CrossMark

The evolution of web archiving

Miguel Costa[1] · Daniel Gomes[2] · Mário J. Silva[3]

Abstract Web archives preserve information published on the web or digitized from printed publications. Much of this information is unique and historically valuable. However, the lack of knowledge about the global status of web archiving initiatives hamper their improvement and collaboration. To overcome this problem, we conducted two surveys, in 2010 and 2014, which provide a comprehensive characterization on web archiving initiatives and their evolution. We identified several patterns and trends that highlight challenges and opportunities. We discuss these patterns and trends that enable to define strategies, estimate resources and provide guidelines for research and development of better technology. Our results show that during the last years there was a significant growth in initiatives and countries hosting these initiatives, volume of data and number of contents preserved. While this indicates that the web archiving community is dedicating a growing effort on preserving digital information, other results presented throughout the paper raise concerns such as the small amount of archived data in comparison with the amount of data that is being published online.

Keywords Web archiving · Digital preservation · Survey

✉ Miguel Costa
migcosta@gmail.com

1 Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

2 Foundation for National Scientific Computing, Lisbon, Portugal

3 INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

1 Introduction

The world wide web has a democratic nature, where everyone can publish all kinds of information using different types of media. News, blogs, wikis, encyclopedias, photos, interviews and public opinions are just a few examples of this vast list. Part of this information is unique and historically valuable. For instance, the speech of a president after winning an election or the announcement of an imminent invasion of a foreign country, might become as valuable in the future as ancient manuscripts are today. However, since the web is so dynamic, a large amount of information is lost everyday. Several studies quantify this loss: 80 % of web pages are not available in their original form after 1 year [1]; 13 % of web references in scholarly articles disappear after 27 months [2]; 11 % of social media resources, such as the ones posted on Twitter, are lost after 1 year [3]. All this information will likely vanish in a few years, creating a knowledge gap about the present for future generations. We are already experiencing unsatisfied information needs due to missing pages or old formats of documents that are not readable by the latest software version.[1] Pioneers of the Internet, such as Vint Cerf, recently warned about the danger of future generations who will have little or no record of the twenty-first century.[2] International organizations are also concerned with the web ephemerality problem. The UNESCO recognized the importance of digital preservation in 2003, by stating that the disappearance of digital information constitutes an impoverishment of the heritage of all nations [4]. In 2010, the UNESCO endorsed the Universal Declaration on Archives, which states that archives play an essential role in the development of societies by safeguard-

---

1 http://en.wikipedia.org/wiki/Digital_obsolescence.

2 http://www.bbc.com/news/science-environment-31450389.

ing and contributing to individual and community memory [5]. It is, therefore, important to preserve these data, not only for historical and social research [6–12], but also to support current technology, such as assessing the trustworthiness of statements [13], detecting web spam [14], improving web information retrieval [15] or forecasting events [16].

At least 68 web archiving initiatives undertaken by national libraries, national archives, private companies and consortia of organizations are acquiring and preserving parts of the web. Together, they hold more than 534 billion files (17 PB) and this number continues to grow as new initiatives arise. Some country code top-level domains and thematic collections are being archived regularly,[3] while other collections related to important events, such as September 11, are created at particular points in time.[4] Web archives also contribute to the preservation of content born in non-digital formats that were afterwards digitized and published online, such as The Times Archive[5] with news since 1785. As result, web archives contain often millions or billions of archived documents and cover decades or even centuries in the case of digitized publications. The historic interest in these documents is also growing as they age, becoming a unique source of past information for widely diverse areas, such as sociology, history, anthropology, politics, journalism, linguistics or marketing.

However, despite the existence of web archives since 1996 and their joint efforts to preserve digital information, information about web archiving initiatives and the services they provide is scarce. Without knowing the status of current web archiving it is impossible to understand its strengths, limitations and the developments that are still needed to turn these document repositories into useful sources of information. Without knowing the preferences, trends and needs of the web archiving community it is difficult to adapt current technology to the emerging challenges and develop strategies to anticipate future problems. Motivated by this lack of knowledge in the research community, we conducted two surveys to gather results about existing web archiving initiatives across the globe. The first survey, already published, provided a comprehensive characterization of world wide web archiving initiatives in 2010 [17]. The second survey was carried out in 2014 and provides an updated characterization of these initiatives. Both surveys analyzed the same metrics, which enabled to study the evolution of the characteristics of web archiving initiatives, such as the location, creation year, selection policy, used formats, number of people engaged, volume of archived data, access type and employed technol-

ogy. We also compared our two surveys against the results obtained from other surveys whenever possible.

The analysis evidences a significant growth in the number of initiatives, countries hosting these initiatives, volume of data and number of contents preserved, which indicates a growing effort that has been employed by the web archiving community to preserve the web. A cause for concern is the small amount of archived data in comparison with the amount of data being published on the web. This will likely originate a knowledge gap about the present time. On the other hand, the amount of archived data is larger and grows faster than the amount processed by any commercial web search engine, which raises scalability challenges in giving efficient and effective data access. In fact, the search tools have not changed in the last years, being essentially based on commonly used web search technology that does not take into account the specificities of web archiving. These tools have a poor performance and greatly affect the finding of historical information [18].

The remainder of this paper is organized as follows. Section 2 describes the background and covers related work. Section 3 describes the methodology for conducting the surveys on web archiving initiatives in 2010 and 2014. Section 4 presents the results obtained in the surveys and the analysis of the advancements made in web archiving during that period. Section 5 finalizes with the conclusions.

2 Related work

Cultural heritage institutions, such as museums, libraries and archives, have been preserving the intangible culture of our society (e.g., folklore, traditions, language) and the legacy of physical artifacts (e.g., monuments, books, works of art). Web archives are a novel form of cultural heritage institutions mandated to preserve similar artifacts. However, the artifacts of web archives are born-digital and digitized contents.

Web archives are a special type of digital libraries. Both share the responsibility of preserving information for future generations. This includes all types of multimedia, such as images and videos, besides the digital counterparts of printed documents. The main difference is that web archives usually grow to a data size that exceeds traditional organization and management of typical digital libraries. Digital libraries are based on meta-data describing manually curated artifacts and catalogs of these artifacts, which are usually used to explore and search digital collections, for instance, through faceted search. However, the experience from the Pandora (National Library of Australia)[6] and the Minerva (Library of Congress)[7] projects showed that this is not a viable option for

---

[3] E.g., Internet Archive available at http://www.archive.org.

[4] E.g., Library of Congress Web Archives available at http://www.loc.gov/minerva.

[5] http://www.thetimes.co.uk/tto/archive/.

[6] http://pandora.nla.gov.au.

[7] http://www.loc.gov/minerva.

World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
> Pointers to the world's online information, subjects , W3 servers, etc.

Help
> on the browser you are using

Software Products
> A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
> Details of protocols, formats, program internals etc

Bibliography
> Paper documentation on W3 and references.

People
> A list of some people involved in the project.

History
> A summary of the history of the project.

How can I help ?
> If you would like to support the web..

Getting code
> Getting the code by anonymous FTP , etc.

Fig. 1 A version of 1992 of the first web site. This earliest version found at CERN describes the world wide web project

web archives. The size of the web makes traditional methods for cataloging too time consuming and expensive, beyond the capability of libraries staff. One of the conclusions from the final report of the Minerva project is that automatic indexing should be the primary strategy for information discovery [19].

The first web site, presented in Fig. 1, was created by Tim Berners-Lee at the European Organisation for Nuclear Research (CERN) and published in August 1991. This site describes the basis of the world wide web and is back online at its original URL.[8] The first web archives appeared only in 1996 and do not contain sites prior to this date with the exception of some pages recovered from backups stored in floppy disks or CDs. The Internet Archive, a USA-based non-profit foundation, was one of the first web archives and has been broadly archiving the web since 1996. It leads the most ambitious initiative. In 2013, the Internet Archive was preserving 240 billion archived documents with a total of about 5 PB of data [20]. In 2014, it held 376 billion archived web pages, which represent 13.8 PB of data. The Pandora and Tasmanian web archives from Australia, and the Kulturarw3 web archive from Sweden, were also created in 1996. Many other initiatives followed since then and a significant effort has been employed by the research community in the web archiving domain. Many of these initiatives are members of the International Internet Preservation Consortium (IIPC), which leads the development of several open-source tools, standards and best practices for web archiving [21]. A time line of some of these initiatives can be obtained online.[9]

Previous initiatives archived a large number of web sites according to some selection policy. In addition to these, there are services that enable any person to permanently archive a web page given a URL, such as Perma.cc,[10] WebCitation[11] or Archive.is.[12] Each archived page receives a unique link, such as a Digital Object Identifier, to direct readers to its original version that will remain available online. Several user needs are met by these services, such as scholars preserving web pages cited in their work [22] or Supreme Courts preserving citations in their published decisions [23].

2.1 Data access

Much of the effort on web archive development focuses on acquiring, storing, managing and preserving data [19]. However, data must also be accessible to users who need to exploit and analyze them. Due to the challenge of indexing all the col-

---

[8] http://info.cern.ch/hypertext/WWW/TheProject.html.

[9] http://timeline.webarchivists.org.

[10] https://perma.cc/.

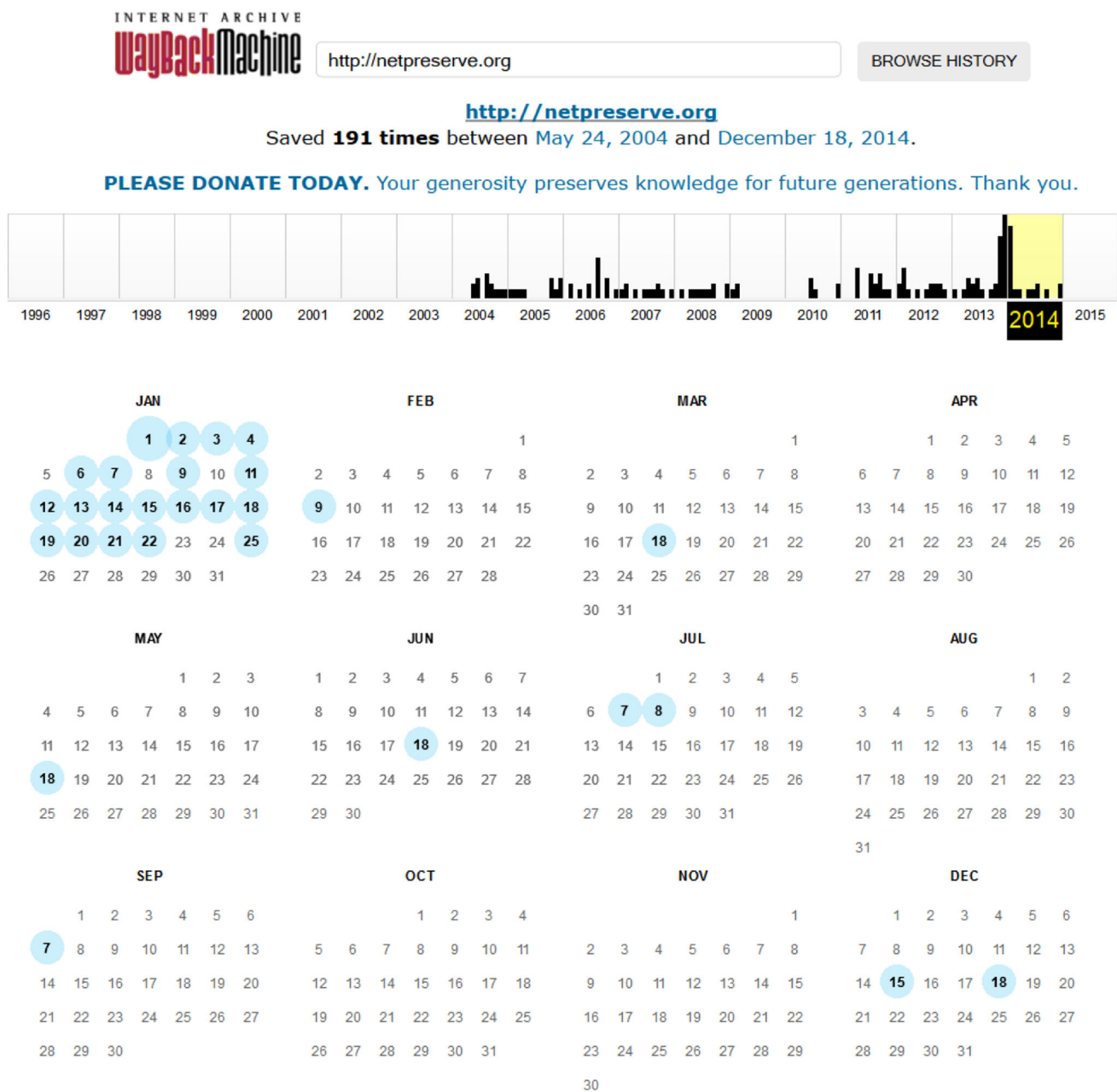[11] http://webcitation.org/.

[12] http://archive.is/.

Fig. 2 User interface of the Internet Archive's Wayback Machine

lected data, the prevalent discovery method in web archives is based on URL search, which returns a list of chronologically ordered versions for a given URL, such as in the Internet Archive's Wayback Machine [24,25]. Figure 2 depicts the user interface of the Wayback Machine after searching a URL. A survey on European web archives reported that 68 %

of web archives support this type of search [26]. However, URL search is limited, as it forces the users to remember the URLs, some of which refer to content that ceased to exist many years ago.

Another type of access is meta-data search, i.e., the search by meta-data attributes, such as category or theme. Meta-
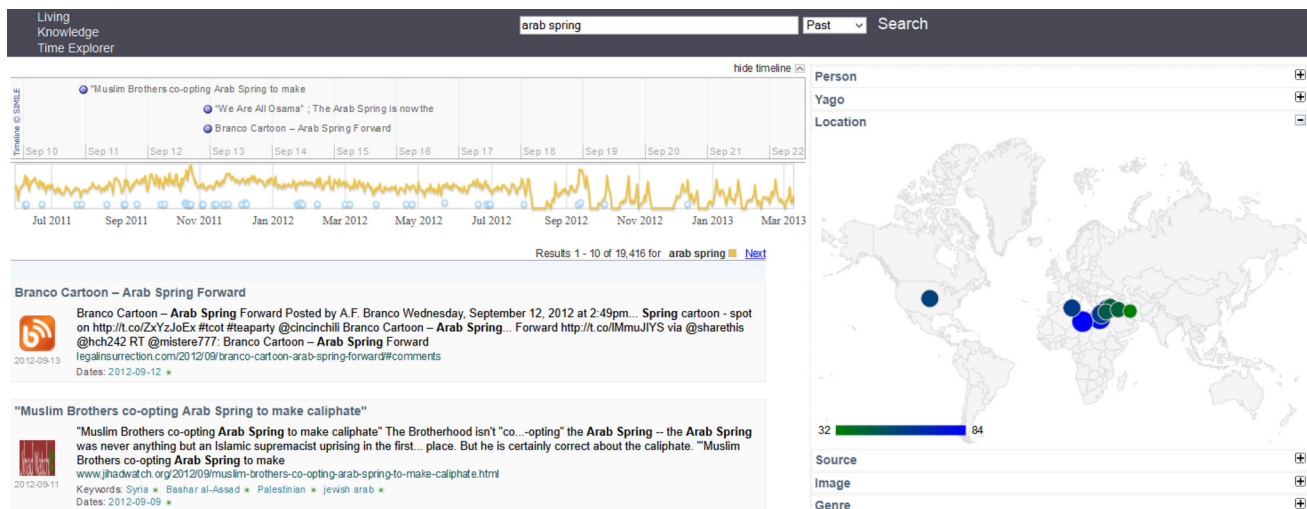
Fig. 3 Time Explorer application

data search is provided by 65 % of European web archives [26]. For instance, the Library of Congress Web Archives[13] supports search on bibliographic records. Some web archives support filtering results by domain and media type, while others organize collections by subject or genre to provide browsing functionality, such as the Pandora Australia's web archive [27]. Most web archives support narrowing the search results by date range.

Full-text search has become the dominant form of information discovery, especially in web search systems such as Google. These systems have a strong influence on the way users search in other settings. This explains why full-text search was reported as the most desired web archive functionality [28] and the most used when supported [29]. Despite the high computational resources required for this purpose, 70 % of the European web archives surveyed support full-text search for at least a part of their collections. Still, previous studies showed that the search services provided by these web archives are poor and frequently deemed unsatisfactory [18,30].

There are several access tools created for web archiving. The site[14] of the International Internet Preservation Consortium (IIPC) has a list with many tools for acquisition, curation, storage and access. Thomas et al. present a comprehensive list of available tools and services that can be used in web archives [31].

2.2 Data analysis

The existing search tools require a substantial human effort when exploring and analyzing complex topics. Hence, ana-

lytical tools are being researched to fulfill informational needs for specific users requiring richer answers such as historians or journalists [32,33]. Such tools would help to explain the stories of the past and predicting future events through the analysis and modeling of the evolution of data. Web archives are an exceptional data source to extract and leverage this evolution. A good example is the work of Leskovec et al. who tracked short units of information (e.g., phrases) from news as they spread across the web and evolve throughout time [34]. This tracking provided a coherent representation of the news cycle, showing the rise and decline of main topics in the media. Another example is the work of Radinsky and Horvitz who mined news and the web to predict future events [16]. For instance, they found a relationship between droughts and storms in Angola that catalyze cholera outbreaks. Anticipating these events may have a huge impact on world populations. Hoffart et al. built a large knowledge base in which entities, facts, and events are anchored in both time and space [35]. Web archives can be the source to extract these data, which will then be used for temporal analysis. For instance, since the veracity of facts is time dependent, it would be interesting to identify whether and when they become inaccurate.

Novel types of interfaces are also being researched to support data analysis over time. The Time Explorer, depicted in Fig. 3, combines several interfaces integrated in the same application designed for analyzing how topics evolve over time [36]. The core of the interface is a time line with the main titles extracted from the news and a frequency graph with the number of news and entities most frequently associated with a given query displayed over the time axis. The interface also displays a list of the most representative entities (people and locations) that occur on matching news and that can be used to narrow the search. The Zoetrope system

---

[13] http://www.loc.gov/webarchiving.

[14] http://www.netpreserve.org/web-archiving/tools-and-software.

also enables exploring archived data [37]. It introduces the concept of lenses that can be placed on any part of a web page to see all its previous versions. These lenses can be filtered by queries and time, and combined with other lenses to compare and analyze archived data (e.g., check traffic maps at 6 p.m. on rainy days). There are other examples, such as the visualization resources offered by the UK web archive,[15] which include *N*-gram charts of the occurrence of terms or phrases over time and tag clouds of content written on web sites. Browser plug-ins that highlight changes between pages, such as the DiffIE Add-on for Internet Explorer, are also of great help for data analysis [38].

2.3 Research projects

Several research projects have been initiated for improving web archiving technologies. The Living Web Archives (LiWA) aimed to provide contributions to make archived information accessible and addressed IR challenges, such as web spam detection, terminology evolution, capture of stream video, and assuring temporal coherence of archived content [39]. LiWA was followed by the Longitudinal Analytics of Web Archive data (LAWA), which aimed to build an experimental testbed for large-scale data analytics [40]. Particular emphasis is given to developing tools for aggregating, querying and analyzing web archive data that have been crawled over extended time periods. The Web Archive Retrieval Tools (WebART) project focus on the development of web archive access tools specifically tailored to facilitate research in humanities and social sciences [41]. The Collect-all ARchives to COmmunity MEMories (ARCOMEM) project was about developing innovative tools and methods to help preserve and exploit the social web [42], while the SCAPE project[16] addressed solutions for large-scale digital preservation. The Memento project adds a temporal dimension to the HTTP protocol so that archived versions of a document can be served by the web server holding that document or by existing web archives if the web server does not contain the requested versions [43]. Users only have to install a browser plug-in, which makes this an easy solution to adopt. Users can also search via the Time Travel portal[17] across several web archives. This portal works as a metasearch engine for web archives. Old versions of web pages can be reconstructed by combining parts returned by web archives that support the Memento's API, which enables the integration of archived content and cooperation among web archives.

3 Methodology

During October 2010, we gathered information from web archiving initiatives across the globe [17]. We read the official sites of known web archive initiatives and published documentation, but had little success because the published information was frequently insufficient or obsolete. Plus, many official sites were exclusively available in the native language of the hosting country (e.g., Chinese) and automatic translation tools were insufficient to obtain the required information. Thus, we decided to contact directly the community to obtain answers to the following questions:

1. What is the name of your web archiving 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)?

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 International Internet Preservation Consortium [21], National Library of Australia in its Preserving Access to Digital Information (PADI) page[18] and International Web Archiving Workshops.[19] We were able to establish contact and obtain direct answers from 33 web archiving initiatives. Finally, we distributed the collected data among 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 clarify our intents and obtain the required 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, we assumed in the third question 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 content in bulk, such as files in ARC format [44]. The post hoc statistical analysis of the obtained answers 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 more accurate information and valuable insights about web archiving initiatives
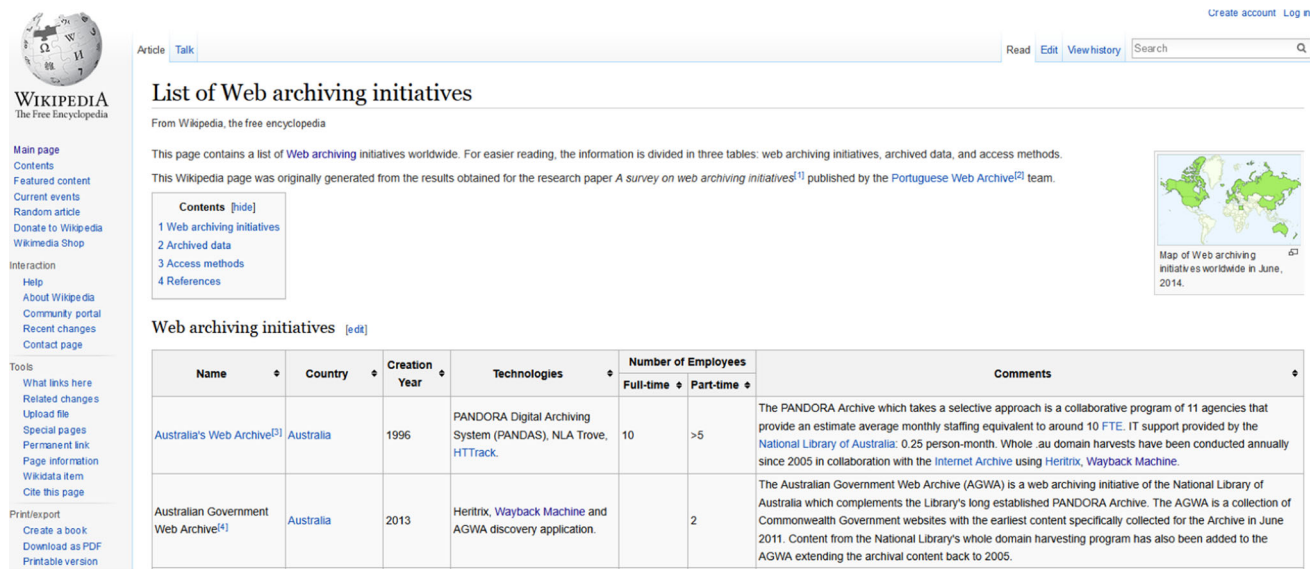
---

[15] http://www.webarchive.org.uk/ukwa/visualisation.

[16] http://www.scape-project.eu.

[17] http://timetravel.mementoweb.org.

[18] http://www.nla.gov.au/padi.

[19] http://iwaw.europarchive.org.

Fig. 4 Wikipedia page with list of web archiving initiatives

than a typical one-shot online survey with closed answers. However, the cost of processing the results for statistical analysis was significantly higher.

This survey was published in 2011 [17]. The data collected and validated enabled the creation of a Wikipedia page named *List of Web Archiving Initiatives*,[20] so that the published information could be collaboratively kept up-to-date. Since then, the web archiving community has been updating this information, making it a useful resource. Figure 4 shows the Wikipedia page that contains three tables populated with information about the web archiving initiatives, such as their name, country, creation year, employed technologies, number of employees, number and volume of archived contents, archived formats, type of crawl and access methods.

To observe how web archiving changed since the first survey, in 2014 we conducted the same analysis on the data published in the Wikipedia page and compared it against the results of 2010. In case of doubt or lack of information, we consulted the official sites of the initiatives or their scientific publications.

3.1 Comparison with other surveys

After our first survey in 2010, three other surveys were conducted on web archiving which obtained related information, such as the access type provided by the initiatives and the technology used to support them. The first survey was conducted by the Internet Memory Foundation over European web archives in 2010, from now on referred to as the IMF2010 survey [26]. The second and third surveys were conducted by the National Digital Stewardship Alliance

(NDSA) in 2011 and 2013, and they covered organizations of the USA involved or planning to archive content from the web [45,46]. These surveys are referred to from now on as the NDSA2011 and NDSA2013 surveys. In this paper, we analyze and compare the results of the surveys whenever possible, despite our surveys having covered world wide web archiving initiatives, while the IMF2010 survey focused just on initiatives from Europe and the NDSA surveys on initiatives from the USA. Still, all surveys took place between 2010 and 2014, which makes their results comparable in time.

4 Results

4.1 Web archiving initiatives

Table 1 shows general statistics about web archiving initiatives surveyed in 2010 and 2014. Web archiving initiatives are very heterogeneous in size and scope. For instance, the web archive (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 web archives exclusively hold content related to their hosting country, region or institution. However, there are a few initiatives, such as the Internet Memory Foundation and the Portuguese Web Archive, that also preserve information related to foreign countries.

Table 1 General statistics of web archiving initiatives

Characteristics	2010	2014	Δ (%)
Total initiatives	42	68	+61.9
Countries hosting initiatives	26	33	+26.9

---

[20] http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives.

We detected an increase in the number of web archiving initiatives, from 42 in 2010 to 68 in 2014. Since the creation and operation of a web archive is complex and costly, several initiatives exist to provide web archiving services (WAS) that can be independently operated by third-party archivists to harvest, build and preserve collections of digital content. These WAS enable focused archiving of web content by organizations, such as universities or libraries, that otherwise could not manage their own archives. In 2014, there were 11 initiatives (16 %) providing WAS against the previous 3 (7 %) offered in 2010. Some of these new WAS are the Aleph Archives,[21] Hanzo Archives[22] and Reed Archives.[23] The oldest WAS are the Archive-It,[24] ArchiveTheNet[25] and Web Archiving Service.[26] Of the 11 WAS, 6 operate in the USA, where most of them offer electronic discovery (ediscovery) services for enterprises, which are required by law since 2006 for the discovery of information in civil litigation or government investigations. In 2014, at least 19 % of the initiatives were using WAS. In 2010, this percentage was 16 %.

4.1.1 Human resources

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 of 2010 show that web archiving engaged at least 112 people in full time and 166 in part-time. The web archive teams were 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). The staff was mostly composed of librarians and computer engineers. The results show that 11 initiatives (26 %) did not have any person dedicated full time. The effort of part-time workers was variable, for instance, at the Library of Congress they spent only a few hours a month. Most of the human resources were invested on data acquisition and quality control. The IMF2010 survey corroborates that web archive teams are small, but the number of staff depends on the phase of the project. Its results show that 38 % of fully operational initiatives count more than five full-time employees, while 67 % that started a project count between two and five employees.

[21] http://aleph-archives.com/.

[22] http://www.hanzoarchives.com/.

[23] http://www.reedarchives.com/.

[24] http://www.archive-it.org.

[25] http://archivethe.net.

[26] http://webarchives.cdlib.org.

Table 2 Staff statistics of web archiving initiatives

Characteristics	2010	2014	Δ
Total people (full time)	112	108	−3.6 %
Total people (part-time)	166	197	+18.7 %
Total people	278	305	+9.7 %
Median people (full time)	2.5	2	−20.0 %
Median people (part-time)	2	2	0.0 %
Average people (full time)	3.5	2.2	−37.1 %
Average people (part-time)	5	4	−20.0 %

In 2014, the size of the teams continued to be highly variable, where initiatives had teams without any person working in full time, such as the University of Texas at San Antonio WA, while other teams had 12 people working in full time, such as the Internet Archive, or 80 people working in part-time, such as the Library of Congress. As shown in Table 2, in 2014, the web archiving initiatives had in total 108 people working in full time and 197 in part-time. There was an increase from 278 to 305 people working in this area. The teams continued to be mostly small, having a median staff of 2 people in full time (average of 2.2) and 2 people in part-time (average of 4). There were 3 initiatives that did not have any person dedicated full time, against the 11 of 2010. Despite the large increase of the number of initiatives, the total number of people working on them increased only slightly, which led to a decrease in the median and average team size. The NDSA2013 survey shows a different reality with less people working in web archiving. The USA initiatives have a median staff of 0.25 people in full time. Only 19 % of the USA initiatives devote at least one person to handle web archiving tasks. The small size of the teams are likely due to the high percentage of initiatives that use WAS instead of running their own web archiving system.

4.1.2 Geographic location

Figure 5a presents the countries that hosted web archiving initiatives in 2010. The 42 initiatives were spread across 26 countries. There were 23 initiatives hosted in Europe, 10 in North America, 6 in Asia and 3 in Oceania. Half of the initiatives were 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 %) hosted at least one web archiving initiative, which is an indicator of the importance of web archiving in developed countries. Most of the countries hosted one (74 %) or two initiatives (22 %). The only country that hosted more than two was the USA with a total of nine initiatives. Although being part of a country, initiatives like the Tasmanian WA (Australia), North Carolina WA (USA) or Digital Heritage

Fig. 5 Countries hosting web archiving initiatives in **a** 2010 and **b** 2014 (in *green*) (color figure online)


(a)

(b)
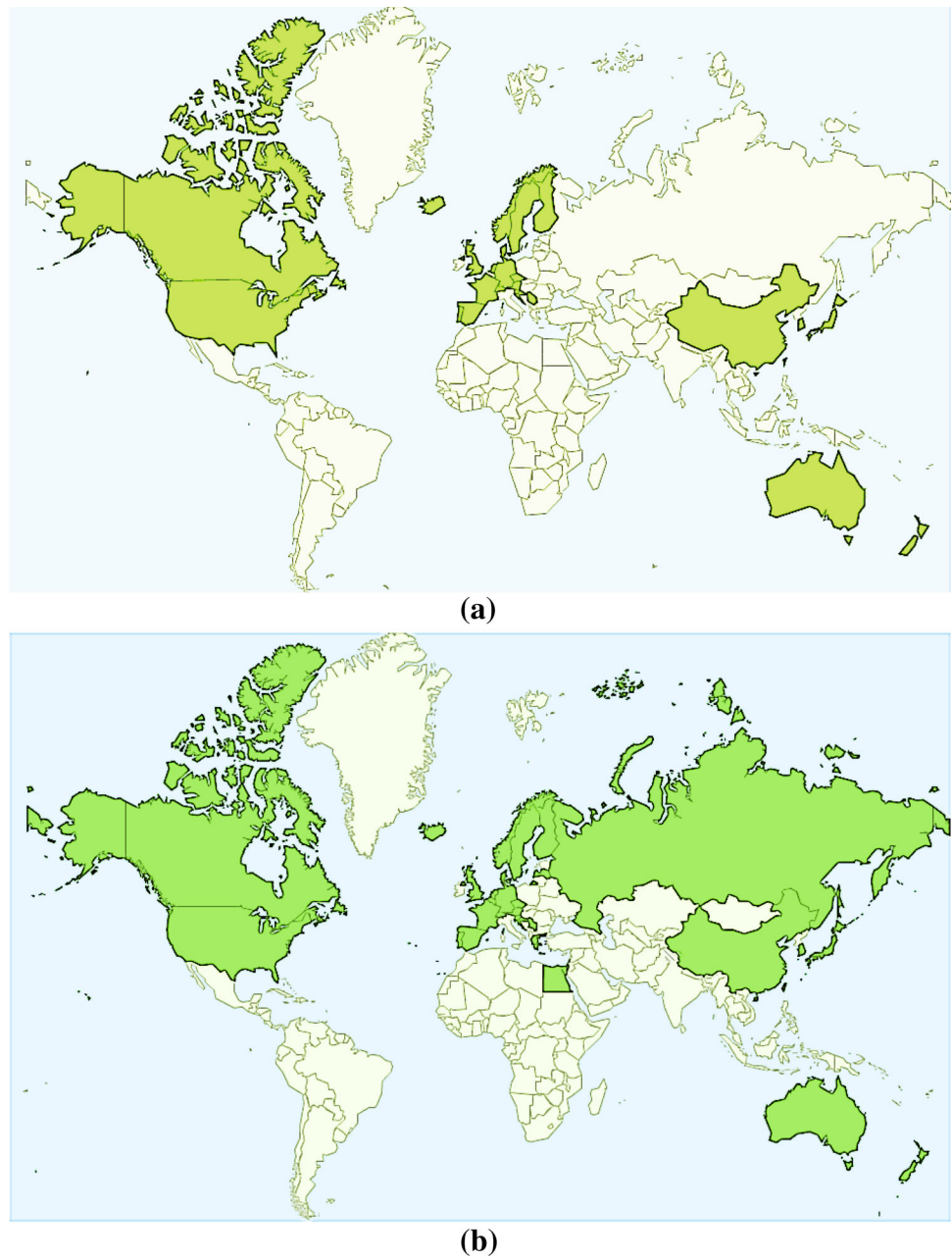
Catalonia (Spain) were hosted at autonomous states and aimed at preserving regional content.

Figure 5b presents the location of all countries hosting web archiving initiatives in 2014. The 68 web archiving initiatives are spread by 33 countries. In 2010, there were only 26 countries hosting web archiving initiatives, which shows a growing awareness of the importance of web archiving all over the world. The USA continues to be the country with the most initiatives, increasing from 9 in 2010 to 19 in 2014. The second country with most initiatives is France, with five initiatives. Germany and Switzerland share the third place with four initiatives each. The distribution of the initiatives over the world is 38 in Europe (previously 23), 22 in North Amer-

ica (previously 10), 8 in Asia (previously 6), 3 in Oceania (equal) and 1 in Africa (previously 0). Notice that some initiatives have more than one location. There were increases in almost all continents, especially in Europe and North America. Africa received its first initiative hosted in Egypt, while South America does not have any yet.

When comparing the number and location of initiatives with other surveys, we detected that many were missing. The IMF2010 survey found 41 European initiatives fully operational in 2010, while we found 38 in 2014. The NDSA2011 and NDSA2013 surveys found 49 and 64 active initiatives in the USA, but we found only 19 in 2014. This difference is mostly due to college and universities, i.e., 36 in 2011 and

Fig. 6 Cumulative number of initiatives created per year
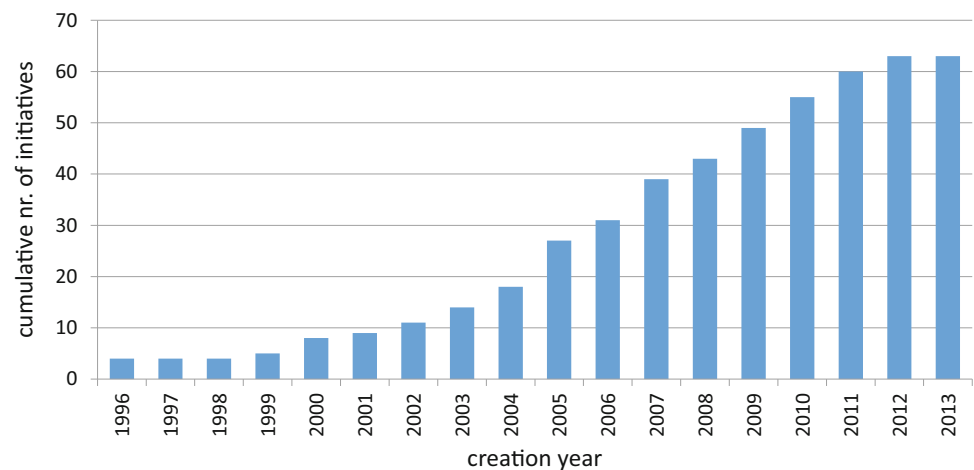
48 in 2013, included in the NDSA surveys and that were not included in our surveys. Future surveys should make an effort to cover all these initiatives. Nevertheless, both NDSA and our surveys show a growing trend of initiatives.

4.1.3 Growth

Figure 6 displays the evolution of the number of web archiving initiatives created per year, including the new initiatives recorded on the Wikipedia page. There was a growth from 4 initiatives in 1996 to 14 initiatives in 2003, which represents an average of 1.8 new initiatives per year. After 2003, many new initiatives appeared to solve the web ephemerality problem. For instance, in 2005 and 2007, nine and eight initiatives were created, respectively. There was an average growth of 5.4 initiatives per year from 2004 to 2012. There is no information on new initiatives created in 2013. One possible explanation for the significant and constant growth since 2003 was the concern raised by the United Nations Educational, Scientific and Cultural Organization (UNESCO) regarding the preservation of the digital heritage [4]. The NDSA2013 survey also shows a constant growth, especially between 2006 and 2013, when there was a great increase of initiatives mainly due to universities starting their web archiving programs. Universities created 39 (out of 67) initiatives during these 8 years, which indicates an emergent awareness in the academic community of the USA about the importance of preserving web content.

4.2 Archived data

4.2.1 Selection policy

Since the resources are scarce and not all the web can be preserved, the selection policy of most web archiving initiatives is to preserve the most relevant parts of the web from their own perspective. In the survey of 2010, all web archives selected specific sites for archiving. This selection is determined by multiple factors such as consent by the authors or relevance for inclusion in thematic collections (e.g., elections or natural disasters). However, 80 % of the web archives exclusively held content related to their hosting country, region or institution. Of the 42 initiatives, 11 (26 %) also performed broad crawls of the web, including all sites hosted under a given domain name or geographical location. The IMF2010 survey reported that 23 % of European web archives run domain crawls, while 71 % performed thematic or selective crawls. The NDSA2011 survey reported that all USA initiatives archived web content from their own institution, as well as content from other organizations or individuals for future research.

Our results show that in 2014, at least 45 initiatives (66 %) performed selective crawls and 20 (29 %) country code top-level domain (ccTLD) or broad crawls of the web. Almost all initiatives continue to exclusively hold content related to their hosting country, region or institution. There are three initiatives that archive ccTLD of other countries besides their own. The Internet Archive and the Internet Memory Foundation share a vision to preserve web content from all over the world. The Portuguese Web Archive preserves content from 4 countries that have Portuguese as their official language.

4.2.2 Volume size

Figure 7 presents the distribution of the size of archived collections measured in total volume of data and number of contents. Notice that one HTML page containing three embedded images results in the archive of four contents. There was an increase of initiatives with collections between 10 and 100 TB in detriment of collections between 1 and 10 TB. While in 2010, 50 % of the initiatives preserved collections smaller than 10 TB and 31 % preserved collections between 10 and 100 TB, in 2014 these percentages were 42 and 40 %, respectively. The percentage of initiatives with

Fig. 7 Size of archived collections measured in: **a** volume of data (terabytes) and **b** number of contents (e.g., images, pages, videos)

(a)

(b)
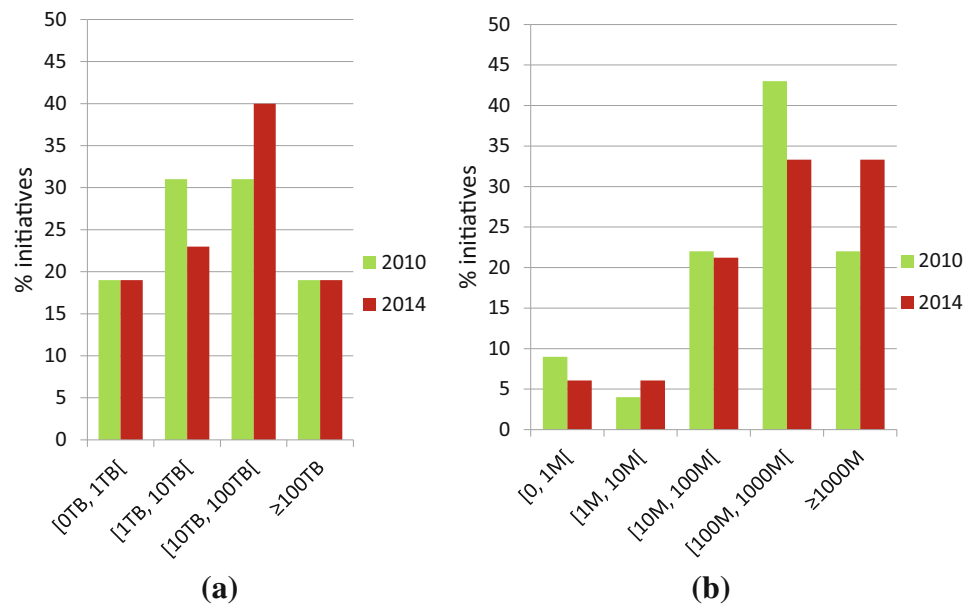
collections larger than 100 TB continues to be 19 %. In accordance with this finding, the percentage of initiatives with collections between 100 and 1000 million contents decreased from 43 to 33 %, mostly because the percentage of initiatives with collections with more than 1000 million contents increased from 22 to 33 %.

World wide web archives preserved from 1996 to 2010 a total of 181,978 million contents (6.6 PB). The Internet Archive by itself held 150,000 million contents (5.5 PB). In 2014, all initiatives had archived together at least 534,604 million contents, which sums around 17 PB of data. This represents an increase from 2010 to 2014 of 294 % on contents and 258 % on volume of data. The Internet Archive continue to be by far the web archive with the largest collection with 376,000 million contents. The information of its volume of data was not available in the Wikipedia page. Hence, we extrapolated from the 2010 results and estimated 13.8 PB of data.

The selection policies of some initiatives intersect, which leads to a replication of archived content [47,48]. For instance, initiatives hosted in the same country may preserve some of the same sites. Initiatives with a broader scope, such as the Internet Archive, preserve some content that are also archived by national initiatives. The overlap of archived content is not contemplated in this paper.

4.3 Access and technologies

4.3.1 Formats to store archived content

Figure 8 presents the evolution of file formats used to store archived content. The ARC format defined by the Internet Archive was the *de facto* standard in 2010 [44]. In 2009, the
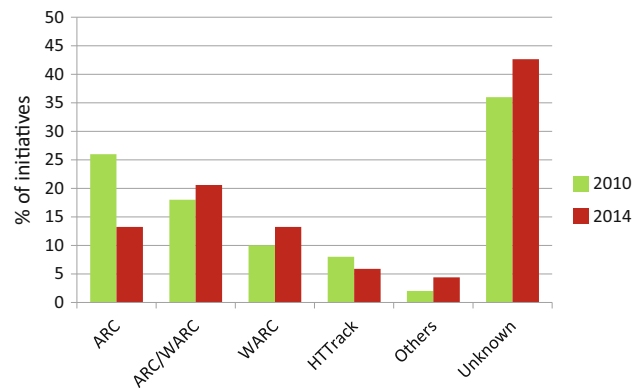

Fig. 8 Usage of file formats to store web content

WARC format was published by the International Organization for Standardization (ISO) as the official standard format for archiving web content and it was exclusively used by 10 % of the initiatives in 2010 [49]. The ARC and WARC formats were dominant in 2010, being used by 54 % of the initiatives.

There was a decrease, from 26 % in 2010 to 13 % in 2014, of initiatives using exclusively the ARC format. These initiatives likely changed to the WARC format that increased 3 % points and the ARC/WARC formats that also increased 3 % points. The ARC and WARC formats continue to be by far the most predominant, being used in 2014 by 47 % of web archiving initiatives against the 54 % in 2010. Besides historical reasons, the widespread of the ARC/WARC formats was motivated by the Archive-Access project, which freely provides open-source tools to process this type of files [50]. There are only 10 % of initiatives using other file formats in 2014, such as the HTTrack format. Still, 43 % of the initiatives did not report the adopted format in the Wikipedia page.
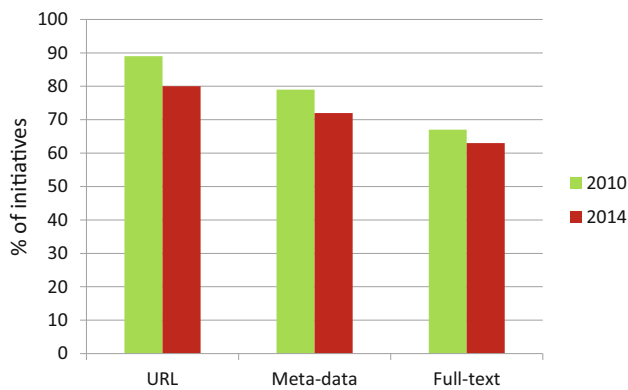
Fig. 9 Search methods provided by web archives

4.3.2 Search methods

Figure 9 presents the search methods provided by the initiatives over their collections in 2010 and 2014. The obtained results of 2010 showed that 89 % of the initiatives support search over multiple versions of a given URL published over time, 79 % enable searching through meta-data and 67 % provide full-text search over archived contents. These results differ from the IMF2010 survey, which reported 68, 65 and 70 % of European initiatives supporting URL, meta-data and full-text search, respectively. The percentage of European web archives offering URL and meta-data search are significantly lower, but slightly higher in full-text search. The NDSA surveys show similar results in 2011 and 2013. The URL search and full-text search are also the most provided search methods. The NDSA surveys reported other methods frequently used, such as browsing by URL and title.

Our results of 2014 are almost the same as in 2010, with a small relative decrease in all search methods. The most predominant is the search by URL, then the search by meta-data and last, by full-text search. There were two initiatives that provided full text, but only to a part of their collections (one 30 % and the other 15 %).

4.3.3 Access restrictions

In 2010, some initiatives held the copyright of the archived contents (e.g., German Bundestag, Canada WA) or explicitly required the consent of the authors before archiving (e.g., UK WA, OASIS of Korea). The Tasmanian WA operated since its inception under the assumption that web sites fall within the definition of books. Thus, no permission to capture from publishers was required. The Internet Archive and the Portuguese Web Archive proactively archive and provide access to contents, but remove access on-demand. On the other hand, for 16 initiatives (38 %) the access to collections was somehow restricted. The Library of Congress, WebArchiv of Czech Republic and Australia Web Archive provided pub-

lic online access to part of their collections. Netarkivet.dk of Denmark provided online access on-demand only for research purposes. The Finnish Web Archive provided online access to meta-data, but not to archived contents. The Bibliothèque nationale de France (BnF), Web@rchive of Austria and Preservation .ES of Spain, granted access exclusively through special rooms on their facilities.

The IMF2010 survey found that 50 % of the European initiatives performed web archiving protected by a law enacted or passed. Regarding the policy for accessing archived data, 41 % of the initiatives provided access for everyone, 28 % online access with restrictions, 18 % on-site access for anyone, 21 % on-site access with restrictions and 21 % did not provide any access of their contents. The NDSA2013 survey indicates that when proving public access to archived web content, 63 % of the USA initiatives neither notified nor sought permission from content owners, 15 % notified content owners, and 21 % sought permission.

The information available on the Wikipedia page about the access restrictions is not sufficient for a statistical analysis. Still, some initiatives recorded their restrictions. The WebArchiv of Czech Republic provides unlimited access only from public terminals in the National Library. The Chinese Web Archive and the Web@rchive of Austria provide access to content in their National Libraries. The Finnish Web Archive also provides on-site access to contents. For the Netarkivet.dk of Denmark, the online access is granted only to researchers and the BnF Web Legal Deposit of France grants access only to authorized users.

4.3.4 Technology

Figure 10 depicts the technologies being used by the initiatives that manage their own systems. In 2010, the Archive-Access tools were dominant (62 %), including the Heritrix, NutchWAX and Wayback Machine projects that support content harvesting, full-text and URL search,
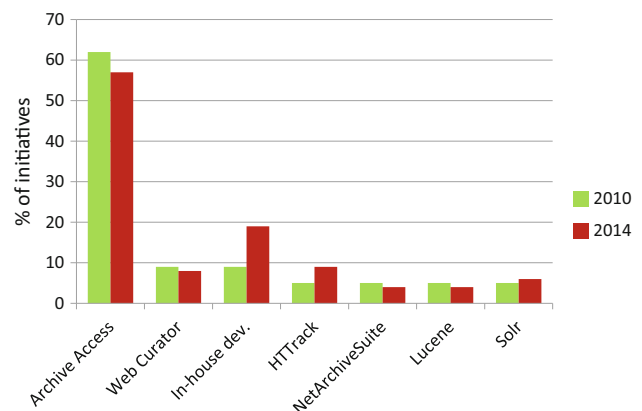

Fig. 10 Technologies used by web archives

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, in 2010, NutchWAX supported full-text search for the Finnish Web Archive (148 million), Canada Web Archive (170 million), Digital Heritage of Catalonia (200 million), California Digital Library (216 million) and BnF (15 % of a collection of 200TB). The IMF2010 survey shows that the European initiatives used similar tools. They used Heritrix to crawl web content (80 %), and for search, they used the Wayback Machine (67.5 %) or NutchWAX (70 %).

Despite the increase from 3 in 2010 to 11 in 2014 of web archive services (WAS), the number of initiatives that used WAS increased just 3 % points, from 16 to 19 %. The Archive-It is the service most used, summing a total of seven initiatives. There was an increase from 9 to 19 % of initiatives doing some in-house development. This software was mostly developed by WAS, such as the Hanzo Archives' access tools, or curation tools developed by libraries, such as the DigiBoard of the Library of Congress Web Archives. These increases contributed to the decrease of the use of Archive-Access tools. Still, the Archive-Access tools continue to predominate, with 57 % of the initiatives using at least one of its tools in 2014, against the 62 % in 2010. Lucene and Solr together continue to be used by 10 % of the initiatives with a growing trend toward Solr.

The NDSA surveys show different results, where the USA initiatives contracted much more WAS. There were 60 % of initiatives in 2011 and 63 % in 2013 that exclusively used WAS. Archive-It is the dominant external service used by approximately 70 % of the initiatives and the California Digital Library WAS is the second most used with 17 %. Regarding technology to capture web content, Heritrix is the most used tool by USA initiatives (29 %), followed by HTTrack (18 %). The Wayback Machine increased from 76 % in 2011 to 89 % in 2013 as the preferred tool to view contents.

5 Conclusion

Web archiving has been gaining interest and recognition from modern societies around the world. Still, there is a lack of knowledge in the research community about the most recent developments in web archiving and the existing initiatives. This paper provides an updated global overview on these issues and discusses evolution trends.

Based on two conducted surveys, we observed that web archiving initiatives are typically hosted by developed countries, but we can find them spread all over the world in almost every continent. Web archives are generally composed of small teams that mainly work on the acquisition and curation of data. Almost all initiatives exclusively hold content related to their hosting country, region or institution, which stresses the need for each country to finance at least one initiative at national level.

Web archiving initiatives have been in existence since 1996 and their number has been growing since then. Particularly, from 2010 to 2014 there was a large increase in the number of initiatives, hosting countries, number of contents and volume of archived data. Currently, web archiving initiatives hold 17 PB (534,604 million contents), which shows a growing awareness of the importance of web archiving all over the world and a continued effort of the community in mitigating the web ephemerality problem.

On the other hand, despite the social and economic impact of losing the information that is being exclusively published on the web, the obtained results show that the human resources invested in web archiving are still scarce and the size of teams are even decreasing. The lack of resources will probably originate a historical void in the future about our current time. Our results already show that only a small part of the web has been preserved.

The web archiving community is adopting common data formats and tools. The ARC and WARC are the predominant data formats to store archived content, but in the last years there was a shifting from ARC to WARC likely to take advantage of the new format enhancements, which enables, for instance, to manage duplicated content and record contextual meta-data. Regarding technology, most initiatives continue to use Lucene-based solutions to support full-text search, such as NutchWAX or Solr, the Wayback Machine to support URL search and display archived content, and Heritrix to crawl web content. This continuity could be explained by the significant number of developers and web archive initiatives that contribute to enhance these projects.

The predominant methods for discovering archived content have remained the URL, meta-data and full-text search. However, the respondents of the surveys mentioned that the existing technology provides unsatisfactory search results and full text, which is the preferred method by the users, is hard to implement. Moreover, recent studies show that these technologies provide poor search results, making difficult for users to find the desired information. With the fast growth of archived data, this problem is only exacerbated. Hence, the development of efficient and effective search technology is urgent to access the massive data already stored in web archives.

References

1. Ntoulas, A., Cho, J., Olston, C.: What's new on the web? The evolution of the web from a search engine perspective. In: Proc. of the 13th International Conference on World Wide Web, pp. 1–12 (2004)

2. Dellavalle, R., Hester, E., Heilig, L., Drake, A., Kuntzman, J., Graber, M., Schilling, L.: Going, going, gone: lost internet references. Science **302**(5646), 787–788 (2003)

3. SalahEldeen, H., Nelson, M.: Losing my revolution: how many resources shared on social media have been lost? In: Theory and Practice of Digital Libraries, pp. 125–137 (2012)

4. UNESCO: Charter on the preservation of digital heritage. In: Adopted at the 32nd Session of the General Conference of UNESCO (2003). http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf. Accessed 17 Oct 2003

5. UNESCO: Universal declaration on archives. In: Adopted at the ICA Annual General Meeting in Malta (2010). http://www.ica.org/6573/reference-documents/universal-declaration-on-archives.html. Accessed 17 Sept 2010

6. Kitsuregawa, M., Tamura, T., Toyoda, M., Kaji, N.: Socio-sense: a system for analysing the societal behavior from long term web archive. In: Proc. of the 10th Asia-Pacific Web Conference on Progress in WWW Research and Development, pp. 1–8 (2008)

7. Arms, W.Y., Aya, S., Dmitriev, P., Kot, B., Mitchell, R., Walle, L.: A research library based on the historical collections of the Internet Archive. D-Lib Mag. **12**(2) (2006)

8. Arms, W., Huttenlocher, D., Kleinberg, J., Macy, M., Strang, D.: From Wayback Machine to Yesternet: new opportunities for social science. In: Proc. of the 2nd International Conference on e-Social Science (2006)

9. Ackland, R.: Virtual observatory for the study of online networks (VOSON)—progress and plans. In: Proc. of the 1st International Conference on e-Social Science (2005)

10. Foot, K., Schneider, S.: Web Campaigning. The MIT Press, Cambridge (2006)

11. Franklin, M.: Postcolonial Politics, the Internet, and Everyday Life: Pacific Traversals Online. Routledge (2004)

12. Gomes, D., Costa, M.: The importance of web archives for humanities. Int. J. Humanit. Arts Comput. **8**(1), 106–123 (2014)

13. Yamamoto, Y., Tezuka, T., Jatowt, A., Tanaka, K.: Honto? Search: estimating trustworthiness of web information by search results aggregation and temporal analysis. In: Advances in Data and Web Management, pp. 253–264 (2007)

14. Chung, Y., Toyoda, M., Kitsuregawa, M.: A study of link farm distribution and evolution using a time series of web snapshots. In: Proc. of the 5th International Workshop on Adversarial Information Retrieval on the Web, pp. 9–16 (2009)

15. Elsas, J., Dumais, S.: Leveraging temporal dynamics of document content in relevance ranking. In: Proc. of the 3rd ACM International Conference on Web Search and Data Mining, pp. 1–10 (2010)

16. Radinsky, K., Horvitz, E.: Mining the web to predict future events. In: Proc. of the 6th ACM International Conference on Web Search and Data Mining, pp. 255–264 (2013)

17. Gomes, D., Miranda, J., Costa, M.: A survey on web archiving initiatives. In: Proc. of the International Conference on Theory and Practice of Digital Libraries, pp. 408–420 (2011)

18. Costa, M., Couto, F.M., Silva, M.J.: Learning temporal-dependent ranking models. In: Proc. of the 37th Annual ACM SIGIR Conference (2014)

19. Masanès, J.: Web Archiving. Springer, New York (2006)

20. Kahle, B.: Wayback machine: now with 240,000,000,000 (2013). http://blog.archive.org/2013/01/09/updated-wayback/. Accessed 30 Apr 2016

21. Grotke, A.: IIPC—2008 member profile survey results. Technical report, International Internet Preservation Consortium (IIPC) (2008)

22. Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., Tobin, R.: Scholarly context not found: one in five articles suffers from reference rot. PloS One **9**(12), 1–39 (2014)

23. Lazun, M.J.: "Link Rot" and legal resources on the web: a 2013 analysis by the chesapeake digital preservation group. Technical Report, The Chesapeake Digital Preservation Group (2013)

24. Tofel, B.: 'Wayback' for accessing web archives. In: Proc. of the 7th International Web Archiving Workshop (2007)

25. Jaffe, E., Kirkpatrick, S.: Architecture of the Internet Archive. In: Proc. of SYSTOR 2009: The Israeli Experimental Systems Conference, pp. 1–10 (2009)

26. Internet Memory Foundation: Web archiving in Europe. Technical Report, Internet Memory Foundation (2010)

27. Niu, J.: Functionalities of web archives. D-Lib Mag. **18**(3/4) (2012)

28. Ras, M., van Bussel, S.: Web archiving user survey. Technical Report, National Library of the Netherlands (Koninklijke Bibliotheek) (2007)

29. Costa, M., Silva, M.J.: Characterizing search behavior in web archives. In: Proc. of the 1st International Temporal Web Analytics Workshop, pp. 33–40 (2011)

30. Costa, M., Silva, M.J.: Evaluating web archive search systems. In: Proc. of the 13th International Conference on Web Information Systems Engineering, pp. 440–454 (2012)

31. Thomas, A., Meyer, E.T., Dougherty, M., Van den Heuvel, C., Madsen, C., Wyatt, S.: Researcher engagement with web archives: challenges and opportunities for investment. Technical Report, Joint Information Systems Committee (JISC) (2010)

32. Spaniol, M., Masanès, J., Baeza-Yates, R.: The 5th temporal web analytics workshop (tempweb'15). In: Proc. of the Companion Publication of the 24th International Conference on World Wide Web, pp. 863–864 (2015)

33. Spaniol, M., Masanès, J., Baeza-Yates, R.: The 4th temporal web analytics workshop (tempweb'14). In: Proc. of the Companion Publication of the 23rd International Conference on World Wide Web, pp. 863–864 (2014)

34. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 497–506 (2009)

35. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: Yago2: a spatially and temporally enhanced knowledge base from wikipedia. Artif. Intell. **194**, 28–61 (2013)

36. Matthews, M., Tolchinsky, P., Blanco, R., Atserias, J., Mika, P., Zaragoza, H.: Searching through time in the New York Times. In: Proc. of the 4th Workshop on Human–Computer Interaction and Information Retrieval, pp. 41–44 (2010)

37. Adar, E., Dontcheva, M., Fogarty, J., Weld, D.S.: Zoetrope: interacting with the ephemeral web. In: *Proc. of the 21st Annual ACM Symposium on User Interface Software and Technology*, pp. 239–248 (2008)

38. Teevan, J., Dumais, S., Liebling, D., Hughes, R.: Changing how people view changes on the web. In: Proc. of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 237–246 (2009)

39. Masanès, J.: LiWA news #3: living web archives (2011). http://liwa-project.eu/images/videos/Liwa_Newsletter-3.pdf. Accessed March 2011

40. Weikum, G., Ntarmos, N., Spaniol, M., Triantafillou, P., Benczur, A.A., Kirkpatrick, S., Rigaux, P., Williamson, M.: Longitudinal analytics on web archive data: it's about time! In: Proc. of the 5th Conference on Innovative Data Systems Research, pp. 199–202 (2011)

41. Huurdeman, H.C., Ben-David, A., Sammar, T.: Sprint methods for web archive research. In: Proc. of the 5th Annual ACM Web Science Conference, pp. 182–190 (2013)

42. Risse, T., Peters, W.: ARCOMEM: from collect-all ARchives to COmmunity MEMories. In: Proc. of the 21st International Conference Companion on World Wide Web, pp. 275–278 (2012)

43. Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L.L., Ainsworth, S., Shankar, H.: Memento: time travel for the web. CoRR (2009). arXiv:0911.1112

44. Burner, M., Kahle, B.: Arc file format (1996). http://www.archive.org/web/researcher/ArcFileFormat.php. Accessed Sept 1996

45. NDSA Content Working Group: Web archiving survey report. Technical Report, National Digital Stewardship Alliance (2012)

46. Bailey, J., Grotke, A., Hanna, K., Hartman, C., McCain, E., Moffatt, C., Taylor, N.: Web archiving in the United States: a 2013 survey. Technical Report, National Digital Stewardship Alliance (2014)

47. Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How much of the web is archived? In: Proc. of the 11th Annual International ACM/IEEE joint Conference on Digital Libraries, pp. 133–136 (2011)

48. AlSum, A., Weigle, M.C., Nelson, M.L., Van de Sompel, H.: Profiling web archive coverage for top-level domain and content language. Int. J. Digit. Libr. **14**(3–4), 149–166 (2014)

49. ISO 28500:2009: Information and documentation—WARC file format (2009). http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717. Accessed 30 Apr 2016

50. IIPC: Internet Archive ARC access tools (2009). http://archive-access.sourceforge.net/. Accessed 30 Apr 2016