A survey on web archiving initiatives

Daniel Gomes, João Miranda, and Miguel Costa

Foundation for National Scientific Computing (FCCN)**
Av. do Brasil, 101
1700-066 Lisboa, Portugal
(daniel.gomes, joao.miranda, miguel.costa)@fccn.pt

Abstract. Web archiving has been gaining interest and recognized importance for modern societies around the world. However, for web archivists it is frequently difficult to demonstrate this fact, for instance, to funders. This study provides an updated and global overview of web archiving. The obtained results showed that the number of web archiving initiatives significantly grew after 2003 and they are concentrated on developed countries. We statistically analyzed metrics, such as, the volume of archived data, archive file formats or number of people engaged. Web archives all together must process more data than any web search engine. Considering the complexity and large amounts of data involved in web archiving, the results showed that the assigned resources are scarce. A Wikipedia page was created to complement the presented work and be collaboratively kept up-to-date by the community.

1 Introduction

The web was invented to exchange data between scientists but it quickly became a crucial mean of publication. However, the web is extremely ephemeral. Most of its information becomes unavailable and is lost forever after a short period of time. It was observed that 80% of the pages are updated or disappear after 1 year [49]. Even printed publications suffer from the effects of web data transience because they frequently cite online resources that became unavailable [52]. Besides losing important scientific and historical information, the transience of the information published on the web causes common people to lose their memories as individuals (e.g. photos shared exclusively through the web). Broken links also degrade the performance of popular web applications and services, such as shared bookmarks, search engines or social networks, leading their users to dissatisfaction.

The web needs preservation initiatives to fight ephemerality. It must be ensured that the information besides being accessible worldwide, prevails across time to transmit knowledge for future generations. Web archives are innovative systems that acquire, store and preserve information published on the web. Notably, they also contribute to preserve contents born in non-digital formats that were afterwards digitized and published online. Web archives enable numerous new use cases. Journalists can look for information to document articles, software engineers can search for documentation to

** Co-funded by: MCTES/UMIC; POS_C and EU

fix legacy systems, webmasters can recover past versions of their site's pages or historians can analyze web pages as they do for paper documents.

This study presents a survey that draws a picture of worldwide initiatives to preserve information published on the web. We gathered results about 42 web archiving initiatives and analyzed metrics, such as, the volume of archived data, used formats or number of people engaged. Considering the complexity and large amounts of data involved in web archiving, the results showed that the resources being assigned are still scarce.

During our research we observed that the publicly available information about web archives is frequently obsolete or inexistent. A complementary contribution of this study was the creation of a Wikipedia page named *List of Web Archiving Initiatives*[1], so that the published information can be collaboratively kept up-to-date.

2 Related Work

The National Library of Australia maintains a page listing the 17 major archiving initiatives to preserve web heritage around the world [36]. The book *Web Archiving* discusses issues related to the preservation of the web and refers to several initiatives [28]. The Web Archiving Workshop began in 2001 and yearly presents updated work about this field [19].

The Joint Information Systems Committee (JISC) published three studies about web archiving. One addressed the legal issues relating to the archiving of Internet resources in the United Kingdom, European Union, USA and Australia, and presented recommendations about the policies that should be adopted in the UK [6]. The second study discussed the feasibility of collecting and preserving the web and presented a review about 8 web archiving initiatives [8] and the most recent one analyzed the researchers engagement with web archives [11].

Shiozaki and Eisenschitz reported on a questionnaire survey of 16 national libraries designed to clarify how they attempt to justify their web archiving activities [51]. The conclusion was that national libraries envisage that the benefits brought by their initiatives are greater than the costs and they are struggling to respond to legal risks (e.g. legislation, contracting and opt-out policies).

The International Internet Preservation Consortium (IIPC) was founded in 2003 and is composed by institutions that collaborate to preserve Internet content for future generations [14]. In 2008, the IIPC published the results of a survey conducted to derive profiles of its members. The survey addressed issues such as membership type, staff, used tools, legal issues and selection criteria.

During December 2010, in the context of the European research project Living Web Archives, the Internet Memory Foundation conducted a survey to characterize web archiving institutions and analyze the main problems of this field in Europe [22]. Statistics regarding institution type, legal context, management and archiving policies were provided.

The 18^{th} Conference of Directors of National Libraries in Asia and Oceania published a report containing the answers obtained through a questionnaire about web

[1] `http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives`

archiving submitted to participant countries [34]. The answers were provided as free text and do not enable a rigorous quantitative analysis. However, they provide a rich qualitative overview about web archiving in this region of the world addressing legal frameworks, main challenges to overcome, collaborations, system descriptions, policies concerning acquisition, access and preservation. In 2010, there were 6 web archiving projects.

Our study presents an updated overview about web archiving initiatives across the world. It is the most comprehensive scientific study about web archiving. The methodology adopted differs from previous work because it was designed to obtain both quantitative and qualitative results through an interactive process with the respondents[2].

3 Methodology

Initially, this research aimed to obtain answers to the following questions about each web archiving initiative:

1. What is the name of your web archive initiative (please state if you want to remain anonymous)?
2. How many people work at your web archive (in person-month)?
3. Which is the amount of data that you have archived (number of files, disk space occupied)?

During October 2010, we tried to gather this information from the official sites and published documentation but we did not succeed because the published information was frequently insufficient or obsolete. Plus, many official sites were exclusively available on the native language of the hosting country (e.g. Chinese) and automatic translation tools were insufficient to obtain the required information. We decided to contact directly the community to complement our results. The questions were sent to a web archive discussion list, published on the site of the Portuguese Web Archive and disseminated through its communication channels (Twitter, Facebook, RSS). We obtained 27 answers. Then, we sent direct e-mails to the remaining web archives referenced by the IIPC [14], National Library of Australia [36] and Web Archiving Workshops [19]. We were able to establish contact and obtain direct answers from 33 web archiving initiatives. Finally, we sent the obtained results to the respondents for validation.

The methodology used in this research enabled web archivists to openly present information about their initiatives. For some situations, we had to actively interact with the respondents to obtain the desired information. We observed that terminology and language barriers led to different interpretations of the questions by the respondents, who involuntarily provided inaccurate answers. For instance, in question 3, we assumed that each archived file was the result of a successful HTTP download (e.g. page, image or video) but some respondents interpreted it as the number of files created to store web contents in bulk (ARC files [4]). The posterior statistical analysis of the results enabled the detection of abnormal values and correction of these errors through interaction with the respondents. We believe that the adopted methodology enabled the extraction of

[2] We would like to express our deep gratitude to everyone who collaborated with our survey.

Table 1. List of web archives (WA). The names of the initiatives were shortened but the references contain the official ones. The description of initiatives marked with * was exclusively gathered from publicly available information.

Initiative short name	Hosting country	Creation year	Staff		Main scope of archived content
			Full-time	Part-time	
Australia's WA [37]	Australia	1996	4	4.25	National
Tasmanian WA [54]	Australia	1996	0	1	Regional
Web@rchive [1]	Austria	2008	0	2	National
DILIMAG [18]	Austria	2007	2	0	German literature magazines
Canada WA [25]	Canada	2005	0	2	National governmental
Chinese WA* [38]	China	2003	n.a.	n.a.	National
Croatian WA [30]	Croatia	2004	4	3	National
WebArchiv [45]	Czech Republic	2000	5	0	National
Netarkivet.dk [53]	Denmark	2005	0	18	National
Finnish WA [57]	Finland	2008	2	2	National
BnF [39]	France	2006	9	0	National
INA* [17]	France	2009	n.a.	n.a.	National audiovisual
Internet Memory [23]	France, Netherlands	2004	21	0	International & service provider
Baden-Württemberg [2]	Germany	2003	7.5	0	German literature
German Bundestag* [10]	Germany	2005	n.a.	n.a.	German parliament
Iceland* [31]	Iceland	2004	n.a.	n.a.	National
WA Project [33]	Japan	2004	10	2	National
OASIS [40]	Korea	2001	3	11	National
Koninklijke Bibliotheek [46]	Netherlands	2006	1	1	National
New Zealand WA [41]	New Zealand	1999	3	10	National
National Library Norway* [42]	Norway	n.a.	n.a.	n.a.	National
Portuguese WA [12]	Portugal	2007	4	1	National
WA of Čačak [50]	Serbia	2009	0	1	Regional
WA Singapore* [35]	Singapore	n.a.	n.a.	n.a.	National
Slovenian WA [16]	Slovenia	2007	1	0	National
Preservation .ES [43]	Spain	2006	2	2	National
Digital Heritage Catalonia [26]	Spain	2006	4	0	Regional
Kulturarw3* [44]	Sweden	1996	n.a.	n.a.	National
WA Switzerland [55]	Switzerland	2008	0	3	National
NTUWAS [47]	Taiwan	2007	0	3	National
WA Taiwan* [32]	Taiwan	2007	n.a.	n.a.	National
UK WA [3]	UK	2004	n.a.	0	National
UK Gov WA[56]	UK	2004	4	2	National governmental
Internet Archive [21]	USA	1996	12	0	International & service provider
Columbia University [7]	USA	2009	3	1	Thematic: human rights
North Carolina [48]	USA	2005	0	3	Regional
Latin American* [62]	USA	2005	n.a.	n.a.	International focused on Latin America
WA Pacific Islands [61]	USA	2008	0	4	International focused on Pacific Islands
Library of Congress [27]	USA	2000	6	80	National
Harvard University Library [15]	USA	2006	0	6	Institutional
California Digital Library [5]	USA	2005	4	1	International & service provider
University of Michigan [58]	USA	2000	0	2	Institutional

more accurate information and valuable insights about web archiving initiatives worldwide, than a typical one-shot online survey with closed answers. However, the cost of processing the results for statistical analysis was significantly higher.

4 Web archiving initiatives

Table 1 presents the 42 web archiving initiatives identified across the world ordered alphabetically by their hosting country. Web archiving initiatives are very heterogeneous in size and scope. The WA of Čačak aims to preserve sites related to this Serbian city, while the Internet Archive has the objective of archiving the global web. The obtained results show that 80% of the archives exclusively hold content related to their hosting country, region or institution. However, initiatives hosted in the USA like the Latin

American WA, Internet Archive or the WA Pacific Islands also preserve information related to foreign countries. The creation and operation of a web archive is complex and costly. The Internet Archive, Internet Memory and California Digital Library provide web archiving services that can be independently operated by third-party archivists. The services are named Archive-it[3], ArchiveTheNet[4] and Web Archiving Service[5], respectively. These services enable focused archiving of web contents by organizations, such as universities or libraries, that otherwise could not manage their own archives. For instance, the Archive-it service is used by the North Carolina, ArchiveTheNet is used by the UK Government WA and the Web Archiving Service by the University of Michigan.

The measurement of human resources engaged in web archiving activities was not straightforward (question 2). Most respondents could not provide an effort measurement in person-month. The presented reasons were that the teams were too variable and some services were hired to third-party organizations out of their control. Instead, most of the respondents described their staff and hiring conditions. The obtained results show that web archiving engages at least 112 people in full-time and 166 in part-time. The total of 277 people that preserve and provide access to the past of the web since its inception contrasts with the resources invested to provide access to a snapshot of the current web. For instance, Google by itself has 24 400 full-time employees, from which 9 508 work in research and development and 2 768 in operations [60]. The web archive teams are typically small, presenting a median staff of 2.5 people in full-time (average of 3.5) and 2 people in part-time (average of 5) and are mostly composed by librarians and information technology engineers. The results show that 11 initiatives (26%) don't have any person dedicated full-time. The effort of part-time workers is variable, for instance, at the Library of Congress they spend only a few hours a month. Most of the human resources are invested on data acquisition and quality control.

Figure 1 presents the location of countries that host web archiving initiatives. The 42 initiatives are spread across 26 countries. There are 23 initiatives hosted in Europe, 10 in North America, 6 in Asia and 3 in Oceania. Half of the initiatives are hosted in countries belonging to the Organisation for Economic Co-operation and Development (OECD). From the 34 countries that belong to the OECD, 21 (62%) host at least one web archiving initiative, which is an indicator of the importance of web archiving in developed countries. Most of the countries host one (74%) or two initiatives (22%). The only country that hosts more is the USA with 8 initiatives. Although being part of a country, initiatives like the Tasmanian WA (Australia), North Carolina (USA) or Digital Heritage Catalonia (Spain) are hosted at autonomous states and aim at preserving regional content.

Figure 2 presents the evolution of the number of web archiving initiatives created per year. The first web archive named Internet Archive was founded by Brewster Kale in 1996. Three initiatives followed in 1996: the Australia's Web Archive and the Tasmanian's Web Archive from Australia, and Kulturarw3 from Sweden. Only 5 new initiatives arose during the following 6 years. However, since 2003 there was a significant and constant growth with the creation of 31 initiatives, reaching 6 initiatives per year

[3] http://www.archive-it.org
[4] http://archivethe.net
[5] http://webarchives.cdlib.org

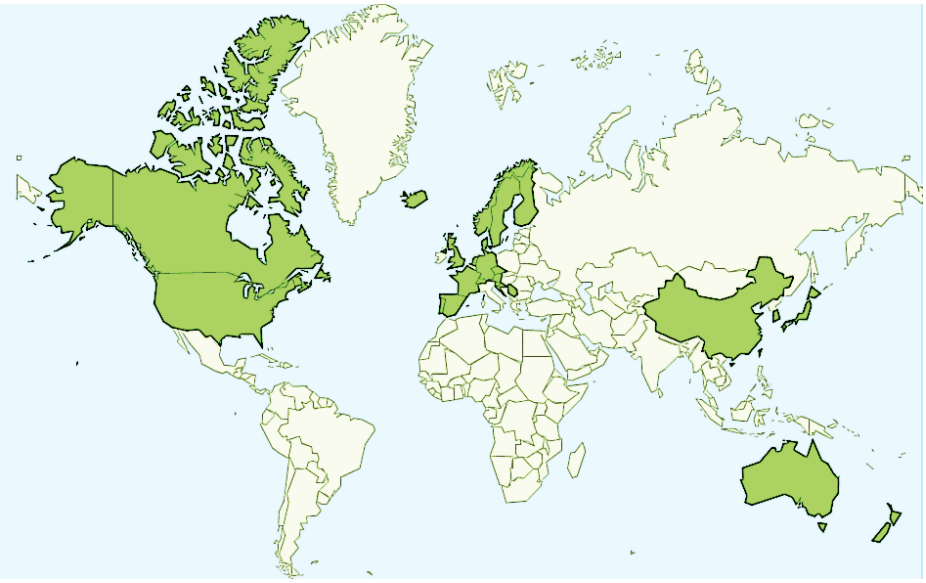

Fig. 1. Countries hosting web archiving initiatives.

in 2004 and 2005. One possible explanation for this fact was the concern raised by the United Nations Educational, Scientific and Cultural Organization (UNESCO) regarding the preservation of the digital heritage [59].

5 Archived data

All web archives select specific sites for archiving. This selection is determined by factors such as consent by the authors or relevance for inclusion in thematic collections (e.g. elections or natural disasters). Eleven initiatives (26%) also perform broad crawls of the web, including all the sites hosted under a given domain name or geographical location.

Figure 3 presents the distribution of the archived collections measured in total volume of data and number of contents. For instance, one HTML page containing three embedded images results in the archive of four contents. The objective of this measurement was to characterize web archives regarding the total amount of data they held. Selective web archiving is frequently focused on preserving individual sites. Thus, the number of archived sites could also be an interesting metric. However, the size of web sites significantly varies and the number of archived sites by itself is not descriptive of the volume of archived data. Therefore, we decided not to include this metric to simplify the questionnaire. The results show that 50% of the collections are smaller than 10 TB and are composed by less than 1 000 million contents (78%). The volume of data correspondent to the creation of replicas to ensure preservation was not considered

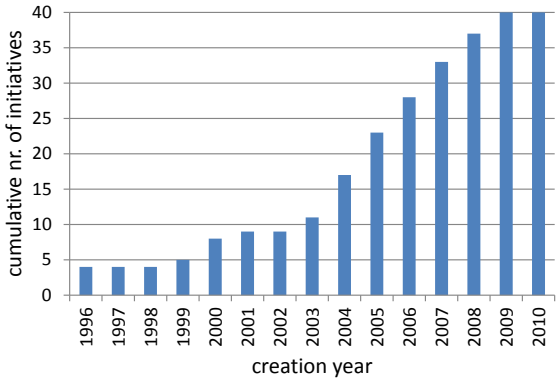

Fig. 2. Cumulative number of initiatives created per year.

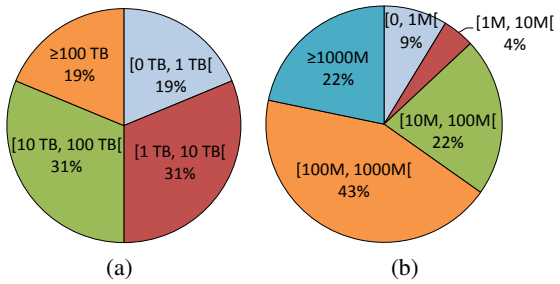

Fig. 3. Size of archived collections in: (a) Volume of data (Terabytes) (b) Number of contents (e.g. images, pages, videos).

in this measurement. The average content size was 46 KB and ranged between 14.2 KB and 119.4 KB. There are several reasons for this difference. Some web archives are focused on specific contents which are typically large, such as video, PDF documents or images. Web archives use different formats for archiving web data that may contain additional meta-data or use compression. Another reason is that the size of contents tends to grow [29]. Therefore, older archived contents tend to be smaller than recent ones. Web archives worldwide preserved since 1996 a total of 181 978 million contents (6.6 PB). The Internet Archive by itself holds 150 000 million contents (5.5 PB). The size of the current web cannot be accurately determined. However, in 2008 Google announced that one single snapshot of the web comprised 1 trillion unique URLs (10^{12}) [13]. Notice that this number refers only to web pages and does not include contents, such as images or videos, that are also addressed by web archives. The obtained results show that the amount of archived data is small in comparison with the volume of data that is permanently being published on the web.

Figure 4 presents the distribution of the file formats used to store archived content. The ARC format was defined by the Internet Archive and applied as a *de facto* standard

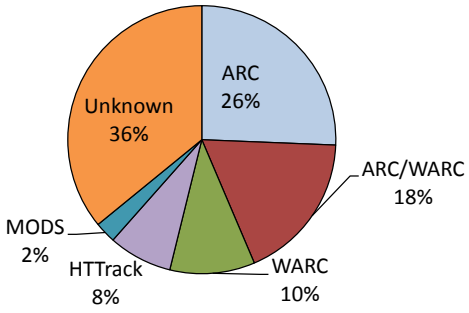

Fig. 4. Usage of file formats to store web contents.

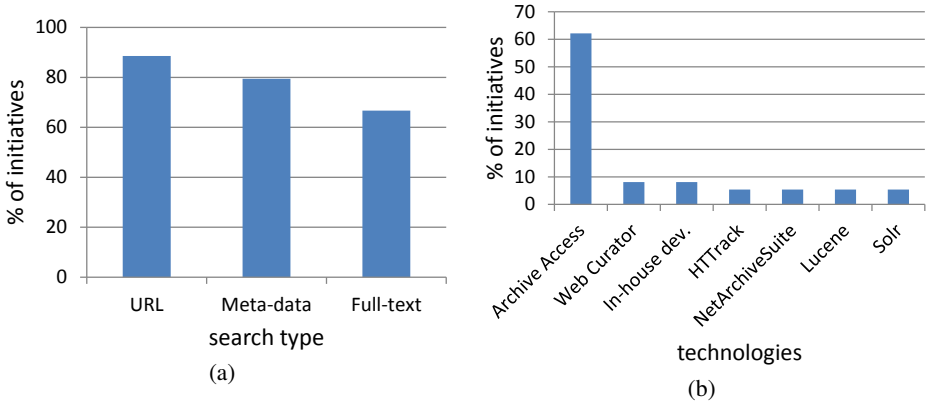

Fig. 5. Provided access to archives: (a) Access type (b) Used technologies.

[4]. In 2009, the WARC format was published by the Internet Organization for Standardization (ISO) as the official standard format for archiving web contents [24] and it is already exclusively used by 10% of the initiatives. The ARC and WARC formats are dominant, being used by 54% of the initiatives. The usage of standard formats for web archiving facilitates the collaborative creation of tools, such as search engines or replication mechanisms, to process the archived data. Besides historical reasons, the widespread of the ARC/WARC formats was motivated by the creation of the Archive-Access project that freely provides open-source tools to process this type of files [20].

6 Access and Technologies

Figure 5(a) presents the types of search provided by the initiatives over their collections. The obtained results show that 89% of the initiatives support access to the history of a given URL, 79% enable searching meta-data and 67% provide full-text search over archived contents. There are 21 initiatives (50%) that provide full online access

to search mechanisms and archived content. Some initiatives hold the copyright of the archived contents (e.g. German Bundestag, UK WA, Canada WA) or explicitly require the consent of the authors before archiving (UK WA, OASIS). The Tasmanian WA operated since its inception under the assumption that web sites fall within the definition of book. Thus, no permission to capture from publishers is required. The Internet Archive and the Portuguese WA proactively archive and provide access to contents but remove access on-demand. On the other hand, for 16 initiatives (38%) the access to the collections is somehow restricted. The Library of Congress, WebArchiv and Australia's WA provide public online access to part of their collections. Netarkivet.dk provides online access on-demand only for research purposes. The Finnish WA provides online access to meta-data but not to archived contents. BnF, Web@rchive and Preservation .ES grant access exclusively through special rooms on their facilities. Maintaining the accessibility level of the original information is mandatory to make web archives useful for citizens. If a content is publicly available on the current web, it should continue to be publicly available when it becomes a historical content. However, this policy collides with national legislations that restrict access or even inhibit proactive web archiving. The web broke economical and geographical barriers to information but legislations are raising them against historical content. It is economical unattainable for most people to travel, possibly to a foreign country, to investigate if an information published in the past exists in a web archive.

Figure 5(b) depicts the technologies being used by the initiatives that manage their own systems. Notice that 16% of the initiatives use software as a service to manage their collections. The Archive-Access tools are dominant (62%), including the Heritrix, NutchWAX and Wayback projects, that support content harvesting, full-text and URL search, respectively. However, respondents frequently mentioned that full-text search was hard to implement and that the performance of NutchWAX was unsatisfactory, being one reason for the partial indexing of their collections. Nonetheless, NutchWAX supports full-text search for the Finnish WA (148 million), Canada WA (170 million), Digital Heritage of Catalonia (200 million), California Digital Library (216 million) and BnF (estimated 2 100 million). Australia's WA supports full-text search over 3 100 million contents indexed using an in-house developed system named Trove. It was estimated that the largest web search engine is Google and that it indexes 38 000 million pages [9]. Creating a search engine over the archived so far (181 978 million contents), would imply indexing 4.7 times more data.

7 Conclusions

The preservation of digital heritage is crucial to modern societies because web publications are extremely transient. This study identified 42 web archiving initiatives created around the world since 1996. Web archives are typically hosted on developed countries and are composed by small teams that mainly work on the acquisition and curation of data. Most of the initiatives carefully select contents from the web to be archived. There are 3 organizations that provide web archiving services. The total amount of archived data so far reaches 6.6 PB (181 978 million contents). However, efficient search mechanisms are required to enable access to this information, which raises new technological

challenges. The largest web search engine indexes only 20% of this amount of data. An additional problem are the legal barriers that restrict access to historical web contents and diminish the visibility and importance of web archives to modern societies. Open access to historical web data would enable the creation of federated search mechanisms across web archives and the development of new applications by third-parties that would contribute to explore the potential of this valuable source of historical information. New laws regarding digital preservation and extension of the legal deposit to web contents have been approved. As future work, we intend to analyze the current legal situation worldwide regarding web archiving and its impact on cultural heritage.

Despite the social and economic impact of losing the information that is being permanently and exclusively published on the web, the obtained results show that the growing resources invested in web archiving are still relatively scarce. This fact will probably originate a historical void regarding our current times.

References

1. Austrian National Library. Österreichische Nationalbibliothek - Web archiving. http://www.onb.ac.at/ev/about/webarchive.htm, March 2011.
2. Bibliotheksservice-Zentrum Baden-Württemberg. Willkommen im Bibliotheksservice-Zentrum Baden-Württemberg. http://www.bsz-bw.de/index.html, March 2011.
3. British Library. UK Web Archive. http://www.webarchive.org.uk/ukwa/, March 2011.
4. M. Burner and B. Kahle. WWW Archive File Format Specification. http://pages.alexa.com/company/arcformat.html, September 1996.
5. California Digital Library. Web Archives: yesterday's web; today's archives. http://webarchives.cdlib.org/, March 2011.
6. A. Charlesworth. Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia. http://www.jisc.ac.uk/media/documents/programmes/preservation/archiving_legal.pdf, 2003.
7. Columbia University Libraries. Web Resources Collection Program. https://www1.columbia.edu/sec/cu/libraries/bts/web_resource_collection/, March 2011.
8. M. Day. Collecting and preserving the World Wide Web. http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf, 2003.
9. M. de Kunder. WorldWideWebSize.com | The size of the World Wide Web. http://www.worldwidewebsize.com/, March 2011.
10. Deutscher Bundestag. Deutscher Bundestag: Web-Archiv. http://webarchiv.bundestag.de/cgi/kurz.php, March 2011.
11. M. Dougherty, E. Meyer, C. Madsen, C. Van den Heuvel, A. Thomas, and S. Wyatt. Researcher engagement with web archives: State of the art. Technical report, Joint Information Systems Committee (JISC), 2010.
12. Foundation for National Scientific Computing. Portuguese Web Archive: search the past. http://www.archive.pt/, March 2011.
13. Google Inc. Official Google Blog: We knew the web was big... http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, July 2008.
14. A. Grotke. IIPC - 2008 Member Profile Survey Results. http://www.netpreserve.org/publications/IIPC_Survey_Report_Public_12152008.pdf, December 2008.
15. Harvard University Library. Web Archive Collection Service - Harvard University Library. http://wax.lib.harvard.edu/collections/home.do, March 2011.

16. Historical Archives of Ljubljana. Zgodovinski arhiv Ljubljana. http://www.zal-lj.si/, March 2011.
17. Ina. Ina.fr - A la une : vidéo, radio, audio et publicité - Actualités, archives du jour de la radio et de la télévision en ligne. http://www.ina.fr/, March 2011.
18. Innsbruck Newspaper Archive at the Univ. of Innsbruck and Dept. for Digitisation & Digital Preservation at the Univ. of Innsbruck Lib. Digitale Literatur Magazine:. http://dilimag.literature.at/default.alo, March 2011.
19. International Web Archiving Workshop. Index. http://iwaw.europarchive.org/, March 2011.
20. Internet Archive. Nutchwax - Home Page. http://archive-access.sourceforge.net/, March 2008.
21. Internet Archive. Internet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine. http://www.archive.org/, March 2011.
22. Internet Memory Foundation. Web Archiving in Europe. http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf, 2010.
23. Internet Memory Foundation. Welcome to Internet Memory Foundation website. http://internetmemory.org/en/, March 2011.
24. I. ISO. 28500: 2009 Information and documentation-WARC file format, 2009.
25. Library and Archives Canada. Home - Library and Archives Canada. http://www.collectionscanada.gc.ca/index-e.html, March 2011.
26. Library of Catalonia. PADICAT, Patrimoni Digital de Catalunya. http://www.padicat.cat/, March 2011.
27. Library of Congress. Web Archiving (Library of Congress). http://www.loc.gov/webarchiving/, March 2011.
28. J. Masanès. *Web Archiving*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
29. J. Miranda and D. Gomes. Trends in Web characteristics. In *7th Latin American Web Congress (LA-Web 2009)*, Merida, Mexico, November 2009.
30. National and University Library in Zagreb. Hrvatski arhiv weba, HAW. http://haw.nsk.hr/, March 2011.
31. National and University Library of Iceland. Vefsafn - English. http://vefsafn.is/index.php?page=english, March 2011.
32. National Central Library, Taiwan. Web Archive Taiwan. http://webarchive.ncl.edu.tw/nclwa98Front/, March 2011.
33. National Diet Library. Web Archiving Project. http://warp.da.ndl.go.jp/search/, March 2011.
34. National Diet Library, Japan - Conference of Directors of National Libraries in Asia and Oceania 2010. Report on questionnaire survey on web-archiving - Document 3. http://www.ndl.go.jp/en/cdnlao/meetings/pdf/report_Japan1_doc3.pdf, 2010.
35. National Library Board Singapore. Web Archive - National Library Board, Singapore. http://was.nl.sg/, March 2011.
36. National Library of Australia. PADI - Preserving Access to Digital Information. http://www.nla.gov.au/padi/, March 2011.
37. National Library of Australia. Pandora Archive - Preserving and Accessing Networked DOcumentary Resources of Australia. http://pandora.nla.gov.au/, March 2011.
38. National Library of China. Web Information Collection and Preservation - WICP (Chinese Web Archive). http://210.82.118.162:9090/webarchive, March 2011.
39. National Library of France. BnF - Digital legal deposit. http://www.bnf.fr/en/professionals/digital_legal_deposit.html, March 2011.
40. National Library of Korea. About OASIS - About OASIS. http://www.oasis.go.kr/intro_new/intro_overview_e.jsp, March 2011.

41. National Library of New Zealand. New Zealand Web Archive - National Library of New Zealand. http://www.natlib.govt.nz/collections/a-z-of-all-collections/nz-web-archive, March 2011.
42. National Library of Norway. Nasjonalbiblioteket ǁ index. http://www.nb.no/, March 2011.
43. National Library of Spain. Biblioteca Nacional de España. Ministerio de Cultura. http://www.bne.es/es/LaBNE/PreservacionDominioES/, March 2011.
44. National Library of Sweden. Swedish Websites - Kungliga biblioteket. http://www.kb.se/english/find/internet/websites/, March 2011.
45. National Library of the Czech Republic. WebArchiv. http://en.webarchiv.cz/, March 2011.
46. National library of the Netherlands. Web archiving. http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html, March 2011.
47. National Taiwan University Library. NTU Web Archiving System, NTUWAS. http://webarchive.lib.ntu.edu.tw/eng/default.asp, March 2011.
48. North Carolina State Archives and State Library of North Carolina. North Carolina State Government Web Site Archives. http://webarchives.ncdcr.gov/, March 2011.
49. 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.
50. Public Library Čačak. Web Archive of Cacak - English - Digitalizacija i digitalne biblioteke. http://digital.cacak-dis.rs/english/web-archive-of-cacak/, March 2011.
51. R. Shiozaki. Role and justification of web archiving by national libraries - A questionnaire survey. http://lis.sagepub.com/content/41/2/90, 2009.
52. D. Spinellis. The decay and failures of web references. *Communications of the ACM*, 46(1):71–77, 2003.
53. State and University Library. netarkivet.dk. http://netarkivet.dk/index-da.php, March 2011.
54. State Library of Tasmania. Our Digital Island. http://odi.statelibrary.tas.gov.au/, March 2011.
55. Swiss National Library. Swiss National Library NL -e-Helvetica. http://www.nb.admin.ch/nb_professionnel/01693/index.html?lang=en, March 2011.
56. The National Archives. UK Government Web Archive | The National Archives. http://www.nationalarchives.gov.uk/webarchive/, March 2011.
57. The National Library of Finland. Finnish Web Archive. http://verkkoarkisto.kansalliskirjasto.fi/, March 2011.
58. The Regents of the University of Michigan. University of Michigan Web Archives. http://bentley.umich.edu/uarphome/webarchives/webarchive.php, March 2011.
59. UNESCO. Charter on the Preservation of Digital Heritage. Adopted at the 32nd session of the General Conference of UNESCO, October 17, 2003. http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf.
60. United States Securities and Exchange Commission. Form 10-K. http://www.sec.gov/Archives/edgar/data/1288776/000119312511032930/d10k.htm, December 2010.
61. University of Hawaii at Manoa Library. Web Archiving Project for the Pacific Islands | University of Hawaii at Manoa Library. http://library.manoa.hawaii.edu/research/archiveit/, March 2011.
62. University of Texas at Austin. Latin American Web Archiving Project, LAWAP. http://lanic.utexas.edu/project/archives/, March 2011.