow. Furthermore test data requirements have been specified. The implementation work of D5.1 has been uploaded into the LiWA code repository and integrated into the test bed.

Deviations from Annex I and impact
There were no major deviations from Annex I in WP5.

Use of resources
All resources were used in accordance with the budget for Year 1. The person months incurred per partner are shown in the above table.

Corrective action
No corrective action necessary.
WP 6: Architecture development and Integration Activities "Living Web archive"

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start date or</td>
<td>Month 1</td>
</tr>
<tr>
<td>starting event:</td>
<td></td>
</tr>
<tr>
<td>Work package title</td>
<td>Architecture and Integration Activities &quot;Living Web archive&quot;</td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH EA MPG MTA BeG HANZO NLP MZK</td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3)</td>
<td>18 28 4 7 6 12</td>
</tr>
<tr>
<td>Expended PM per beneficiary (Year 1)</td>
<td>3.3 7.45 2 1.35 0 4.5</td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
The goal of this workpackage is to integrate the methods developed in the project into existing Web archiving infrastructures. Therefore, the objectives of the first project year include the following goals:

- Setup of the LiWA testbed based on a subset of the IIPC framework
- Development of testing strategy and test cases
- Integration of LiWA methods into the infrastructure of European Archive
- Documentation of the integration experiences

Task 6.1: Integration strategy development (EA, HANZO, LUH) ➔ D6.1, D6.3, D6.5
In the first period, the WP6 team initiated a preliminary work on LiWA's architectural design by organizing several conference calls and workpackage meetings to define the integration strategy. The technical meetings between EA and HANZO reviewed a scalable architecture for loosely coupled services integration between LiWA partners, based on the REST design pattern. Initial planning for LiWA instances of EA and HANZO Enterprise infrastructures has also been discussed. The achievements of this work have been materialized into the first WP6 deliverable D6.1: Report on Integration Strategy, Testing Plan and Testbed Architecture.

LUH coordinated the production of the deliverable. The content has been identified and the structure of the document has been developed and revised. LUH contributed to the discussion about the overall architecture. Furthermore, test cases and API for the integration of the Semantic Evolution Detection technology from Wp5 has been developed. General testing methodologies have also been defined.

EA and HANZO contributed to the discussion about the integration strategy (based on REST protocol), to the redaction of the deliverable and to the review of the overall architecture.

EA participated to the technical meetings on the content of the deliverable and contributed to its writing and reviewing. EA also gave a presentation of the LiWA architecture at the IWAW workshop, as well as a brief overview on the Wp6 activity to the IIPC members, during the project meeting at Aarhus - Denmark.

HANZO contributed to the preparation of the deliverable and reviewed the design of the integration strategy.

MPG contributed to Deliverable D6.1, particularly with respect to incorporating results achieved in Wp4 into the overall architecture. In this aspect, specification of interfaces, Coherence Analyzer functionality, and test cases have been undertaken.
Task 6.2: Integration Testing (EA, LUH) \( \Rightarrow \) D6.2, D6.3

EA set up the testbed framework on a dedicated machine at European Archive, including an Apache Web server, Heritrix crawler, Oracle and MySQL database servers, Cruise Control system, and the crawl storage access tools, fulfilling the requirements of the deliverable **D6.2: Integration Testing Testbed (Prototype)**.

LUH set up the SVN based repository as part of the testbed, as defined in D6.1. Initial integration steps of all workpackages’ sources have been performed. LUH contributed to the ongoing discussion about crawling metadata.

The experience of this work has been documented into the second written deliverable of WP6 – **D6.3: Integrated Prototypes Progress Report V1**. EA was in charge of writing the deliverable, which includes also the description of the LiWA testbed framework.

HANZO contributed to the development of strategy for the testbed architecture; analysis/appraisal of existing frameworks and systems.

MPG contributed to Deliverable D6.3 and to the integration of results achieved in Wp4 into the European Archive architecture. In this aspect, the LiWA **Temporal Coherence Processor V1.0 (D4.1)** has been integrated into the project’s Heritrix crawler. The specification of performance indicators for the report on “Advances in Web Archiving Technologies” (**Deliverable D6.5**) has also been initiated.

Task 6.3: Integration of selected methods and services for the European Archive infrastructure (EA, LUH, MPG) \( \Rightarrow \) D6.6, D6.7, D6.8, D6.9, D6.10, D6.11

EA started the work on this task with preliminary integration and tests of the Rich Media Capture module from Wp2 (streaming capture). This module was implemented as an Heritrix plug-in for European Archive infrastructure. The work on integrating the LiWA **Temporal Coherence Processor** (database setup and configuration) was also initiated, together with MPG.

Task 6.4: Integration of selected methods and services for the HanzoWeb infrastructure (HANZO, LUH, MPG) \( \Rightarrow \) D6.9, D6.7, D6.8, D6.10, D6.11

Not started; not due in Year 1.

Deliverables

- **D6.2: Integration testing Test-bed** delivered in Month 10.
- **D6.3 Integrated Prototypes Progress Report V1** delivered in Month 12.

Contributions to

- **D10.3 “Joint development roadmap with IIPC”** delivered in Month 12.

Deviations from Annex I and impact

There were no major deviations from Annex I in WP6.
Use of resources
All resources were used in accordance with the budget for Year 1. The person months incurred per partner are shown in the above table.

Corrective action (if applicable)
No corrective action necessary.
WP 7: LiWA Application “Streaming”

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 7</th>
<th>Start date or starting event:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>LiWA Application “Streaming”</td>
<td>Activity type</td>
<td>RTD</td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
<td>Participant short name</td>
<td>LUH EA MPG MTA BeG HANZO NLP MZK</td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3)</td>
<td>16 7 7</td>
<td>Expended person-months per beneficiary (Year 1)</td>
<td>2.4 1.4</td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
The goal of this WP is to build an application on top of the LiWA services, which showcases the advantages of the advanced Web archiving approach for library and archiving organizations that have to increasingly deal with the archiving of audiovisual content. In more detail the WP has the following objectives:

- Understand the needs for tools and services in support of archiving and library institutions and other stakeholders for managing the capturing of audio-visual content
- Ensure that the requirements of archiving streaming and other audio-visual content are comprehensively and soundly considered in creating the LiWA services
- Building an application that enables archiving organizations to archive such content as part of their Web archiving activities

Task 7.1 Design of the LiWA Application “Streaming” (BeG, MZK) ➔ D7.1, D6.4

After the delivery of D1.1. in June, Sound and Vision and MZK focussed on the creation of a mock-up (D 7.1 Mockup of LiWA Application “Streaming”) . To this end, Sound and Vision translated the functional requirements in D1.1. into a functional design.

A first sketch of the mock-up was presented to the consortium at the management board meeting in Aarhus, Denmark. In a second step, we reviewed various user interfaces as well as current research (notable http://iwaw.net/08/) to come to a mock up. This was created in plain html. The first html design was made available for testing and review in early October.

Task 7.2 Creation of a Mockup (BeG, MZK) ➔ D7.1, D6.4

All partners (notably MZK) commented on this design using a structured feedback form provided by Sound and Vision. The final version will be released in November as scheduled in the DoW.

---

From the DoW: A mockup of the Application “Streaming” will be constructed based on the design decisions of Task 7.1 and the requirements identified in WP1. The mockup will focus on showing the planned functionality and the possible user interaction of the considered application. The mockup will be demonstrated and discussed with potential users of the application and demonstrated at the Review. The collected feedback will be used to refine the design of the application.
Sound and Vision has main involvement in WP7, all other partners commented on the designs. The Mockup was finalised in November with the delivery of “D7.1 Mockup ‘Streaming Web’ Walk through”. The application is available from http://www.slidingdoors.nl/LiWA/. Sound and Vision is started to build an environment to carry out crawls and implement LiWA technology. In collaboration with NLP, work started on defining the use of metadata standards within the project.

Task 7.3: Identification and preparation of datasets (BeG, NLP, MZK) ➔ D6.4, D6.8, D6.10, D7.2
Not yet started; not due in Year 1.

Task 7.4 Implementation of Application V1 (BeG, MZK) ➔ D7.2
Not yet started; not due in Year 1.

Task 7.5 Evaluation of Application “Streaming” V1 (NLP) ➔ D6.8
Not yet started; not due in Year 1.

Task 7.6 Implementation Application “Streaming” V2 (BeG, MZK) ➔ D7.3
Not yet started; not due in Year 1.

Task 7.7 Evaluation of Application “Streaming” V2 (NLP) ➔ D6.10
Not yet started; not due in Year 1

Deliverables Year 1
- D7.1 Mockup of LiWA Application “Streaming” delivered in Month 10.

Deviations from Annex I and impact
There were no major deviations from Annex I in WP7.

Use of resources
All resources were used in accordance with the budget for Year 1. The person months incurred per partner are shown in the above table.

Corrective action
No corrective action necessary.
WP 8: LiWA Application “Social Web Archiving”

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 8</th>
<th>Start date or starting event:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>LiWA Application “Social Web Archiving”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1  2  3  4  5  6  7  8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH  EA  MPG  MTA  BeG  HANZO  NLP  MZK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3)</td>
<td>8  6  6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expended PM per beneficiary (Year 1)</td>
<td>3  1.5  0.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Progress towards objectives and details per task

Objectives
The goal of this WP is to demonstrate a dramatic improvement in archive structure and content which more accurately and evenly captures the rapidly evolving and increasingly diverse content of the social Web.

Task 8.1 Design of the LiWA Application “Social Web Archiving” (HANZO, MZK) ➔ D8.1, D6.4
The first phase of work involved initial planning and review of the specification of the creating and developing a Mock Up of Social Web Archiving (8.1) and the direction of the next stage of work. This was followed by a workshop we conducted at the Work Package leaders meeting held in Saarbrucken in May 2008. LiWA partner’s requirements and design feedback were gathered. Partners then progressed on to developing the description of Workflow Management Tool. The Description of the Workflow depicts the best methods that are possible to integrate the LiWA modules.
During the course of collation of the user requirements and design feedback (Task 8.1), we have found that a need for Web archiving workflow UI already existed across the project. We have decided to, therefore, modify our plans and work on defining a general framework for providing such a UI, by designing UI widgets that will be usable to display information provided by the LiWA modules, and where needed, action and feedback from users will be incorporated. This includes, for instance, the information provided on temporal coherence of archived sites and spam detection for quality checking purposes. Such UI widgets will also be usable for the spam assessment interface (WP3) and the semantic browser (WP5). The adversarial crawl management module (WP2) has also a potential use of such UI framework.

Task 8.2 Creation of a Mockup (HANZO, MZK) ➔ D8.1, D6.4
Work began on the revised design and specification for the Mock Up where the current description of workflow was detailed and the future needs and potential uses of proposed Liwa services was documented. It illustrates the integration of the Liwa services into the existing workflow and processes of one partner (NLP) who have long experience (8 years) in processing web archives. It demonstrates how the various inputs from the LiWA module can be used in a real case scenario, with a specific focus on complex and dynamic websites that typically constitutes the social web.

Deliverables Year 1
- D8.1 Mock-up of LiWA Application “Social Web Archiving” delivered in Month 12.

**Deviations from Annex I and impact**

D8.1 was delayed by 2 months from its original due date in Month 10. The delay was caused due to the need to amend the work plan. After the initial information gathering and review process we found that Web archiving workflow UI already existed across the project. We then had to revise, modify and review the revised plans and set about developing a general framework for providing the appropriate UI. There was no impact on any other tasks or deliverables.

**Use of resources**

**Corrective action**

Plans for WP 8 changed to include revised framework for providing web archiving workflow UI. This had an impact on nature of work on the work package and delivery of D8.1 was therefore delayed by 2 months, but has since been submitted in Month 12.
**WP 10: Dissemination and Exploitation**

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 10</th>
<th>Start date:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Dissemination and Exploitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH EA MPG MTA BeG HANZO NLP MZK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3)</td>
<td>3 8 3 3 3 2 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expended PM per beneficiary (Year 1)</td>
<td>0.6 5.2 1</td>
<td>0.6 0.35</td>
<td></td>
</tr>
</tbody>
</table>

**Progress towards objectives and details per task**

**Objectives**

The objective of WP10 is to promote the results (scientific and technological) of LiWA targeting three sets of stakeholders in digital preservation from the very active ones in Web archiving to those who are only planning to engage in this domain (Active Web Archiving Institutions, Short Term Engaging Institutions, Long Term Engaging Institutions, see section B3.2.1). In more detail the activities in this WP will pursue the following goals:

- Informing the targeted stakeholders as well as the respective scientific communities about the LiWA activities and raising awareness scientific publications as well as the creation and distribution of promotional material;
- Actively involving the targeted communities and stakeholders in digital preservation into LiWA activities by strong dissemination activities.
- Preparing the exploitation and sustainability of the LiWA results by monitoring and consulting the other LiWA activities with this goal in mind and by the preparation of an adequate exploitation plan;

Coordination with IIPC development effort will be made that will result in a coordinated roadmap for development. Where possible, LiWA will contribute open-source code in the form of contribution to IIPC projects or new modules.

**Task 10.1 Public Web site (EA)**

Before the official launch of the project in February 2008, LUH developed an initial version of the project Website (D10.1) in coordination with EA. A final version of the web site was developed to provide an easier access to the participants (publication facilities). Research topics were highlighted with the use of short videos and a topical organisation of the site interface (EA). A particular attention was given to the needs of the core user group and of the reviewers by preparing new sections and functionalities designed for them.

**Task 10.2 Dissemination plan (EA, HANZO)**

During the second quarter of the project, EA in consultation with Hanzo proposed a first draft of the “Initial dissemination and exploitation plan” (D10.2). The third quarter of the project, saw the finalisation of the D10.2 “Initial dissemination and exploitation plan” (EA in consultation with Hanzo), thanks to partners and user community input. The deliverable D10.2 was then reviewed internally by involved partners and more particularly by NLP team.
Task 10.3 Dissemination activities (All)  ➔ D10.5
Liwa partners presented the project activities in several major conferences and proposed papers on the first output of their Work Packages research results.

- All partners participated in the 8th International Web Archiving Workshop (IWAW), Aarhus, Denmark (18th-19th Sept. 2008). Partners presented LIWA WP activities and organized LiWA research sessions for the IWAW workshop to explain LiWA project and its current advancement. A particular attention was given to the users community as the IIPC members were also meeting during the ECDL 2008 conference. The concordance of the three events (ECDL, IWAW and IIPC work sessions) allowed Liwa partners (All) to present and discuss the Liwa WP activities with IIPC members and to organize a common technical meeting.
- LUH gave a presentation of the project and a lecture at the University of Stuttgart and wrote an article for the DPE Preservation bulletin.
- LUH made a contribution to the biannual newsletter and the quarterly bulletin of Digital Preservation Europe (DPE).
- The project activities were also presented at I-Know, International Conference on Knowledge Management, Austria in September 2008 and during the “Web Science” course on November 28, 2008 at RWTH Aachen University, Germany through presentations and lectures given by MPG team.
- EA team provided dissemination tools and material such as presentation templates, project reports and letterhead with the newly designed logo.
- An ad-hoc mailing list was created and each of the research WP contributed to the writing and the review of the first Liwa newsletter which was made available on the web site in January and distributed through mailing lists and personalized mailing to core user group members.

Task 10.4 Ensure results translate into a contribution to the IIPC reference Web archiving platform (EA, HANZO, LUH, MPG, MTA) ➔ D10.3
The core user group members were finalized and the organisation of a coordinated development roadmap with IIPC was set and confirmed through the writing and reviewing of the D10.3: “Coordinated development roadmap IIPC”.

Task 10.5 Web archives community platform (EA, HANZO, BeG, NLP, MZK) ➔ D10.4 9
Not started. Not due in Year 1.

Deliverables
- D10.1: Project Web site delivered in Month 1.
- D10.2: Initial dissemination and exploitation plan delivered in Month 6.
- D10.3: Coordinated development roadmap with IIPC delivery planned in Month 12.

Deviations from Annex I and impact
Preparation of the first version of D10.3 has been slightly delayed to be completed after the first IIPC-LIWA meeting to facilitate practical coordination between IIPC and LIWA.

Use of resources
All resources were used in accordance with the budget for Year 1 with a minor deviation in the design costs for the LiWA website and PR logo and templates. The design was realized via a subcontract which was not originally foreseen in Annex I, but which represents only a minor task with very little change to the global budget. The person months incurred per partner are shown in the above table.

**Corrective action**
Regarding the delay in the delivery of D10.3, further meetings have been organised with IIPC TC to complete the final version of the roadmap.
## 4. Deliverables and milestones tables

### Deliverables for Year 1 (excluding the periodic and final reports)

<table>
<thead>
<tr>
<th>Del. no.</th>
<th>Deliverable name</th>
<th>WP no.</th>
<th>Lead beneficiary</th>
<th>Nature</th>
<th>Dissemination level</th>
<th>Delivery date from Annex I</th>
<th>Delivered Yes/No</th>
<th>Actual / Forecast delivery date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>D10.1</td>
<td>Project Web site</td>
<td>10</td>
<td>2</td>
<td>O</td>
<td>PU</td>
<td>1</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1.1</td>
<td>Requirement Analysis Report “Living Web archive”</td>
<td>1</td>
<td>5</td>
<td>R</td>
<td>PU</td>
<td>5</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D10.2</td>
<td>Initial dissemination and exploitation plan</td>
<td>10</td>
<td>R</td>
<td>CO</td>
<td>6</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D6.1</td>
<td>Report on Integration Strategy, Testing Plan and Test-bed Architecture</td>
<td>6</td>
<td>1</td>
<td>R</td>
<td>CO</td>
<td>7</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D10.3</td>
<td>Coordinated development roadmap with IIPC</td>
<td>10, 6</td>
<td>2</td>
<td>R</td>
<td>PU</td>
<td>9</td>
<td>No</td>
<td>12</td>
<td>Preparation of the first version of 10.2 has been slightly delayed to be completed after the first IIPC-LIWA meeting to facilitate practical coordination between IIPC and LIWA.</td>
</tr>
<tr>
<td>D6.2</td>
<td>Integration testing Test-bed</td>
<td>6</td>
<td>1</td>
<td>P</td>
<td>PP</td>
<td>10</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D7.1</td>
<td>Mockup of LIWA Application “Streaming”</td>
<td>7</td>
<td>5</td>
<td>P</td>
<td>PU</td>
<td>10</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Del. no.</td>
<td>Deliverable name</td>
<td>WP no.</td>
<td>Lead beneficiary</td>
<td>Nature</td>
<td>Dissemination level</td>
<td>Delivery date from Annex I</td>
<td>Delivered</td>
<td>Actual / Forecast delivery date</td>
<td>Comments</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------------------------------------------</td>
<td>--------</td>
<td>------------------</td>
<td>--------</td>
<td>---------------------</td>
<td>----------------------------</td>
<td>-----------</td>
<td>---------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>D8.1</td>
<td>Mock-up of LiWA Application “Social Web Archiving”</td>
<td>8</td>
<td>6</td>
<td>P</td>
<td>PU</td>
<td>10</td>
<td>No</td>
<td>12</td>
<td>After the initial information gathering and review process we found that Web archiving workflow UI already existed across the project. We then had to revise, modify and review the revised plans and set about developing a general framework for providing the appropriate UI.</td>
</tr>
<tr>
<td>D2.1</td>
<td>Enhanced Capture Technology V1</td>
<td>2</td>
<td>6</td>
<td>P</td>
<td>RE</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D6.3</td>
<td>Integrated Prototypes Progress Report V1</td>
<td>6</td>
<td>2</td>
<td>R</td>
<td>PU</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3.1</td>
<td>Archive Filtering Technology V1</td>
<td>WP 3</td>
<td>4</td>
<td>P</td>
<td>RE</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D4.1</td>
<td>Coherence Analysis Technology V1</td>
<td>WP 4</td>
<td>3</td>
<td>P</td>
<td>RE</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D5.1</td>
<td>Terminology Extraction Technology V1</td>
<td>WP 5</td>
<td>1</td>
<td>P</td>
<td>RE</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milestone no.</td>
<td>Milestone name</td>
<td>Work package no</td>
<td>Lead beneficiary</td>
<td>Delivery date from Annex I</td>
<td>Achieved Yes/No</td>
<td>Actual / Forecast achievement date</td>
<td>Comments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-----------------------------</td>
<td>----------------</td>
<td>-----------------------------------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M9.1</td>
<td>Quality Management Plan</td>
<td>9</td>
<td>LUH</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. Project management

WP 9: Project Management

<table>
<thead>
<tr>
<th>Workpackage number</th>
<th>WP 9</th>
<th>Start date or starting event:</th>
<th>Month 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Project Management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>MGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>LUH, EA, MPG, MTA, BeG, HANZO, NLP, MZK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planned PM per beneficiary (Years 1-3)</td>
<td>12 1 1 1 0.5 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expended PM per beneficiary (Year 1)</td>
<td>3.6 0.3 0 0 0 0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Objectives
The overall objectives for WP9 are to ensure effective planning, coordination and implementation of the project activities, oversee project progress and identify and address project related opportunities.

- Specifically for Year 1, the main objectives were to establish an implement decision-making, communication and reporting structures to facilitate controlling and guide project activities.

Tasks 9.1. Planning and Scheduling (LUH, EA, MPG)
The project kick-off meeting took place in Month 1 in Hannover, with further Governing Board and WP planning meetings held in Months 4 (Saarbruecken), 6 (Paris), 9 (Aarhus) and 12 (Hilversum). Preparations were also made for the first project review to take place in March 2009.

Task 9.2. General project management activities (LUH)
This task primarily relates to the day-to-day management of the project and its oversight. A reporting schedule was created and has been effectively implemented. An internal wiki is available for project communication and scheduling and is being actively used by the consortium. Mailing lists for the easy and efficient communication with the whole consortium or parts of it have been established. Support in financial and administrative matters of the project has been continuously provided for partners as necessary and the first tranche of prefinancing was distributed in Month 1.

The project Quality Plan as milestone M9.1 has been prepared and put online to the wiki with all definitions about quality procedures to be used throughout the project. The Plan has been applied to ensure consistent quality of all deliverables, including an internal auditing and peer review process.

Regular reporting in the form of quarterly progress reports and biannual reports on expended resources was conducted within the consortium to facilitate the controlling of the project’s progress. Support was provided for all beneficiaries in the preparation of the current periodic activity and management report (D9.1 and D9.2) for Year 1.

**Deliverables Year 1**
- **D9.1 Periodic Activity Report** to be delivered in Month 12
- **D9.2 Periodic Management Report** to be delivered in Month 12

D9.1. and D9.2 will be combined in the Project Periodic Report due for submission after the end of Year 1. A draft version will be submitted for the project review in March 2009. The finalised report will be submitted by the deadline of 60 days following the end of the reporting period.

**Milestones**

The quality management plan was finalised and circulated among the consortium by Month 4 of the project.

**Deviations from Annex I and impact**

There were no major deviations from Annex I in WP9.

**Use of resources**
All resources were used in accordance with the budget for Year 1. The person months incurred per partner are shown in the above table.

**Corrective action**
No corrective action necessary

**Problems which have occurred and how they were solved or envisaged solutions;**
Not applicable.

**Changes in the consortium, if any;**
As of December 2008, Dr. Thomas Risse has replaced Dr. Claudia Niederee as project coordinator and contact person for the project. A contract amendment was requested in December 2008 accordingly.
Project planning and status

Update of Gantt Chart

Figure 1: Timing for WP 1 – WP 6
Impact of possible deviations from the planned milestones and deliverables, if any
The deviations of delayed deliverables D10.3 and D8.1 have no overall impact on the project.

Any changes to the legal status of any of the beneficiaries, in particular non-profit public bodies, secondary and higher education establishments, research organisations and SMEs
There were no changes in legal status of any of the beneficiaries in Year 1.

Development of the Project website, if applicable
Before the official launch of the project in February 2008, LUH developed an initial version of the project Website (D10.1) in coordination with EA. A final version of the web site was developed to provide an easier access to the participants (publication facilities). Research topics were highlighted with the use of short videos and a topical organisation of the site interface (EA). A particular attention was given to the needs of the core user group and of the reviewers by preparing new sections and functionalities designed for them. The URL of the web page is: http://www.liwa-project.eu/
Use of foreground and dissemination activities during this period (if applicable).

Liwa partners presented the project activities in several major conferences and proposed papers on the first output of their Work Packages research results.

- All partners participated in the 8th International Web Archiving Workshop (IWAW), Aarhus, Denmark (18th-19th Sept. 2008). Partners presented LiWA WP activities and organized LiWA research sessions for the IWAW workshop to explain LiWA project and its current advancement.
- LUH gave a presentation of the project and a lecture at the University of Stuttgart and wrote an article for the DPE Preservation bulletin.
- LUH team made a contribution to the biannual newsletter and the quarterly bulletin of Digital Preservation Europe (DPE).
  

- The project activities were also presented at I-Know, International Conference on Knowledge Management, Austria in September 2008 and during the “Web Science” course on November 28, 2008 at RWTH Aachen University, Germany through presentations and lectures given by MPG team.
- An ad-hoc mailing list was created and each of the research WP contributed to the writing and the review of the first Liwa newsletter which was made available on the web site in January and distributed through mailing lists and personalized mailing to core user group members.

Communication between partners, internal communication

Mailinglists
For easy and targeted communication among all teams within the LiWA project, a number of internal mailing lists have been set up. The liwa-all list is used to send emails to all project participants. For administrative information the liwa-admin list is used. The liwa-tech list consist mainly of technicians, researchers and people interested in the technology development. Finally the liwa-apps list is used for discussions about applications.

Internal Project Wiki
As one of the major tools for communication within the project LiWA exploits a Wiki (https://wiki.liwa-project.eu/). The main benefits of the wiki are its ease of use, flexibility in the structuring information, versioning to help track any changes. The content of the LiWA wiki covers all areas of the project.

Meetings
Meetings are another major instrument in LiWA to ensure a regular exchange of achievements and discussions about open issues. Therefore regular physical meetings were held. In the following an overview about the meetings is given. Agenda, meeting minutes, participants list and presentations can be found in the “meeting area” of the LiWA wiki.

<table>
<thead>
<tr>
<th>Meeting</th>
<th>Date</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiWA Kick-Off Meeting</td>
<td>04 - 05 February 2008</td>
<td>Hannover, Germany</td>
</tr>
</tbody>
</table>
Teleconferencing / Skype Conferences
Teleconferencing is a powerful tool for organising short meetings to complement physical meetings. They can be organised with short notice, participants only need a plain telephone set to participate and do not need to spend time travelling. Within LiWA mostly Skype is used for teleconferencing. An overview of major conference calls is given below.

<table>
<thead>
<tr>
<th>Date</th>
<th>Who</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.03.08</td>
<td>WP Leaders</td>
<td>Project Status</td>
</tr>
<tr>
<td>29.04.08</td>
<td>WP Leaders</td>
<td>Project Status</td>
</tr>
<tr>
<td>08.05.08</td>
<td>WP 1 Partners</td>
<td>WP 1 Discussion</td>
</tr>
<tr>
<td>16.05.08</td>
<td>WP 5 Partners</td>
<td>WP 5 Discussion</td>
</tr>
<tr>
<td>19.06.08</td>
<td>WP Leaders</td>
<td>Project Status</td>
</tr>
<tr>
<td>15.12.08</td>
<td>WP Leaders</td>
<td>Project Status</td>
</tr>
</tbody>
</table>

External communication

Coordination with IIPC
In the domain addressed by LiWA, web archiving, the International Internet Preservation Consortium (IIPC) is playing an important role both to organize the community of heritage institutions and to develop open source tools. LiWA has been thought as a virtual ‘R&D’ extension of this effort since the very beginning. Although LiWA has its own research agenda, we try our best to position it as useful project for the Web Archiving community and given the prominence of this IIPC in this community, this includes following and coordinating closely where possible with IIPC’s own development effort. The organization of a joint meeting in September 2008 and the preparation of a Joint Workshop in 2009 for developers reflect this willingness of both groups to work together. A first version of coordinated roadmap has been developed in D10.3.

Cooperation with other projects
Beside the coordinated activities with IIPC contacts has been established to the European project PLANETS, IMPACT and DigitalPreservationEurope (DPE). Since the beginning of LiWA in 2008 the consortium regularly published an article in the DPE publications. With IMAPCT possible cooperation in the area of text analysis has been identified, which could be useful for the detection of terminology evolution.
### TABLE 3.1 Personnel, subcontracting and other major direct cost items for Beneficiary 1 for the period

<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Personnel costs</td>
<td>9,569,53</td>
<td>Task 1.1, Task 1.2, Task 1.3</td>
</tr>
<tr>
<td>3</td>
<td>Personnel costs</td>
<td>20,212,05</td>
<td>Task 3.1, Task 3.2</td>
</tr>
<tr>
<td>4</td>
<td>Personnel costs</td>
<td>6,215,54</td>
<td>Task 4.2</td>
</tr>
<tr>
<td>5</td>
<td>Personnel costs</td>
<td>85,515,59</td>
<td>Task 5.1, Task 5.2, Task 5.3, Task 5.4, Task 5.6</td>
</tr>
<tr>
<td>6</td>
<td>Personnel costs</td>
<td>16,521,93</td>
<td>Task 6.1, Task 6.2, Task 6.3</td>
</tr>
<tr>
<td>9</td>
<td>Personnel costs (MGT)</td>
<td>19,779,87</td>
<td>Task 9.1, Task 9.2, Task 9.3, Task 9.4</td>
</tr>
<tr>
<td>10</td>
<td>Personnel costs</td>
<td>3,702,15</td>
<td>Task 10.1, Task 10.3</td>
</tr>
<tr>
<td>1, 5, 9</td>
<td>Travel &amp; Subsistence</td>
<td>1,270,95</td>
<td>LiWA Meeting Saarbrucken, 04-06.05.2008, T. Risse, W. Siberski, T. Iofciu, G. Zenz, N. Tahmasebi</td>
</tr>
<tr>
<td>10</td>
<td>Travel &amp; Subsistence</td>
<td>23,92</td>
<td>Presentation of LiWA at GI-Digital Library Meeting, Karlsruhe, T. Risse</td>
</tr>
<tr>
<td>10</td>
<td>Travel &amp; Subsistence</td>
<td>423,44</td>
<td>Presentation of LiWA at the University of Stuttgart, 26-27.06.2008, T. Risse, N. Tahmasebi</td>
</tr>
<tr>
<td>10</td>
<td>Travel &amp; Subsistence</td>
<td>670,16</td>
<td>Presentation of LiWA at the University of Stuttgart, 26-27.06.2008, T. Risse, N. Tahmasebi</td>
</tr>
<tr>
<td>1, 3, 4, 5, 6, 9</td>
<td>Travel &amp; Subsistence</td>
<td>2,658,27</td>
<td>LiWA Meeting Paris, 09-12.07.2008, T. Risse, G. Zenz, W. Nejdl, N. Tahmasebi</td>
</tr>
<tr>
<td>5, 10</td>
<td>Travel &amp; Subsistence</td>
<td>703,80</td>
<td>Presentation of LiWA at the IFIP WG 2.6 Meeting, T. Risse</td>
</tr>
<tr>
<td>1, 3, 4, 5, 6, 9</td>
<td>Travel &amp; Subsistence</td>
<td>2,990,40</td>
<td>LiWA Meeting and ECDL Conference, Aarhus, 19-19.09.2008, T. Risse, W. Siberski, N. Tahmasebi</td>
</tr>
<tr>
<td>1, 3, 4, 5, 6, 9</td>
<td>Travel &amp; Subsistence</td>
<td>109,67</td>
<td>LiWA Meeting Hilversum, 19-21.01.2009, T. Risse, G. Zenz, T. Iofciu</td>
</tr>
<tr>
<td>9</td>
<td>Travel &amp; subsistence (MGT)</td>
<td>186,95</td>
<td>Project meeting (management aspects), Hilversum, 19-20.01.2009, N. McGuinness</td>
</tr>
<tr>
<td>1, 3, 4, 5, 6, 9</td>
<td>Other costs</td>
<td>1048,60</td>
<td>LiWA Meeting Costs, Aarhus, 15-17.09.2008, meeting costs</td>
</tr>
<tr>
<td>10</td>
<td>Other costs</td>
<td>29,00</td>
<td>LiWA website registration</td>
</tr>
<tr>
<td>9</td>
<td>Other costs (MGT)</td>
<td>281,09</td>
<td>LiWA Kick-Off Meeting, Hannover, 04-05.02.2008, meeting costs</td>
</tr>
</tbody>
</table>

**TOTAL DIRECT COSTS** 172,241.01
<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Personnel costs</td>
<td>26,046,57</td>
<td>Task 1.1, Task 1.2, Task 1.3</td>
</tr>
<tr>
<td>2</td>
<td>Personnel costs</td>
<td>12,172,14</td>
<td>Task 2.1, Task 2.3</td>
</tr>
<tr>
<td>4</td>
<td>Personnel costs</td>
<td>5,704,39</td>
<td>Task 4.1, Task 4.2</td>
</tr>
<tr>
<td>6</td>
<td>Personnel costs</td>
<td>36,082,07</td>
<td>Task 6.1, Task 6.2, Task 6.3</td>
</tr>
<tr>
<td>10</td>
<td>Personnel costs</td>
<td>33,945,62</td>
<td>Task 10.1, Task 10.2, Task 10.3, Task 10.4</td>
</tr>
<tr>
<td>9</td>
<td>Personnel costs (MGT)</td>
<td>2,199,38</td>
<td>Task 9.1, Task 9.2</td>
</tr>
<tr>
<td>10</td>
<td>Subcontracting (MGT)</td>
<td>4,500,00</td>
<td>Website design, video preparation, logo and presentation templates by a designer</td>
</tr>
<tr>
<td>all</td>
<td>Travel &amp; subsistence</td>
<td></td>
<td>Attendance of Kick-off meeting in Hanover, 03-05.02.2008</td>
</tr>
<tr>
<td>1, 10</td>
<td>Travel &amp; subsistence</td>
<td></td>
<td>IIPC General Assembly in Canberra, 07-11.04.2008</td>
</tr>
<tr>
<td>1, 2, 6, 10</td>
<td>Travel &amp; subsistence</td>
<td></td>
<td>LIWA Project Meeting in Saarbrücken, 05-06.05.2008</td>
</tr>
<tr>
<td>2, 6, 10</td>
<td>Travel &amp; subsistence</td>
<td></td>
<td>LIWA Project Meeting, IIPC meeting, IWAW Workshop and LIWA Conference in Aarhus, 15-22.09.08</td>
</tr>
<tr>
<td>2, 6, 9, 10</td>
<td>Travel &amp; subsistence</td>
<td>9,266,95</td>
<td>LIWA Project Meeting in Hilversum, 19-21.01.09</td>
</tr>
<tr>
<td>9</td>
<td>Other Costs: Coordination Meeting Costs</td>
<td>2,546,56</td>
<td>Governing Board Meeting; Organisation of the LIWA Meeting in Paris, 10-11.07.2008</td>
</tr>
</tbody>
</table>

**TOTAL DIRECT COSTS** | **132,463,68**
<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Personnel costs</td>
<td>104,149,97</td>
<td>Evaluation methodology and visualization strategy for data cleansing and noise filtering based on change recognition in Web sites</td>
</tr>
<tr>
<td>4</td>
<td>Personnel costs</td>
<td></td>
<td>Dating methods and cross-crawl time reconciliation for archive coherence; definition and modelling of coherence; simulation of coherence optimizing crawling strategies; participation in work package meetings</td>
</tr>
<tr>
<td>5</td>
<td>Personnel costs</td>
<td></td>
<td>Representation formats for handling semantic evolution in Web archiving; work on experimental testbeds for terminology evolution; participation in work package meetings</td>
</tr>
<tr>
<td>6</td>
<td>Personnel costs</td>
<td></td>
<td>Specification of interfaces, coherence analyzer functionality, and test cases; integration of the LIWA temporal coherence processor V1.0 into the testbed</td>
</tr>
<tr>
<td>10</td>
<td>Personnel costs</td>
<td></td>
<td>Presentations at the intl. conf. on knowledge management &quot;I-Know&quot; (Graz, Austria) and the intl. Web archiving workshop &quot;IWAW&quot; (Arhus, Denmark); lecture on Web archiving held at RWTH Aachen in scope of the &quot;Web Science&quot; lecture series</td>
</tr>
<tr>
<td></td>
<td>Travel &amp; subsistence</td>
<td>3,737,34</td>
<td>Hannover, Paris, Graz, Aarhus (work package meetings and presentations)</td>
</tr>
<tr>
<td></td>
<td><strong>TOTAL DIRECT COSTS</strong></td>
<td><strong>107,887,31</strong></td>
<td></td>
</tr>
</tbody>
</table>
### Table 3.4 Personnel, subcontracting and other major direct cost items for Beneficiary 4 (MTA SZAKI) for the period

<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Personnel costs</td>
<td>47,247,31</td>
<td>Task 3.1, Task 3.2, Task 3.3, Task 3.4, Task 3.5</td>
</tr>
<tr>
<td>4</td>
<td>Personnel costs</td>
<td>3,641,97</td>
<td>Task 4.2</td>
</tr>
<tr>
<td>6</td>
<td>Personnel costs</td>
<td>3,585,36</td>
<td>Contribution to D6.1</td>
</tr>
<tr>
<td>3</td>
<td>Travel</td>
<td>1,068,68</td>
<td>The Future of Web Search: Beyond Text Conference (03-05.04.2008, Andorra, Spain)</td>
</tr>
<tr>
<td>3</td>
<td>Travel</td>
<td>1,602,03</td>
<td>LiWA meeting (05.-06.05.2008, Saarbrucken, Germany)</td>
</tr>
<tr>
<td>3</td>
<td>Travel</td>
<td>867,83</td>
<td>LiWA meeting (10.-11.07.2008, Paris, France)</td>
</tr>
<tr>
<td>3</td>
<td>Travel</td>
<td>824,27</td>
<td>Meeting and ECDL conference, IWAW workshop (14-18.09.2008, Aarhus, Denmark)</td>
</tr>
<tr>
<td>3</td>
<td>Travel</td>
<td>378,72</td>
<td>WP meeting (20.01.2008, Hilversum, Netherlands)</td>
</tr>
<tr>
<td>3,4,6</td>
<td>Consumables</td>
<td>522,47</td>
<td>Diverse consumables</td>
</tr>
</tbody>
</table>

**TOTAL DIRECT COSTS** 60,351,23

### Table 3.5 Personnel, subcontracting and other major direct cost items for Beneficiary 5 (BeG) for the period

<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 7</td>
<td>Personnel costs</td>
<td>64,024</td>
<td>Activity WP1 and WP7</td>
</tr>
<tr>
<td>1, 7</td>
<td>Travel &amp; subsistence</td>
<td>963,36</td>
<td>kickoff meeting 05-05-08 Hannover</td>
</tr>
<tr>
<td>1, 7</td>
<td>Travel &amp; subsistence</td>
<td>325,08</td>
<td>Meeting Saarbrücken 05 June 08</td>
</tr>
<tr>
<td>1, 7</td>
<td>Travel &amp; subsistence</td>
<td>1,127,46</td>
<td>Meeting Arhus 14.09-19-09</td>
</tr>
<tr>
<td>1, 7</td>
<td>Other</td>
<td>810</td>
<td>Meeting LiWA Hilversum 20-01-09</td>
</tr>
</tbody>
</table>

**TOTAL DIRECT COSTS** 67,249,90
<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>Personnel costs</td>
<td>71,595.43</td>
<td>D1.1 - requirement analysis; T2.1, T2.2 link detection module work, Java Script engine to D2.1, D6.; Building a customized version of WebKit especially for link extraction; Demonstrator tool showing click based link extraction.</td>
</tr>
<tr>
<td>6, 8</td>
<td>Personnel costs</td>
<td>46,307.49</td>
<td>Contribute to WP6.1, 1.2, 1.4, 1.6; integration strategy with EA; Design Mock up for social web archive, Development of strategy for test bed architecture; analysis/appraisal of existing framework and systems and participation, D.8.1: Social web archiving - work on design and specification for Mock Up, ready for review</td>
</tr>
<tr>
<td>10</td>
<td>Personnel costs</td>
<td>3,161.73</td>
<td>Participation in LiWA Joint meeting; Presentation at IWAW 08</td>
</tr>
<tr>
<td>all</td>
<td>Travel &amp; subsistence</td>
<td>2,603.52</td>
<td>Hannover meeting Feb 2008</td>
</tr>
<tr>
<td>8, 2, 6</td>
<td>Travel &amp; subsistence</td>
<td>489.61</td>
<td>LiWA meeting Paris 9-12 July 08</td>
</tr>
<tr>
<td>2, 6</td>
<td>Travel &amp; subsistence</td>
<td>1,605.96</td>
<td>LiWA meeting Saarbrucken 5-7 May 08</td>
</tr>
<tr>
<td>6</td>
<td>Travel &amp; subsistence</td>
<td>402.74</td>
<td>LiWA meeting Paris with EA -25-27 march 08</td>
</tr>
<tr>
<td>2</td>
<td>Travel &amp; subsistence</td>
<td>434.62</td>
<td>Meeting in Uk 24 April</td>
</tr>
<tr>
<td>2, 6, 8</td>
<td>Travel &amp; subsistence</td>
<td>455.98</td>
<td>Meeting in Paris with EA 15-17 June 08</td>
</tr>
<tr>
<td>2, 6</td>
<td>Travel &amp; subsistence</td>
<td>763.64</td>
<td>Meeting in Uk 24 June</td>
</tr>
<tr>
<td>2, 8</td>
<td>Travel &amp; subsistence</td>
<td>1,565.00</td>
<td>Aarhus September 2008 meeting and Demo</td>
</tr>
<tr>
<td>2, 8</td>
<td>Travel &amp; subsistence</td>
<td>357.64</td>
<td>Meeting in Hilversum January 09</td>
</tr>
<tr>
<td>2, 8</td>
<td>Travel &amp; subsistence</td>
<td>519.76</td>
<td>Aarhus September 08 meeting and Demo</td>
</tr>
</tbody>
</table>

**TOTAL DIRECT COSTS** 130,263.12
### Table 3.7 Personnel, subcontracting and other major Direct cost items for Beneficiary 7 (NLP) for the period

<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Personnel costs</td>
<td>4283.99</td>
<td>Task 1.1, 1.2, D1.1</td>
</tr>
<tr>
<td>1</td>
<td>Travel &amp; subsistence</td>
<td>421.06</td>
<td>Attendance of kick-off meeting, Hannover, 03-06.02.2008</td>
</tr>
<tr>
<td>2</td>
<td>Travel &amp; subsistence</td>
<td>619.21</td>
<td>Attendance of LIWA meeting, Paris, 9.-11.07.2008</td>
</tr>
<tr>
<td></td>
<td>Personnel costs</td>
<td>1514.16</td>
<td>Task 8.1, 8.2</td>
</tr>
<tr>
<td></td>
<td>Personnel costs</td>
<td>365.74</td>
<td>D10.2, LIWA Newsletter</td>
</tr>
<tr>
<td></td>
<td>Travel &amp; subsistence</td>
<td>1624.48</td>
<td>Attendance of LIWA meeting, Hilversum, 19-21.01.2009</td>
</tr>
<tr>
<td></td>
<td>TOTAL DIRECT COSTS</td>
<td>8.828,64</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3.8 Personnel, subcontracting and other major Direct cost items for Beneficiary 8 (MZK) for the period

<table>
<thead>
<tr>
<th>Work Package</th>
<th>Item description</th>
<th>Amount</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Personnel costs</td>
<td>4054.53</td>
<td>Analysis of target group requirements, writing Use Cases and User Requirements for the social web application, preparation of Requirement Analysis Report “Living Web archive”</td>
</tr>
<tr>
<td>1</td>
<td>Travel &amp; subsistence</td>
<td>570.26</td>
<td>Attendance of kick-off meeting, Hannover, 03-06.02.2008</td>
</tr>
<tr>
<td>1</td>
<td>NLP-MZK meeting</td>
<td>10.74</td>
<td>NLP-MZK meeting, Prague, 26.3.2008</td>
</tr>
<tr>
<td>1</td>
<td>Travel &amp; subsistence</td>
<td>58.87</td>
<td>NLP-MZK meeting, Prague, 28.-30.5.2008</td>
</tr>
<tr>
<td>1</td>
<td>Travel &amp; subsistence</td>
<td>2043.71</td>
<td>LIWA meeting, Paris, 10.-11.07.2008</td>
</tr>
<tr>
<td>7</td>
<td>Personnel costs</td>
<td>1492.31</td>
<td>Mockup for the LIWA Application “Streaming”, Identification of archiving problematic web sites for testing</td>
</tr>
<tr>
<td>7</td>
<td>Travel &amp; subsistence</td>
<td>1163.75</td>
<td>Web archiving conference/workshop, Aarhus, 15.-21. 9. 2008</td>
</tr>
<tr>
<td>7</td>
<td>Travel &amp; subsistence</td>
<td>2238.16</td>
<td>LIWA WP-Meeting, Hilversum, 20.-21. 1. 2009</td>
</tr>
<tr>
<td>8</td>
<td>Personnel costs</td>
<td>804.40</td>
<td>Description of workflow model for integrating LIWA moduls</td>
</tr>
<tr>
<td></td>
<td>TOTAL DIRECT COSTS</td>
<td>12.436,71</td>
<td></td>
</tr>
</tbody>
</table>
7. Financial statements – Form C and Summary financial report
To be prepared online.

8. Certificates

<table>
<thead>
<tr>
<th>Beneficiary</th>
<th>Organisation short name</th>
<th>Certificate on the financial statements provided?</th>
<th>Any useful comment, in particular if a certificate is not provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LUH</td>
<td>No</td>
<td>Expenditure threshold not reached</td>
</tr>
<tr>
<td>2</td>
<td>EA</td>
<td>No</td>
<td>Expenditure threshold not reached</td>
</tr>
<tr>
<td>3</td>
<td>MPG</td>
<td>No</td>
<td>Expenditure threshold not reached</td>
</tr>
<tr>
<td>4</td>
<td>MTA SZTAKI</td>
<td>No</td>
<td>Overall EC contribution not above threshold</td>
</tr>
<tr>
<td>5</td>
<td>BeG</td>
<td>No</td>
<td>Overall EC contribution not above threshold</td>
</tr>
<tr>
<td>6</td>
<td>HANZO</td>
<td>No</td>
<td>Expenditure threshold not reached</td>
</tr>
<tr>
<td>7</td>
<td>NLP</td>
<td>No</td>
<td>Overall EC contribution not above threshold</td>
</tr>
<tr>
<td>8</td>
<td>MZK</td>
<td>no</td>
<td>Overall EC contribution not above threshold</td>
</tr>
</tbody>
</table>