

Archiving the Web sites of Athens University of Economics and Business

Vassilis Plachouras, Chrysostomos Kapetis, Michalis Vazirgiannis

Athens University of Economics and Business

vplachouras@aeub.gr, mkap@aeub.gr, mvazirg@aeub.gr

Περίληψη

Η αναγκαιότητα της αρχειοθέτησης του περιεχομένου των ιστότοπων έχει αναγνωριστεί από το ξεκίνημα του Παγκόσμιου Ιστού, λόγω της ευκολίας με την οποία μπορεί να υπάρχει απώλεια της διαθέσιμης πληροφορίας. Η αρχειοθέτηση ιστοπεριεχομένου μπορεί να πραγματοποιηθεί τόσο σε παγκόσμιο επίπεδο, όσο και σε εθνικό και τοπικό επίπεδο. Η εργασία αυτή περιγράφει την πιλοτική αρχειοθέτηση ιστοπεριεχομένου από τους ιστότοπους του Οικονομικού Πανεπιστημίου Αθηνών (ΟΠΑ). Περιγράφουμε την αρχιτεκτονική του συστήματος που έχει εγκατασταθεί, καθώς και τα χαρακτηριστικά των δεδομένων που έχουμε συλλέξει σε μια περίοδο τριών μηνών. Επίσης τονίζουμε την σημαντικότητα του ρόλου της βιβλιοθήκης του ΟΠΑ, η οποία, ως ένα σύγχρονο κέντρο πληροφόρησης οφείλει να υιοθετήσει τις κατάλληλες στρατηγικές και να θεσπίσει τις αντίστοιχες πολιτικές για την αξιοποίηση του συστήματος αρχειοθέτησης, με απώτερο στόχο την θεματική αρχειοθέτηση ιστοπεριεχομένου και την δημιουργία ψηφιακών συλλογών προσαρμοσμένων στις εκάστοτε ανάγκες και απαιτήσεις. Τέλος, περιγράφουμε μελλοντικές κατευθύνσεις με στόχο τη βελτίωση των παρεχόμενων υπηρεσιών από την αρχειοθέτηση των ιστοσελίδων.

Λέξεις κλειδιά: Αρχειοθέτηση ιστοπεριεχομένου, συλλογή δεδομένων, ανάκτηση πληροφορίας

Abstract

The importance of Web archiving has been recognized since the early days of the World Wide Web, due to the ease with which online information can be lost. Web archiving efforts can take place both at a global level, as well as at national and local levels. In this work, we describe ongoing work on archiving the Web sites of the Athens University of Economics and Business (AUEB). We describe the architecture of a system we have deployed, as well as the data we have collected

during a period of three months from the Web sites of AUEB. We also discuss the important role of AUEB Library as an information hub within the university with respect to Web archiving. Indeed, by adopting appropriate strategies and policies, the Library can generate focused Web archives and digital collections on demand according to its needs. Finally, we discuss future directions of work for enhancing the services that the Web archive offers.

Keywords: Web archiving, crawling, information retrieval

1. Introduction

The Internet and the World Wide Web (Web) greatly facilitate the access to electronic information. Publishing information online has become as simple as uploading a file to a server, or adding a document to a database. In the context of academic institutes, the Web has also become a crucial tool for the dissemination of information and research output, as well as for educational purposes. It is now common that scientific articles contain references to online resources such as Web pages. However, there is no guarantee that the resources that were available at the time of writing will be available in the future too.

Indeed, online information published on the Web is ephemeral, because it is not necessarily replicated on printed material, but it is most likely only stored on hard disks of servers. Hence, changes in the content of a Web site in most of the cases result in the loss of the previous version of Web content. Hardware failures may also render the information inaccessible.

The necessity of preservation of Web resources has been identified from the early days of the Web (Masanés, 2006). The Internet Archive (Internet Archive, 2010), a nonprofit organization founded in 1996, aims to keep record of all available Web pages and their changes, as well as collections of videos and other media. Feise (2000) describes the initial design and implementation of the Wayback Machine, which enables the online browsing of archived versions of Web pages in the same way one browses the Web. Jaffe and Kirkpatrick (2009) also describe the architecture of the Internet Archive and discuss hardware requirements for a petabyte-scale Web archive.

Archiving the whole Web is a very challenging task, and it is difficult to keep up with its increasing scale. These observations have led to national initiatives for preserving the Web content related to specific countries. Different strategies are adopted for data collection, ranging from full crawls in Sweden (Arvidson *et al.*, 2002) to the selective collection of Web pages in UK (Bailey and Thompson, 2006) and Australia (Cathro *et al.*, 2001). The former approach aims to provide complete coverage of a domain taken at regular intervals. A disadvantage of this approach is that there is no indication about the changes of the Web content between the crawls and the consistency of the collected data (Spaniol *et al.*, 2009). The latter approach results in higher quality collections restricted to selected Web sites. Focused crawlers, such as THESUS (Halkidi *et al.*, 2003), can be potentially used for creating thematic collections using the selective approach in Web Archiving. In addition to

collecting Web content at the national level (Abiteboul *et al.*, 2001; Lampos *et al.*, 2004; Gomes *et al.*, 2006), Web archiving can be performed by an institute for the preservation of its public online information as well as for minimizing the risk of its reputation (University of Melbourne, 2010).

In this paper, we present an ongoing work to create a Web archive for the Web sites of Athens University of Economics and Business (AUEB). Based on open-source software, which has been developed for the purpose of Web archiving, we have deployed a system for collecting the Web pages from the Web sites of AUEB, as well as for searching the collected data. We also characterize the collected data and present an initial study of the changes of Web pages over time. Moreover, we discuss the role of the university's Library in leveraging a Web archiving system and the resulting data collections.

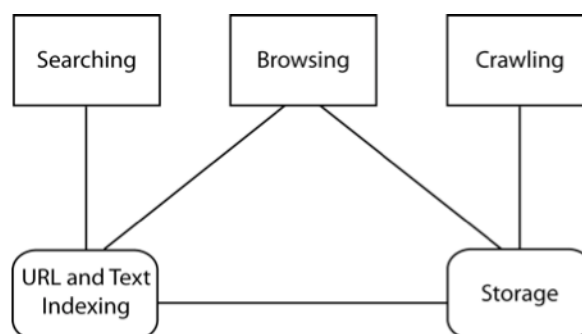
The remainder of the paper is organized as follows. Section 2 describes the system architecture and the software we have employed to implement the prototype system. In Section 3, we characterize the collected data from a period of four months (February to May 2010). In Section 4, we discuss the role of the university's library in the context of Web archiving. Finally, we close with a discussion of directions for future work and concluding remarks in Section 5.

2. System Architecture and Implementation

The design and implementation of the Web archiving prototype for the Web sites of AUEB aim to collect and preserve the Web content of AUEB. The archived content is made available to users by means of two services. The first one is URL search, which corresponds to the functionality of the Internet Archive's Wayback Machine, allowing to browse the archived versions of Web pages. The second service is full-text search in the archived Web pages, allowing to search the text of the different versions of Web pages rather than just the most recent one. Figure 1 shows the main components of the Web archiving prototype.

First, a crawler, starting from a set of seed URLs, performs the collection of the Web content and recursively downloads URLs, following their links to discover new URLs. We

Figure 1: *Components of the Web archiving prototype.*



allow the crawler to fetch any file that is available on the Web sites of AUEB. We also download files which are linked from Web pages of AUEB, but do not belong to the domain

aueb.gr, in order to allow a better browsing experience of the archives. The crawler we employ is Heritrix, an open-source crawler developed by the Internet Archive (Mohr *et al.*, 2004).

The crawled Web content is stored compressed on hard disk in files containing several records. Due to the moderate size of the Web sites of the University, we have opted for storing the crawled Web content on a local filesystem, and keeping backups on external hard disks.

Figure 2: User interface for URL search, showing the archived versions for <http://www.aueb.gr> between February 1st and May 1st 2010.




Once the Web content is downloaded, we need to index it before making it available for searching. We build two types of indexes. The first one allows the efficient location and retrieval of the contents of a given URL in the archive. For example, this index stores for each version of the processed URLs a canonicalized version of the URL, the mime-type of the page, the corresponding HTTP status code, as well as information on the position of the URL's contents in the storage. We build and maintain this index using an open-source version of the Wayback Machine (Tofel, 2007). The second type of index allows the full-text search over the archived content (Stack, 2005) and we employ NutchWax, an adapted version of an open-source search engine for indexing and retrieving from Web archives. Figure 2 presents the user interface for the URL search, where the user can enter a URL in the right textbox of the interface and search for all the corresponding archived versions between a given range of dates. The results are presented in chronological order, and the user may click on any of the dates to browse the corresponding URL as it was on that date. Figure 3 shows a screenshot of the full-text search service. The user can enter a keyword query in the textbox on the left and search for the pages that contain the given keywords within the specified range of dates. The results are presented in a way similar to typical Web search engines.

3. Data collections

We have collected the content of the Web sites of AUEB several times since February 2010. Here, we describe four crawls we performed between February and May 2010. We initiated each of the crawls from the same set of seed pages, corresponding to top-level Web pages of Web sites in the AUEB domain, using the Heritrix crawler. We changed the configuration of the crawler in order to optimize the crawling of two Web sites, a large forum

for undergraduate computing science students, and the Web site of a department in which the crawler was trapped in following the links of a calendar. We present results for all Web sites, as well as for the Web sites that were not affected by the crawler reconfiguration.

Figure 3: User interface for text search, showing results for the query 'web mining'.



The screenshot shows the AUEB Web Archive search interface. At the top, there is a search bar with the text 'web mining' and a 'Search' button. To the right of the search bar is a 'Take me back' button. Below the search bar, there are fields for 'From date' and 'To date'. The main content area displays search results for the query 'web mining'. The first result is 'DB-NET - Research team on Data & Web Mining' with a link to 'http://www.db-net.aueb.gr/ [html] (16796 bytes) - 2010-02-19 - other versions - more from www.db-net.aueb.gr'. The second result is 'eClass του Οικονομικού Πανεπιστημίου Αθηνών | ΕΞΟΥΣΙΑ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΟΝ ΠΑΓΚΟΣΜΙΟ' with a link to 'http://eclass.aueb.gr/courses/INF131/ [html] (11298 bytes) - 2010-02-19 - other versions'. The third result is 'The Research Group of Data and WEB Mining' with a link to 'http://www.cs.aueb.gr/english/research-labs/lab-db-net.shtml [html] (20734 bytes) - 2010-02-19 - other versions - more from www.cs.aueb.gr'. The fourth result is 'Web Information Management - A.U.E.B' with a link to 'http://wim.aueb.gr/research.htm [html] (19255 bytes) - 2010-02-19 - other versions - more from wim.aueb.gr'.

3.1 Description of Crawls

Table 1 presents statistics about the crawls. For each crawl, we show the date on which it was started, the number of URLs that were not fetched, and the number of fetched URLs grouped by the HTTP status codes returned by the Web servers.

The crawler did not fetch all discovered URLs, either because of network problems, or because the corresponding robots.txt files did not allow the crawling of specific URLs. We have observed that the majority of URLs not allowed to be crawled by the robots.txt file correspond to images, Javascript code, and other files used for rendering the Web sites. The fact that we are not allowed to crawl those URLs only has an impact on the presentation of the archived Web pages during browsing the archives. For the fetched URLs, we report the HTTP status codes grouped in four categories: successful (2xx), redirections (3xx), client errors (4xx), and server errors (5xx). For example, requests for URLs that result in a 404 response code, denoting that the requested URL does not exist on the Web server, belong to the group client errors (4xx).

When we consider all crawled Web sites, we can see from Table 1 that the total number of URLs is smaller for the second crawl C2. This is explained because the server that performed the crawl crashed and the crawling stopped early. Overall, we see that for all crawled Web sites, the number of URLs that result in redirections, client and server errors, does not vary significantly across crawls C1, C3 and C4. When we consider only the Web sites for which there was no change in the crawl configuration, the obtained numbers also remain stable for all four crawls.

Regarding the types of the files crawled from the URLs with successful status codes (2xx), we have observed that the four most frequent types of files are text/html,

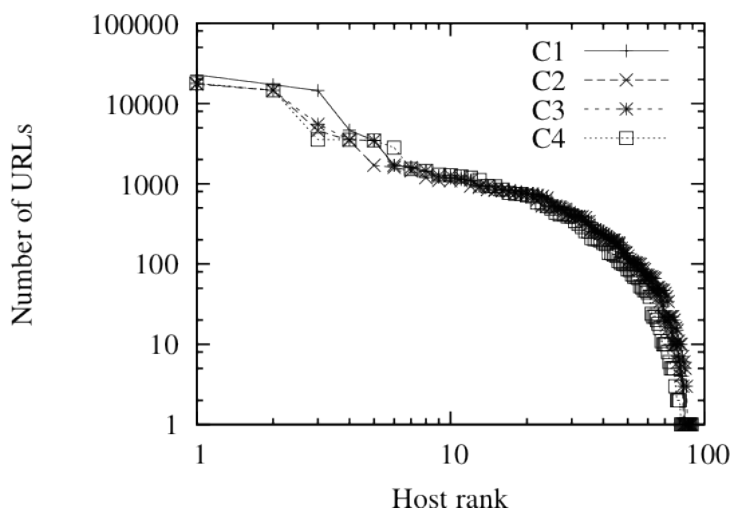
application/pdf, image/jpeg and image/gif. While their ordering varies across the four crawls, they correspond to at least 88% of the successfully crawled URLs when we consider all crawled Web sites, and over 90% when we consider Web sites for which we did not change the crawler configuration. This observation is important because it implies that we can cover most of the Web content, without requiring the management and handling of many different types of files.

Last, we look at the amount of storage required, which is a crucial factor for the sustainability of the Web archive, as it continuously grows and sets one of the most important requirements for infrastructure. The amount of data crawled from the domain aueb.gr ranges between 10GB (for C2) and 14.9GB (for C1). The data is stored on disk compressed and requires between 8.8GB (for C2) and 9.7GB (for C1), resulting in space savings of 12% for C2 and 35% for C1, respectively. In the case of the crawl C3, we only stored on disk the Web pages that did not change since the previous crawl. Consequently, we crawled 13.1GB from the domain aueb.gr, but we only stored on disk 1.6GB, resulting in space savings of 88%. We have confirmed this finding in subsequent crawls, where we only stored the content that changed. Hence, we conclude that archiving the domains of AUEB requires only moderate storage capacity and can be achieved with low cost.

3.2 Distribution of pages on hosts

Figure 4 shows in logarithmic scale the distribution of successfully fetched URLs for each different host name in the aueb.gr domain. The Y-axis corresponds to the number of URLs fetched for each host name, and the X-axis corresponds to the rank of the host name. The figure is drawn for all crawled Web sites. Overall, we can see that the distribution of the number of crawled pages per host does not change significantly across the different crawls. However, we have observed fluctuations regarding the number of URLs crawled from the largest sites.

Figure 4: Distribution of number of pages crawled per Web site.



The largest site corresponds to a forum of undergraduate students in the Department of Informatics. This site is the largest because of the large volume of old posts, as well as the presentation of the content (i.e. there are URLs for different versions of the pages, such as text-only versions of the pages, or RSS feeds of posts, or accessing the Web site using secure HTTP). Indeed, the large number of crawled pages from the site in C1 is related to the crawling of printer-friendly pages, as well as the RSS feeds related to the posts. The drop in the number of crawled URLs from this site in C2 is related to the early stop of the crawl. In subsequent crawls, we reconfigured the crawler for this particular Web site to ignore printer-friendly pages and RSS feeds.

The fluctuation in the size of the second host is related to the blocking of our crawler, which in C1 crawled a significant number of URLs by following links in a calendar appearing in most of the Web pages. Hence, we did not collect any data in C2 from this particular host. We resumed the crawling of this particular site in C3 and C4 after reconfiguring the crawler not to follow the calendar links.

A third host that presents significant fluctuation is the Web site of another department of the university, where the hosting Web server failed resulting in a loss of data. It is worth noting that in this particular case, we have been in contact with the Web site administrators, and we have provided a copy of the Web site from the Web archive. Finally, a fourth host that exhibited fluctuation in the number of URLs is the Web site of the university's Library which updated its Web site before performing crawl C4. Hence, our Web archiving system holds both the old and the updated version of the Library's Web site.

3.3 Web content changes

One of the reasons for the archiving of Web content is the fact that the Web content is updated and previous versions are lost. Hence, it is important to measure the extent to which the content of URLs changes over time. Table 2 shows statistics comparing consecutive crawls. For example, column "C1 / C2" compares the crawls C1 and C2. The table presents statistics for all Web sites, as well as for the Web sites that the crawler was not reconfigured. The statistics are the number of URLs that exist in the first and the second crawl respectively, and the number of URLs that appear in both crawls. We further break down the latter in the number of URLs that do not change across the two crawls, and the number of URLs the contents of which change. We detect changes by comparing a hash function over the contents of URLs. Hence, even a difference in a single byte results in the URL being counted as having different contents.

From Table 2, we can see that when we consider all crawled Web sites, there is a significant number of URLs that change between consecutive crawls. This is explained by the fact that the large Web forum for which the crawler was reconfigured, involved dynamic pages, which showed information related to the time of browsing, for example, the number of online users in the forum at the time the page was generated. When we consider only the Web sites for which the crawler was not reconfigured, we can see that the rate of changes in the content of URLs is lower. We have also observed that most of the changes in Web pages are due to dynamically generated HTML pages. Regarding the comparison of C1 and C2, out of the 31176 URLs that change, 94% correspond to dynamically generated Web pages.

Overall, the collection of Web pages is a challenging task, even for Web sites of moderate size. Because crawling is not instantaneous, the servers can suffer from network connectivity problems or power failures, as in the case of the crawl C2. In addition, Web sites may result in infinite number of generated URLs, hence trapping the crawler. Such issues need to be addressed in subsequent crawls by refining the configuration of the crawler, or provisioning the crawling server to recover from network and power failures.

Crawl Name	C1	C2	C3	C4
Start Date	2010-02-26	2010-03-20	2010-04-26	2010-05-13
All Web sites				
Not Fetched	2882	2684	17097	8661
2xx (Successful)	139212	96495	113811	141084
3xx (Redirection)	3200	2364	3020	3135
4xx (Client Error)	10867	8248	10141	10808
5xx (Server Error)	24	24	15	16
Excluding reconfigured Web sites				
Not Fetched	1730	1604	1786	1895
2xx (Successful)	68825	67826	67768	66497
3xx (Redirection)	3048	2241	2775	2887
4xx (Client Error)	7572	7238	7358	6504
5xx (Server Error)	24	24	15	14

Table 1: Summary statistics of the four performed crawls. The table reports the start date, number of URLs from aueb.gr, number of URLs not fetched and the distribution of HTTP Status codes.

Crawls	C1 / C2	C2 / C3	C3 / C4
All Web sites			
$u \in C_i \setminus C_{i+1}$	50747	6017	13990
$u \in C_{i+1} \setminus C_i$	8030	23333	41263
$u \in C_i \cap C_{i+1}$	88465	90578	99821
u same	56749	63865	59805
u changes	31716	26613	40016
Excluding reconfigured Web sites			
$u \in C_i \setminus C_{i+1}$	3848	3487	5347
$u \in C_{i+1} \setminus C_i$	2849	3429	4076
$u \in C_i \cap C_{i+1}$	64977	64339	62421
u same	54691	55127	53193
u changes	10286	9212	9228

Table 2: Overlap and content changes across consecutive crawls.

In addition to the crawls of the AUEB Web sites, we have also performed a crawl of Greek universities in April 2010. The crawl is not exhaustive, meaning that we stopped it before all Web pages are crawled. We have observed that the collected data follow similar trends as these collected from AUEB Web sites. In particular, we have collected 211880 and 305316 Web pages from the Web sites of the University of Athens (UOA) and National Technical University of Athens (NTUA), respectively. These numbers correspond to 90% of all requests we made to each of the domains, and they are similar to the corresponding percentage computed for the AUEB crawls (for example, 89% for C1 in Table 1). We have

also observed that the distribution of crawled mime types is similar in the case of the AUEB crawls and the UOA and NTUA crawls, where the four most frequent mime types are the same and account for more than 90% of crawled URLs. Hence, we expect that the rate of changes in crawled URLs will be similar for AUEB, UOA, and NTUA crawls, because it mainly depends on the updates in HTML pages, and in particularly the dynamically generated ones.

4. Web archiving and the Library of AUEB

The Web archiving system we have described in the previous sections gives the capability to the Library of AUEB to implement one of its basic aims, which is the archiving and preservation of digital content of the university's Web pages. A second important possibility is the integration of the Web archive to the digital repository of the Library in the form of digital collections. The most important benefits for the Library as a central information hub of the university are the following:

- Preservation of the history of the institute
- Use of historical data for future research
- Preservation and protection of cultural heritage and scientific production

Using a Web archiving system creates new prospects for digital information collections. Operating and extending a Web archiving system, the Library of AUEB aims to build, archive, and preserve digital collections with Web content selected from a set of high-quality Web sites based on topics related to the Library's own needs. To better illustrate the significance of thematic archiving of Web content, we can consider the importance of the ability to selectively collect pages from Web sites that publish on daily-basis information related to the current fiscal crisis in our country. Constructing such an archive would preserve a significant part of the information that may not be available in the future when the country exits from the current situation and recovers growth.

Creating thematic digital collections raises additional challenges to the technological ones, and it is not sufficient to only have a Web archiving system and a retrieval system to search for information in the archived content. One issue that needs to be addressed is the intellectual property of the content published on the Web sites to be archived. Archiving Web content requires the permission of its creators, and obtaining the corresponding permission may be a very time-consuming process, especially when there is no relevant legislation defining rules and processes for such cases.

An additional factor that needs to be considered is the infrastructure required for the archiving and preservation of Web content. Web archiving systems require significant storage resources, especially when the archived Web sites are very large, because archiving takes place at regular intervals and the variation of the size of a Web site may not be predictable. It is important to note that it is not required to store content that has not changed since the last time it was archived, however, Web archiving in the long term requires an increasing amount of storage space. This requirement becomes more important, because archiving a Web site

also involves storing images and other types of media that are available on the Web site in order to be able to reconstruct latter an accurate and complete snapshot of the Web site.

We believe that the most important factor for the success of such an effort is the adoption of relevant strategies and policies regarding the selection of the Web sites to be archived, and the use of standards for the creation of metadata, indexing and classification of the archived Web content as an integral part of the Library's collections.

The above discussion shows the important role of the Library for the success and the use of a Web archiving system. The success of such an effort is based on the close cooperation of librarians and computing science specialists. Librarians should have the responsibility for the selection and the classification of the content to be archived, while computing science specialists should provide the systems and the monitoring for the most efficient and uninterrupted function of archiving systems.

5. Future Work and Concluding Remarks

In this work we have described a prototype Web archiving system implemented to collect and preserve the content of the Web sites of AUEB. It represents ongoing work, which can be extended in several different ways with respect to the quality of the data collection, as well as the implementation of services in addition to URL and text searching.

Crawling is not a instant process, as already mentioned in Section 3, but it requires a period of time due to several constraints. For example, the crawler does not have infinite bandwidth to download data, and the number of requests to a server is limited in order not to overload the remote servers. In addition, a crawler trap may delay the crawler from downloading useful content by generating, intentionally or unintentionally, an infinite number of URLs that the crawler can follow. These factors have an effect on the quality and the coherence of the crawled data, because the content of URLs may change during the crawling process (Spaniol *et al.*, 2009). As discussed in Section 3.3, the majority of URLs do not change between crawls. Indeed, URLs that correspond to images, or scientific publications, are not likely to change once they are published. Consequently, the repeated crawling of these URLs does not bring new information but it confirms that the content from those URLs is still accessible. In such cases, the priority of such URLs for the crawler can be lower, resulting in lower probability of reducing the quality or the coherence of the crawl.

Another challenge that typical crawlers, and Web archiving systems, face is the acquisition of data hidden behind forms. While such data may be available for crawling when browsing is allowed by following links, this is not always the case, and hence, Web archives do not store a significant amount of information that is available currently online.

Regarding the storage requirements, we expect to achieve better compression by exploiting the fact that the contents of some URLs do not change significantly, and only storing the differences between consecutive versions, adapting techniques from indexing of multi-version documents (Berberich *et al.*, 2008; He *et al.*, 2009).

In the Web archiving propotype of AUEB, we have implemented two services, one for URL search, and a second one for full-text search. These two services, however, are not the only ones that can be implemented on top of a Web archive. For example, several

applications may exploit the historical information about Web pages in a Web archive to compute a degree of page trustworthiness, or to provide a summary of historical changes to users (Jatowt *et al.*, 2008).

Overall, Web archiving is a topic of great importance for the preservation of today's cultural and scientific online information. In this context, Web archiving fits in the role of a library as an information hub, providing access to online resources that otherwise they are likely to disappear. Our experience in archiving the Web sites of AUEB shows that it is possible to archive institutional Web content in a sustainable way using existing technology, but further improvements are expected by adopting advanced techniques for storage and retrieval.

References

- Abiteboul, S., Cobena, G., Masanes, J. and Sedrati, G. (2002). A first experience in archiving the French Web. In Proceedings of the 6th European Conference on Digital Libraries (ECDL), pages 1-15.
- Arvidson, A., Persson, K. and Mannerheim, J. (2000). The Kulturarw3 Project - The Royal Swedish Web Archiw3e - An example of "complete" collection of web pages. In Proceedings of the 66th International Federation of Libraries and Institutions (IFLA) Council and General Conference. Available from <http://archive.ifla.org/IV/ifla66/papers/154-157e.htm>. Accessed on 27/05/2010.
- Bailey, S. and Thompson, D. (2006). UKWAC, Building the UK's First Public Web Archive. *D-Lib Magazine*, 12(1).
- Berberich, K., Bedathur, S. and Weikum, G. (2008). Tunable Word-Level Index Compression for Versioned Corpora. In Proceedings of Workshop on Efficiency Issues on Information Retrieval (EIIR'08).
- Cathro, W., Webb, C. and Whiting, J. (2001). Archiving the web: The PANDORA archive at the National Library of Australia. In Preserving the Present for the Future, Proceedings of Conference on strategies for the Internet.
- Feise, J. (2000). Accessing the History of the Web: A Web Way-Back Machine. In Proceedings of the 6th International Workshop and 2nd International Workshop on Open Hypertext Systems and Structural Computing, pages 38-45.
- Gomes, D., Freitas, S. and Silva, M. (2006). Design and selection criteria for a national web archive. In Proceedings of the 10th European Conference on Digital Libraries (ECDL), pages 196-207.
- Halkidi, M., Nguyen, B., Varlamis, I., Vazirgiannis, M. (2003). THESUS: Organizing Web document collections based on link semantics. *VLDB Journal* 12(4):320-332.
- He, J., Yan, H. and Suel, T. (2009). Compact Full-Text Indexing of Versioned Document Collections. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09), pages 415-424, ACM, New York, NY, USA.
- Internet Archive (2010). Available from <http://www.archive.org>. Accessed on 25/05/2010.

- Jaffe, E. and Kirkpatrick, S. (2009). Architecture of the Internet Archive. In Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference, pages 1-10.
- Jatowt, A., Kawai, Y., Ohshima, H. and Tanaka, K. (2008). What can history tell us?: towards different models of interaction with document histories. In Proceedings of 19th ACM Conference on Hypertext and Hypermedia, pages 5-14.
- Lamos, C., Eirinaki, M., Jevtuchova, D. and Vazirgiannis, M. (2004). Archiving the Greek Web. In Proceedings of the 4th International Web Archiving Workshop (IWAW).
- Masanés, J. (editor) (2006). Web archiving. Springer Berlin Heidelberg.
- Mohr, G., Stack, M., Ranitovic, I., Avery, D. and Kimpton, M. (2004). An Introduction to Heritrix: An open source archival quality web crawler. In Proceedings of the 4th International Web Archiving Workshop (IWAW).
- Spaniol, M., Denev, D., Mazeika, A., Weikum, G. and Senellart, P. (2009). Data quality in web archiving. In Proceedings of the 3rd Workshop on Information credibility on the web, pages 19-26.
- Stack, M. (2005). Full Text Search of Web Archive Collections. In Proceedings of the International Web Archiving Workshop (IWAW).
- Tofel, B. (2007). Wayback for Accessing Web Archives. In Proceedings of the 7th International Web Archiving Workshop (IWAW).
- University of Melbourne (2010). Web archiving at the University of Melbourne. Available from <http://www.unimelb.edu.au/records/web-archiving/>. Accessed on 25/05/2010.