Design and Selection Criteria for a National Web Archive

Daniel Gomes, Sérgio Freitas, and Mário J. Silva

University of Lisbon, Faculty of Sciences
1749-016 Lisboa, Portugal

dcg@di.fc.ul.pt, sfreitas@lasige.di.fc.ul.pt, mjs@.di.fc.ul.pt

Abstract. Web archives and Digital Libraries are conceptually similar, as they both store and provide access to digital contents. The process of loading documents into a Digital Library usually requires a strong intervention from human experts. However, large collections of documents gathered from the web must be loaded without human intervention. This paper analyzes strategies to select contents for a national web archive and proposes a system architecture to support it. [1]

1 Introduction

Publishing tools, such as Blogger, enabled people with limited technical skills to become web publishers. Never before in the history of mankind so much information was published. However, it was never so ephemeral. Web documents such as news, blogs or discussion forums are valuable descriptions of our times, but most of them will not last longer than one year [21]. If we do not archive the current web contents, the future generations could witness an information gap in our days. The archival of web data is of interest beyond historical purposes. Web archives are valuable resources for research in Sociology or Natural Language Processing. Web archives could also provide evidence in judicial matters when ephemeral offensive contents are no longer available online. The archival of conventional publications has been directly managed by human experts, but this approach can not be directly adopted to the web, given its size and dynamics. We believe that web archiving must be performed with minimum human intervention. However, this is a technologically complex task. The Internet Archive collects and stores contents from the world-wide web. However, it is difficult for a single organization to archive the web exhaustively while satisfying all needs, because the web is permanently changing and many contents disappear before they can be archived. As a result, several countries are creating their own national archives to ensure the preservation of contents of historical relevance to their cultures [6].

Web archivists define boundaries of national webs as selection criteria. However, these criteria influence the coverage of their archives. In this paper, we analyze strategies for selecting contents for a national web archive and present a system's architecture

[1] This study was partially supported by FCT under grant SFRH/BD/11062/2002 (scholarship) and FCCN.

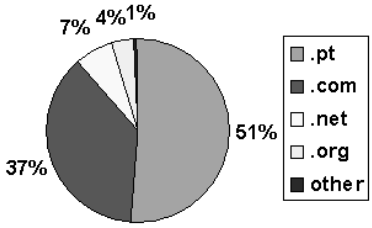

Fig. 1. Distribution of documents per domain from the Portuguese web.

to support a web archiving system. This architecture was validated through a prototype named Tomba. We loaded Tomba with 57 million documents (1.5 TB) gathered from the Portuguese web during 4 years to update the indexes of a search engine and made this information publicly available through a web interface (available at tomba.tumba.pt). The main contributions of this paper are: i) the evaluation of selection strategies to populate a web archive; ii) a system's architecture to support a web archive.

In the following Section we discuss strategies to populate a web archive. In Section 3, we present the architecture of the Tomba prototype. Section 4 presents related work and in Section 5 we conclude our study and propose future work.

2 Selecting

Web archivists define strategies to populate web archives according to the scope of their actions and the resources available. An archive can be populated with contents delivered from publishers or harvested from the web. The delivery of contents published on the web works on a voluntary basis in The Netherlands but it is a legislative requirement in Sweden [20]. However, the voluntary delivery of contents is not motivating for most publishers, because it requires additional costs without providing any immediate income. On the other hand, it is difficult to legally impose the delivery of contents published on sites hosted on foreign web servers, outside a country's jurisdiction. The absence of standard methods and file formats to support the delivery of contents is also a major drawback, because it inhibits the inclusion of delivery mechanisms in popular publishing tools. Alternatively, a web archive can be populated with contents periodically harvested from the country's web. However, defining the boundaries of a national web is not straightforward and the selection policies are controversial.

We used the Portuguese web as a case study of a national web and assumed that it was composed by the documents hosted on a site under the .PT domain or written in the Portuguese language hosted in other domains, linked from .PT [10]. We used a crawl of 10 million documents harvested from the Portuguese web in July, 2005 as baseline to compare the coverage of various selection policies.

2.1 Country code Top Level Domains

There are two main classes of top-level domains (TLD): generic (gTLDs) and country code (ccTLDs). The gTLDs were meant to be used by particular classes of organizations (e.g. COM for commercial organizations) and are administrated by several institutions world wide. The ccTLDs are delegated to designated managers, who operate them according to local policies adapted to best meet the economic, cultural, linguistic, and legal circumstances of the country. Hence, sites with a domain name under a ccTLD are strong candidates for inclusion in a web archive. However, this approach excludes the documents related to a country hosted outside the ccTLD. Figure 1 presents the distribution of documents from the Portuguese web per domain and shows that 49% of its documents are hosted outside the ccTLD .PT.

2.2 Exclude blogs

Blogs have been introduced as frequent, chronological publications of personal thoughts on the web. Although the presence of blogs is increasing, most of them are rarely seen and quickly abandoned. According to a survey, "the typical blog is written by a teenage girl who uses it twice a month to update her friends and classmates on happenings on her life" [5], which hardly matches the common requirements of a document with historical relevance. On the other hand, blogs are also used to easily publish and debate any subject, gaining popularity against traditional web sites. Blogs that describe the life of citizens from different ages, classes and cultures will be an extremely valuable resource for a description of our times [8].

We considered that a site is a blog if it contained the string "blog" on the site name and observed that 15.5% of the documents in the baseline would have been excluded from a national web archive if blogs were not archived. 67% of the blog documents were hosted under the .com domain and 33% were hosted on blogs under the .PT domain. One reason we found for this observation is that most popular blogging sites are hosted under the .COM domain, which tends to increase the number of documents from a national web hosted outside the country code TLD (Blogspot that holds 63% of the Portuguese blogs).

2.3 Physical location of web servers

The RIPE Network Management Database provides the country where an IP address was firstly allocated or assigned. One could assume that the country's web is composed by the documents hosted on servers physically located on the country. We observed that only 39.4% of the IP addresses of the baseline Portuguese web were assigned to Portugal.

2.4 Select media types

A web archive may select the types of the contents it will store depending on the resources available and the scope of the archive. For instance, one may populate a web archive exclusively with audio contents. Preservation strategies must be implemented

MIME type	avg size (KB)	%docs.
text/html	24	**61.2**%
image/jpeg	32	**22.6**%
image/gif	9	**11.4**%
application/pdf	327	1.6%
text/plain	102	0.7%
app'n/x-shockwave-flash	98	0.4%
app'n/x-tar	**1,687**	0.1%
audio/mpeg	**1,340**	0.04%
app'n/x-zip-compressed	**541**	0.1%
app'n/octet-stream	454	0.1%
other	129	1.8%

Table 1. Prevalence of media types on the Portuguese web.

according to the format of the documents. For instance, preserving documents in proprietary formats may require having to preserve also the tools to interpret them. The costs and complexity of the preservation of documents increases with the variety of media types archived and it may become unbearable. Hence, web archivists focus their efforts on the preservation of documents with a selected set of media types. Table 1 presents the coverage of selection strategies according to the selected media types. We can observe that a web archive populated only with HTML pages, JPEG and GIF images covers 95.2% of a national web.

2.5 Ignore robots exclusion mechanisms

The Robots Exclusion Protocol (REP) enables authors to define which parts of a site should not be automatically harvested by a crawler through a file named "robots.txt" [16] and the meta-tag ROBOTS indicates if a page can be indexed and the links followed [26]. Search engines present direct links to the pages containing relevant information to answer a given query. Some publishers only allow the crawl of the site's home page to force readers to navigate through several pages containing advertisements until they find the desired page, instead of finding it directly from search engine results. One may argue that archive crawlers should ignore these exclusion mechanisms to achieve the maximum coverage of the web. However, the exclusion mechanisms are also used to prevent the crawling of sites under construction and infinite contents such as online calendars [24]. Moreover, some authors create spider traps, that are sets of URLs that cause the infinite crawl of a site [15], to punish the crawlers that do not respect the exclusion mechanisms. So, ignoring the exclusion mechanisms may degrade the performance of an archive crawler.

We observed that 19.8% of the Portuguese web sites contained the "robots.txt" file but the REP forbade the crawl of just 0.3% of the URLs. 10.5% of the pages contained the ROBOTS meta-tag but only 4.3% of them forbade the indexing of the page and 5% disallowed the following of links. The obtained results suggest that ignoring

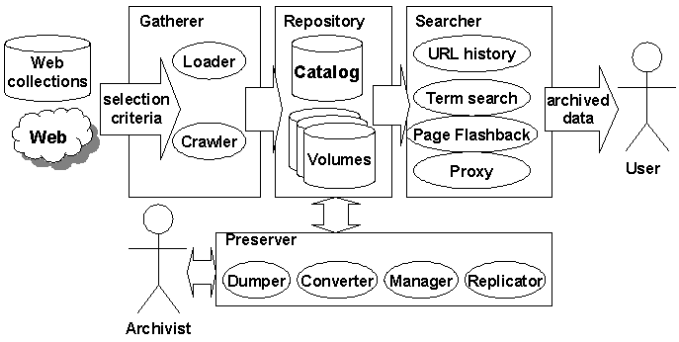

Fig. 2. Architecture of the Tomba web archive.

exclusion mechanisms does not significantly increase the coverage of a national web crawl. However, this behavior may degrade the crawler's performance because exclusion mechanisms are also used to prevent crawlers against hazardous situations.

3 The Tomba web archive

The Tomba web archive is a prototype system developed at the University of Lisbon to research web archiving issues. A web archive system must present an architecture able to follow the pace of the evolution of the web, supporting distinct selection criteria and gathering methods. Meta-data must be kept to ensure the correct interpretation and preservation of the archived data. A collection of documents built through incremental crawls of the web contains duplicates, given the documents that remain unchanged and the different URLs that reference the same document. It is desirable to minimize duplication among the archived data to save storage space without jeopardizing performance. The storage space must be extensible to support the collection growth and support various storage policies according to the formats of the archived documents and the level of redundancy required. The archived data should be accessible to humans and machines, supporting complementary access methods to fulfill the requirements of distinct usage contexts. There must be adequate tools to manage and preserve the archived documents, supporting their easy migration to different technological platforms.

Figure 2 represents the architecture of Tomba. There are 4 main components. The *Gatherer* is responsible for collecting web documents and integrating them in the archive. The *Repository* stores the contents and their correspondent meta-data. The *Preserver* provides tools to manage and preserve the archived data. The *Searcher* allows human users to easily access the archived data. The *Archivist* is a human expert that manages preservation tasks and defines selection criteria to automatically populate the archive.

3.1 Repository

A content is the result of a successful download from the web (e.g. an HTML file), while meta-data is information that describes it (e.g. size). The *Repository* is composed by

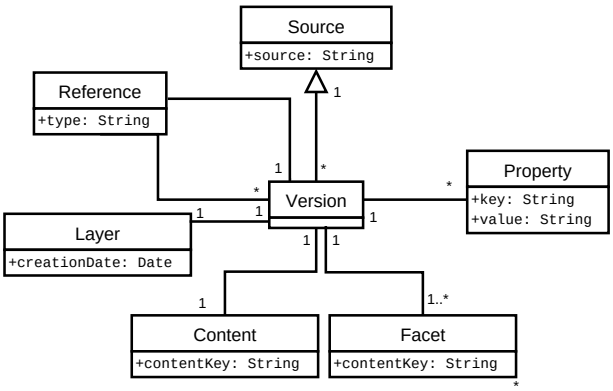

Fig. 3. Data model.

the *Catalog* [3] that provides high performance structured access to meta-data and the *Volumes* [9] that provide an extensible storage space to keep the contents, eliminating duplicates among them.

Figure 3 describes the data model of the Catalog. We assume that the archive is loaded in bulk with snapshots of the web. The *Source* class identifies the origin of the document, for example an URL on the web. Each *Version* represents a snapshot of the information gathered from a Source. The Versions correspondent to the same snapshot of the web are aggregated in *Layers*. A Layer represents the time interval from its creation until the creation of the next one. This way, time is represented in a discrete fashion within the archive, facilitating the identification of web documents that need to be presented together, such as a page and the embedded images. The Property class holds property lists containing meta-data related to a Version. The use of property lists instead of a static meta-data model, enables the incremental annotation of contents with meta-data items when required in the future. The *Content* and *Facet* classes reference documents stored in the Volumes. The former references the documents in their original format and the latter alternative representations. For instance, a Content is an HTML page that has a Facet that provides the text contained in it. In the archive, Facets provide storage for current representations of contents retrieved earlier in obsolete formats. The Repository supports merging the Content, Facets and meta-data of a Version into a single Facet in a semi-structured format (XML), so that each document archived in a Volume can be independently accessed from the Catalog. There are web documents that contain useful information to preserve other ones. For instance, a web page containing the specification of the HTML format could be used in the future to interpret documents written in this format. The *Reference* class enables the storage of associations of Versions that are related to each other.

3.2 Gatherer

The *Gatherer*, composed by the *Loader* and the *Crawler*, integrates web data in the Repository. The *Loader* was designed to support the delivery of web contents by publishers and receive previously compiled collections of documents. The *Crawler* iteratively harvests information from the web, downloading pages and following the linked URLs. Ideally, a page and the embedded or referenced documents would be crawled sequentially to avoid that some of them become unavailable meanwhile. Sequentially crawling all the documents referenced by a page degrades the crawler's performance, because harvesting the documents hosted outside a site requires additional DNS lookups and establishment of new TCP connections. According to Habib and Abrams, these two factors account for 55% of the time spent downloading web pages [12]. Crawling the documents of one site at a time in a breadth first mode and postponing the crawl of external documents until the corresponding sites are visited, is a compromise solution that ensures that the majority (71%) of the embedded documents internal to each site are crawled in a short notice, without requiring additional bandwidth usage [18].

3.3 Preserver

Replication is crucial to prevent data loss and ensure the preservation of the archived documents. The replication of data among mirrored storage nodes must consider the resources available, such as disk throughput and network bandwidth. A new document loaded into the archive can be immediately stored across several mirrors, but this is less efficient than replicating documents in bulk. Considering that an archive is populated with documents crawled from the web within a limited time interval, the overhead of replicating each document individually could be prohibitive. The *Replicator* copies the information kept in a Volume to a mirror in batch after each crawl is finished. The *Dumper* exports the archived data to a file using 3 alternative formats: i) WARC, proposed by the Internet Archive to facilitate the exportation of data to other web archives [17]; ii) an XML based format to enable flexible automatic processing; iii) a textual format with minimum formatting created to minimize the space used by the dump file. The dissemination of the archived documents as public collections is an indirect way to replicate them outside the archive, increasing their chance of persisting into the future. These collections are interesting for scientific evaluations [14] or to be integrated in other web archives. The main obstacles to the distribution of web collections are their large size, the lack of standards to format them in order to be easily integrated in external systems and copyright legislation that requires authorization from the authors of the documents to distribute them. Obtaining these authorizations is problematic for web collections having millions of documents written by different authors. The archived documents in obsolete formats must be converted to up-to-date formats to maintain their contents accessible. The *Converter* iterates through the documents kept in the Repository and generates Facets containing alternative representations in different formats. The *Manager* allows a human user to access and alter the archived information. The meta-data contained in the *Content-Type* HTTP header field identifies the media type of a web document but sometimes it does not correspond to the real media type of the document. On our baseline crawl, 1.8% of the documents identified as plain text were in

Fig. 4. Tomba web interface.

fact JPEG image files. The format of a document is commonly related to the file name extension of the URL that references it. This information can be used to automatically correct erroneous media type meta-data. However, the usage of file name extensions is not mandatory within URLs and the same file name extension may be used to identify more than 1 format. For example, the extension .rtf identifies documents in the application/rtf and text/richtext media types. In these cases, a human expert can try to identify the media type of the document and correct the corresponding meta-data using the Manager.

3.4 Searcher

The *Searcher* provides 3 methods for accessing the archived data: *Term Search*, *URL History* or *Navigation*. The Term Search method finds documents containing a given term. The documents are previously indexed to speed up the searches. The URL History method finds the versions of a document referenced by an URL. The Navigation method enables browsing the archive using a web proxy.

Figure 4 presents the public web interface of Tomba that supports the URL History access method. Navigation within the archive begins with the submission of an URL in the input form of the Tomba home page. In general, multiple different URLs reference the same resource on the web and it may seem indifferent to users to submit

any of them. If only exact matches on the submitted URL were accepted, some documents might not be found in the archive. Hence, Tomba expands each submitted URL to a set of URLs that are likely to reference the same resource, and then searches for them. For instance, if a user inputs the URL www.tumba.pt, Tomba will look for documents harvested from the URLs: www.tumba.pt/, tumba.pt, www.tumba.pt/index.html, www.tumba.pt/index.htm, www.tumba.pt/index.php, www.tumba.pt/index.asp. On the visualization interface, the archive dates of the available versions of a document are displayed on the left frame. The most recent version of the document is initially presented on the right frame and users can switch to other versions by clicking on the associated dates. The versions presented on the left frame enable a quick tracking of the evolution of a document. The documents harvested from the web are archived in their original format. However, they are transformed before being presented to the user to enable mimicking their original layout and allow a user to follow links to other documents within the archive when activating a link on a displayed page. The documents are parsed and the URLs to embedded images and links to other documents are replaced to reference archived documents. When a user clicks on a link, Tomba picks the version of the URL in the same layer of the referrer document and displays it on the right frame along with the correspondent versions on the left frame. A user may retrieve an archived document without modifications by checking the box *original content* below the submission form (Figure 4). This is an interesting feature for authors that want to recoverer old versions of a document. The Page Flashback mechanism enables direct access to the archived versions of a document from the web being displayed on the browser. The user just needs to click on a toolbar icon and the versions of the page archived in Tomba will be immediately presented.

The URL History access method has 3 main limitations. First, users may not know which URL they should submit to find the desired information. Second, the short life of URLs limits their history to a small number of versions. The Tomba prototype was loaded with 10 incremental crawls of the Portuguese web but on average each URL referenced just 1.7 versions of a document. Third, the replacement of URLs may not be possible in pages containing format errors or complex scripts to generate links. If these URLs reference documents that are still online, the archived information may be presented along with current documents. The Term Search and Navigation complement the URL History but they have other limitations. The Term Search finds documents independently from URLs but some documents may not be found because the correspondent text could not be correctly extracted and indexed [7]. The Navigation method enables browsing the archive without requiring the replacement of URLs because all the HTTP requests issued by the user's browser must pass through the proxy that returns contents only for archived documents. However, it might be hard to find the desired information by following links among millions of documents.

4 Related work

According to the National Library of Australia there are 16 countries with well-established national Web archiving programs [20]. The Internet Archive was the pioneer web archive. It has been executing broad crawls of the web and released an open-source crawler

named Heritrix [11]. The National Library of Australia founded its web archive initiative in 1996 [22]. It developed the PANDAS (PANDORA Digital Archiving System) software to periodically archive Australian online publications, selected by librarians for their historical value. The British Library leads a consortium that is investigating the issues of web archival [4]. The project aims to collect and archive 6,000 selected sites from the United Kingdom during 2 years using the PANDAS software. The sites have been stored, catalogued and checked for completeness. The MINERVA (Mapping the INternet Electronic Resources Virtual Archive) Web Archiving Project was created by the Library of the Congress of the USA and archives specific publications available on the web that are related to important events, such as an election [25].

In December 2004 the Danish parliament passed a new legal deposit law that calls for the harvesting of the Danish part of the Internet for the purpose of preserving cultural heritage and two libraries became responsible for the development of the Netarkivet web archive [19]. The legal deposit of web contents in France will be divided among the Institut National de l'Audiovisuel (INA) and the National Library of France (BnF). Thomas Drugeon presented a detailed description of the system developed to crawl and archive specific sites related to media and audiovisual [7]. The BnF will be responsible for the archive of online writings and newspapers and preliminary work in cooperation with a national research institute (INRIA) has already begun [1].

The National Library of Norway had a three-year project called Paradigma (2001-2004) to find the technology, methods and organization for the collection and preservation of electronic documents, and to give the National Library's users access to these documents [2]. The defunct NEDLIB project (1998-2000) included national libraries from several countries (including Portugal) and had the purpose of developing harvesting software specifically for the collection of web resources for an European deposit library [13]. The Austrian National Library together with the Department of Software Technology at the Technical University of Vienna, initiated the AOLA project (Austrian On-Line Archive) [23]. The goal of this project is to build an archive by harvesting periodically the Austrian web. The national libraries of Finland, Iceland, Denmark, Norway and Sweden participate in the Nordic Web Archive (NWA) project [**?**]. The purpose of this project is to develop an open-source software tool set that enables the archive and access to web collections.

5 Conclusions and Future work

We proposed and evaluated selection criteria to automatically populate a national web archive. We observed that no criteria alone provides the solution for selecting the contents to archive and combinations must be used. Some criteria are not selective but their use may prevent difficulties found while populating the archive. In particular, we conclude that populating a national web archive only with documents hosted in sites under the country's Top Level Domain or physically located on the country excludes a large amount of documents. The costs and complexity of the preservation of documents increases with the variety of media types archived. We observed that archiving documents of just three media types (HTML, GIF and JPEG) reduced the coverage of a national

web only 5%. We conclude that this is an interesting selection criterion to simplify web archival, in exchange for a small reduction on the coverage of the web.

We described the architecture of an information system designed to fulfil the requirements of web archiving and validate it through the development of a prototype named Tomba. We loaded Tomba with 57 million documents (1.5 TB) harvested from the Portuguese web during the past 4 years and explored three different access methods. None of them is complete by itself, so they must be used in conjunction to provide access to the archived data.

As future work, we intend to enhance accessibility to the archived information by studying an user interface suitable to access a web archive.

References

1. S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati. A first experience in archiving the french web. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 1–15, London, UK, 2002. Springer-Verlag.
2. K. Albertsen. The paradigma web harvesting environment. In *Proceedings of 3rd ECDL Workshop on Web Archives*, Trondheim, Norway, August 2003.
3. J. Campos. Versus: a web repository. Master thesis, 2003.
4. U. W. A. Consortium. Uk web archiving consortium: Project overview. http://info.webarchive.org.uk/, January 2006.
5. P. D. Corporation. Perseus blog survey. September 2004.
6. M. Day. Collecting and preserving the world wide web. `http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf`, 2003.
7. T. Drugeon. A technical approach for the french web legal deposit. In *5th International Web Archiving Workshop (IWAW05)*, Viena, Austria, September 2005.
8. R. Entlich. Blog today, gone tomorrow? preservation of weblogs. *RLG Diginews*, 8(4), August 2004.
9. D. Gomes, A. L. Santos, and M. J. Silva. Managing duplicates in a web archive. In L. M. Liebrock, editor, *Proceedings of the 21th Annual ACM Symposium on Applied Computing (ACM-SAC-06)*, Dijon, France, April 2006.
10. D. Gomes and M. J. Silva. Characterizing a national community web. *ACM Trans. Inter. Tech.*, 5(3):508–531, 2005.
11. M. S. I. R. Gordon Mohr, Michele Kimpton. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, Bath, UK, September 2004. Internet Archive, USA.
12. M. A. Habib and M. Abrams. Analysis of sources of latency in downloading web pages. In *WebNet*, San Antonio, Texas, USA, November 2000.
13. J. Hakala. Collecting and preserving the web: Developing and testing the nedlib harvester. *RLG Diginews*, 5(2), April 2001.
14. D. Hawking and N. Craswell. Very large scale retrieval and web search. In E. Voorhees and D. Harman, editors, *The TREC Book*. MIT Press, 2004.
15. A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.
16. M. Koster. A standard for robot exclusion. http://www.robotstxt.org/wc/norobots.html, June 1994.
17. J. Kunze, A. Arvidson, G. Mohr, and M. Stack. *The WARC File Format (Version 0.8 rev B)*, January 2006.

18. M. Marshak and H. Levy. Evaluating web user perceived latency using server side measurements. *Computer Communications*, 26(8):872–887, 2003.
19. F. McCown. Dynamic web file format transformations with grace. In *5th International Web Archiving Workshop (IWAW05)*, Viena, Austria, September 2005.
20. National Library of Australia. Padi - web archiving. http://www.nla.gov.au/padi/topics/92.html, January 2006. 18.
21. A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM Press, 2004.
22. M. Phillips. PANDORA, Australia's Web Archive, and the Digital Archiving System that Supports it. *DigiCULT.info*, page 24, 2003.
23. A. Rauber, A. Aschenbrenner, and O. Witvoet. Austrian on-line archive processing: Analyzing archives of the world wide web, 2002.
24. H. Snyder and H. Rosenbaum. How public is the web?: Robots, access, and scholarly communication. Working paper WP-98-05, Center for Social Informatics, Indiana University, Bloomington, IN USA 47405-1801, January 1998.
25. The Library of Congress. Minerva home page (mapping the internet electronic resources virtual archive, library of congress web archiving). http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html, January 2006.
26. The Web Robots Pages. Html author's guide to the robots meta tag. http://www.robotstxt.org/wc/meta-user.html, March 2005.