

Turning pure Web Page Storages into Living Web Archives

Thomas Risse
L3S Research Center
risse@L3S.de

Julien Masanès
European Archive
julien@europarchive.org

András A. Benczúr
Hungarian Academy of Sciences
benczur@sztaki.hu

Marc Spaniol
Max-Planck-Institut für Informatik
mspaniol@mpi-inf.mpg.de

Abstract

Web content plays an increasingly important role in the knowledge-based society, and the preservation and long-term accessibility of Web history has high value (e.g., for scholarly studies, market analyses, intellectual property disputes, etc.). There is strongly growing interest in its preservation by libraries and archival organizations as well as emerging industrial services. Web content characteristics (high dynamics, volatility, contributor and format variety) make adequate Web archiving a challenge.

LiWA will look beyond the pure “freezing” of Web content snapshots for a long time, transforming pure snapshot storage into a “Living” Web Archive. 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. 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.

Keywords: Web Archiving, Rich Media, Spam Detection, Crawl Coherence, Terminology Evolution

1. Introduction

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 is to create innovative methods and services for Web content capture, preservation, analysis and enrichment.

In the following section we first give an overview about the current state in Web archiving. Afterwards we will introduce in more detail the Living Web Archives project followed by an overview of the approaches to address the previously mentioned issues. Furthermore we will give an overview of the applications to be developed within the project. Finally the paper concludes and gives an outlook on the remaining project life time.

1. The Living Web Archives Project

The LiWA project, started in February 2008, brings together a consortium of highly qualified researchers (L3S Research Center, Max Planck Society, Hungary Academy of Science), archiving organizations (European Archive Foundation, Sound and Vision Foundation (NL), National Library of the Czech Republic, Moravian Library) and a commercial company (Hanzo Archives). 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 addresses 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;
- **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”.

2. LiWA Approaches

In the following sub-section we give an overview about the selected approaches in the four research areas covering the three objectives of the LiWA project. These approaches were

developed after getting a detailed understanding of the requirements and the system architecture. 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.

2.1. 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.

2.2. Data Cleansing and Noise Filtering

The ability to identify and prevent spam is a top priority issue for the search engine industry [1] but less studied by Web archivists. The apparent lack of a widespread dissemination of Web spam filtering methods in the archival community is surprising in view of the fact that, under different measurement and estimates, roughly 10% of the Web sites and 20% of the individual HTML pages constitute spam.

Spam filtering is essential in Web archives even if we acknowledge the difficulty of defining the boundary between Web spam and honest search engine optimization. Archives may have to tolerate more spam compared to search engines in order not to loose some content. Also they might want to have some representative spam either to preserve an accurate image of the Web or to provide a spam corpus for researchers. Therefore the main objective of spam cleansing in Web archives is to reduce 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.

The current LiWA solution is based on the lessons learned from the Web Spam Challenges [2]. As it has turned out, the feature set described in [3] and the bag of words representation of the site content [4] give a very strong baseline. Therefore the *LiWA baseline content*

feature set consists of the following language-independent measures: the number of pages in the host, the number of characters in the host name, in the text, title, anchor text etc; the fraction of code vs. text, the compression rate and entropy; and the rank of a page for popular queries. Within this set we use the measures for *in- and outdegree*, *reciprocity*, *assortivity*, *(truncated) PageRank*, *Trustrank* [5] and *neighborhood sizes*, together with the logarithm and other derivatives for most values. Whenever a feature refers to a page instead of the host, we select the home page as well as the maximum PageRank page of the host in addition to host-level averages and standard deviation.

In addition LiWA services intend to provide collaboration tools to share known spam hosts and features across participating archival institutions. A common interface to a central knowledge base will be built in which archive operators may label sites or pages as spam based on own experience or suggested by the spam classifier applied to the local archives.

As a major step in disseminating the special needs of Internet Archives, we propose tasks for a future Web Spam Challenge [6]. We generate new features by considering the temporal change of several crawl snapshots of the same domain [7]. In addition by the needs of collaboration across different archival institutions we also provide training labels over one top level domain and request prediction over a different domain.

2.3. 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[8]. 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. Current results make 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.

2.4. 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 [9] that allows the representation of terminology snapshots at different times (term-concept-graphs). From this we derived that the act of automatically detecting terminology evolution given a corpus can be divided into two subtasks. The first one is to automatically determine, from a large digital corpus, the senses of terms. Such a word sense discrimination module has been implemented and successfully been tested on the Times corpus that covers 200 years of news articles. Current work focuses on the second step – the detection of terminology evolution. In this step the word clusters detected in the first step are tracked over time to detect evolution and to derive mappings.

3. Applications

The LIWA Technologies can be used either at crawl-time or after completion of the crawl, integrated with existing web archiving workflow. In order to test and apply these new methods and results, an integration platform of the modules is being built both by the European Archive Foundation (using open source tools) and by Hanzo Archives.

Two applications scenarios are developed in LIWA to illustrate the possible use of these technologies in real world scenario whose scope is wider than what LiWA specifically addresses.

3.1. LiWA technology for content and context in Sound and Vision archive

The Netherlands Institute for Sound and Vision is one of the largest audio-visual archives in Europe. The cultural heritage preservation policy of the Institute implies that the AV archive should preserve the Dutch audiovisual cultural heritage. As the Internet is increasingly becoming an important source for (user generated) audiovisual cultural heritage content, Sound and Vision has a strong commitment to capture information available on the Web. More specifically, the institute is eager to capture broadcast related websites, including streaming content. However, as capturing streaming content from the web is difficult, until now only a selection of user generated video content is downloaded manually from the Internet. With the streaming content capturing technology developed in the LiWA project, Sound and Vision is able to address the capturing of Dutch cultural heritage content in a much more efficient way.

Besides being a potential provider of audiovisual content, the Web is regarded as a valuable source for gathering contextual information that relates to the collections. Context information is relevant for both documentalists, and also other users interested in a specific broadcast or a broadcasting related topic, such as journalists, teachers or researchers. Typically, these users have to use different interfaces for different sources to search these sources. Ideally, Sound and Vision provides these users with a single interface that allows searching both the digital asset management system of the AV archive (iMMix) and related web content. The LiWA application Streaming demonstrates how broadcast related potential end users could access web content. The archived content will be used as test data for the development of the Sound and Vision context data platform that specifically addresses the linking of web context to the digital asset management system of Sound and Vision.

3.2. Social web application

Social web sites typically contain highly inter-linked content and use dynamic linking, widgets and tools as well as high degree of personalisation. Capturing social web sites is extremely challenging and cannot be fully achieved using current methods and tools. Social web thus represents one of the greatest challenges in web archiving.

With the Social web application, LiWA intends to demonstrate a dramatic improvement in both archive structure and content completeness so that the rapidly evolving and increasingly diverse content of the social Web is captured more accurately and evenly. The aim of the application is to show how the LiWA technology fits in the workflow of an active Web archiving institution, by considering a real-life scenario of the National Library of the Czech Republic. The application is designed as a set of independent modules developed in LiWA as described in section 2. The modules can be readily integrated with existing Web archiving workflow management tools. A Web archiving institution can choose to deploy all of the modules or just some of them, depending on its needs and particular workflow. The application is designed as generic and can be used to enhance archiving of any type of web content, not just social web.

4. Conclusions & Outlook

In this paper we presented important issues in Web archiving and introduced the Living Web Archives project, which aim is to overcome these limitations. For research areas have been identified namely Capturing of Rich and Complex Web content, Data Cleansing and Noise Filtering, Archive Coherence and Archive Interpretability. Promising solutions have already been developed and continuously being enhanced in the second half of the project. Furthermore the presented application showcases will be implemented.

5. Acknowledgments

This work is funded by the European Commission under LiWA (IST FP7 216267).

6. References

- [1] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.
- [2] C. Castillo, K. Chellapilla, and L. Denoyer. Web spam challenge 2008. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.
- [3] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 423–430, 2007.
- [4] I. Bíró, D. Siklósi, J. Szabó, and A. A. Benczúr. Linked latent Dirichlet allocation in web spam filtering. In *AIRWeb '09: Proc. 5th int. workshop on Adversarial information retrieval on the web*, 2009.
- [5] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. of the 30th Int. Conference on Very Large Data Bases (VLDB)*, pp. 576–587, Toronto, Canada, 2004.

- [6] A. A. Benczúr, M. Erdélyi, J. Masanés, and D. Siklósi. Web spam challenge proposal for filtering in archives. In AIRWeb '09: Proc. of the 5th Int. Workshop on Adversarial Information Retrieval on the Web. ACM Press, 2009.
- [7] M. Erdélyi, A. A. Benczúr, J. Masanés, and D. Siklósi. Web spam filtering in internet archives. In Proc. of 5th the Int. Workshop on Adversarial information retrieval on the web (AIRWeb), 2009.
- [8] M. Spaniol, D. Denev, A. Mazeika, P. Senellart and G. Weikum. Data Quality in Web Archiving. In Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW 2009) in conjunction with the 18th World Wide Web Conference (WWW2009), Madrid, Spain, April 20, 2009, pp. 19-26.
- [9] N. Tahmasebi, T. Iofciu, T. Risse, C. Niederee, and W. Siberski; Terminology Evolution in Web Archiving: Open Issues; In Proc. of the 8th Int. Web Archiving Workshop 2008, Aarhus, Denmark