III. DIGITAL PRESERVATION

THE IMPORTANCE OF WEB ARCHIVES FOR HUMANITIES

DANIEL GOMES AND MIGUEL COSTA

Abstract *The web is the primary means of communication in developed societies. It contains descriptions of recent events generated through distinct perspectives. Thus, the web is a valuable resource for contemporary historical research. However, its information is extremely ephemeral. Several research studies have shown that only a small amount of information remains available on the web for longer than one year.*

Web archiving aims to acquire, preserve and provide access to historical information published online. In April 2013, there were at least sixty four web archiving initiatives worldwide. Altogether, these archived collections of web documents form a comprehensive picture of our cultural, commercial, scientific and social history. Web archiving has also an important sociological impact because ordinary citizens are publishing personal information online without preservation concerns. In the future, web archives will probably be the only source of personal memories to many people. We provide some examples of tools that facilitate historical research over web archives highlighting their potential for Humanities.

Keywords: Web Archiving, Digital Preservation, Digital Humanities

1. INTRODUCTION

For centuries, historians have been analysing printed media published in the past, such as books or letters, to research and write history. So, today's ordinary information will be tomorrow's resource for historical research. The web has been replacing printed media and most of the information that characterizes our

International Journal of Humanities and Arts Computing 8.1 (2014): 106–123
DOI: 10.3366/ijhac.2014.0122

www.euppublishing.com/ijhac

current days is being exclusively published online. For instance, web sites are replacing newspapers and books, blogs are replacing diaries, and web photo galleries are replacing photo albums. Thus, the web reflects our current days and it is a crucial resource to support research on Humanities.

The problem is that the information published on the web is extremely ephemeral. Several studies performed and referenced by Daniel Gomes and Mário Silva have shown that one year from now, eighty five per cent of the pages available on the web will have disappeared or been changed.[1] As Tim Berners-Lee, the inventor of the WWW, argued 'There are no reasons at all in theory for people to change URLs (or stop maintaining documents), but millions of reasons in practice'.[2] For instance, sites that are disabled due to lack of funding or pages that are changed to present updated information. The fast and unexpected disappearance of information from the web will prevent future historians from accessing and researching valuable information sources. As it has been performed for printed media for centuries, the information published on the web must be archived and preserved to enable future historical research. UNESCO recognizes digital information as a heritage for future generations and acknowledges that this digital heritage is at risk of being lost.[3] Digital preservation benefits present and future generations, and it is an urgent issue of worldwide concern.

Web archiving aims to acquire, preserve and provide access to historical information published on the web. In April 2013, there were at least sixty four web archiving initiatives worldwide.[4] Web archives also contribute to preserve contents born in non-digital formats that were afterwards digitized and published online. These initiatives hold more than 181 billion web files (6.6 petabytes) gathered since 1996 that provide a comprehensive picture of our cultural, scientific and social recent history.

This article discusses the importance of web archiving for historical research. It discusses how web archives and humanities researchers can collaborate, presents real use cases that illustrate how web-archived information can support future historical research and introduces tools that are already available to facilitate this research.

2. WEB ARCHIVING

Web archiving has spread worldwide and is performed by different types of organizations, such as libraries, universities or companies.[5] The first web archive was founded by Brewster Kahle in 1996 and was named Internet Archive.[6] The non-profit organization who manages it has the stated mission of enabling 'universal access to all knowledge'.

Eric Meyer, et al., discussed actions to be performed in the present days to enable the use of web archives.[7] Tools and methods based on those existent

for the live web are proposed to enable researchers to explore the archived web. The authors identify needs regarding web archives and challenges for individuals, organizations and international bodies to answer them. The findings of a web archive survey of federal depository libraries revealed that libraries prefer to access materials from web archives rather than acquiring them for their collections.[8] However, web archives cannot make an exhaustive preservation of all the published information.[9] Adam Jatowt, et al., presented the results of an online survey conducted with the objective of investigating the users' information needs for temporal support on the web.[10] The results emphasized the users interest in page histories. As web archives become more widely available, studies have been conducted to identify their users' needs, functionalities and collaborations required to support them.[11] Special attention has been paid to enable research over web archives performed by scientists of several areas, from humanities to computer science.[12]

It was not clearly defined which methods and tools should exist to effectively support research over web archives. However, several projects on humanities have already started to use web archives as information sources. The objective of the *Cornell Yesternet* project was to create a research laboratory for social science research based on the Internet Archive's fourty-billion page Web collection. The *Yesternet* project joined social scientists alongside with computer scientists, to study problems like the diffusion of innovation and beliefs or the human behaviour in social networks. They used the Internet Archive collections since 1996 as the main source.[13] The *Virtual Knowledge Studio for the Humanities and Social Sciences* supports researchers in the humanities and social sciences in the creation of new scholarly practices and in their reflection on e-research in relation to their fields.[14] It cooperates with several web archives. *Socio-Sense* is a system for analysing the societal behaviour from long term web archive.[15] It applies structural and temporal analysis methods to historical archived data to obtain insight into the real society. The researchers present excerpts from case studies on consumer behaviour analyses.

The conducted research based on web archived information has already produced interesting results. Masashi Toyoda and Masaru Kitsuregawa extracted the evolution of web communities by comparing four Japanese web archives crawled from 1999 to 2002.[16] Kirsten Foot, et al., examined the linking practices exhibited on archived web sites produced by U. S. Congressional candidates during the 2002 campaign season, focusing on the extent and development of links from candidate web sites to other types of political web sites during the three months prior to the election.[17] Mike Thelwall and Liwen Vaughan examined country balance in the Internet Archive. They concluded there is a bias on the information being preserved and poorer countries are generally under-represented.[18] Although unintentional, researchers using the archive in the future need to be aware of this historical bias. Meghan

Dougherty, et al., discussed ethical issues that arise during a web archiving project, such as selection criteria, privacy boundaries or publication roles.[19] The *K12 Web Archiving Program* introduced web archiving into university and school classrooms.[20] The participants discussed the main principles, concepts and skills required to archive web resources and strategies to incorporate web archiving activities into the classroom. This program involved web archivists, teachers and students on hands-on experience with available web archiving tools.

3. WEB ARCHIVING AND HUMANITIES ARE SYMBIOTIC IN THE DIGITAL ERA

In the twentieth century, the development of telecommunications, such as the phone or TV enabled quick communications but the emission of messages that could be immediately delivered to millions of people across the world was still limited to a small set of people, such as influential politicians. The widespread usage of the web on the twenty-first century changed the world. For the first time in the history of mankind, any person with access to the Internet became able to make a message available worldwide. The number of Internet users grew 566% from 2000 to 2012.[21] In June 2012, thirty four point three per cent of the global population had access to the Internet. Thus, approximately 2,405 million people were able to make a message available worldwide. Nonetheless, the large amount of information that has been produced and made widely available contrasts with the very small amount of it that prevails across time. As information quickly vanishes from the web we may witness a historical gap regarding our current days. As Adam Farquhar from the British Library put it: 'the world has in some ways a better record of the beginning of the twentieth century than of the beginning of the twenty-first'.[22]

Due to the large amount of data involved in the web, web archives must use software components named crawlers.[23] In a nutshell, a crawler starts collecting pages from a set of interesting web addresses to be archived (e.g. home pages of online news). Then, iteratively, it downloads pages from the addresses and follows embedded links to find new content. The crawling ends when all the web addresses are archived or after a pre-determined period of time. However, pages are permanently being updated and, in the same way that it happens for printed media, web archives cannot assure that all the information published on the visited sites was archived. Moreover, there are sites build with technologies that prevent their archive and preservation. Authors that publish their works on the web should follow recommendations to enable their long-term preservation.[24] The crawled information is stored on a repository composed by several computers and it is indexed to create data structures that enable its fast search. The access to the archived information is retained for a period of time to prevent concurrent accesses with the original sites. When this period of

time expires, the information becomes available through web archives access mechanisms.

Web archives enable researchers to search in a few seconds millions of documents written from different perspectives. As web documents can be published by anyone, they provide heterogeneous and original first-person testimonies about historical events. We believe that web archives can provide information that contributes to improve historical research on digital humanities. On their turn, humanities researchers can provide valuable contributions by applying their field knowledge to select and organize web information of manifest interest to be preserved. Researchers can identify sites containing valuable information to be preserved and suggest them for preservation. For instance, the *Web Archive of Catalonia* provides a form to receive suggestions of sites related to this autonomous community of Spain.[25] Moreover, researchers can contribute by generating additional meta-data and organizing information that is already archived. For instance, a researcher can create a web page that documents an historical event and cite web archived documents as information sources. This action would increase the visibility, dissemination and reuse of archived web documents. A researcher could identify a past artistic phenomenon that was documented on the web and create a thematic collection of archived pages about it. This action would enrich and facilitate access to web archived information.

4. USE CASES FOR HISTORICAL RESEARCH

The amount of data of the digitized collections made freely and publicly available on the Internet by the Library of Congress is about seventy four terabytes[26], while the amount of data made available in the same conditions by the Internet Archive is 5,500 terabytes[27]. Therefore, the world's largest web archive, that began archiving information only since 1996, is already seventy four times bigger than the world's largest library. These results suggest that web archives provide a much larger amount of information for research than traditional archives of written media. They hold diverse types of information that support a wide scope of use cases for historical research illustrated and this Section illustrates some of them with real examples.

4.1 International events

Figure 1 presents a thematic collection of 61 million web documents about the historic devastation and massive relief effort due to the Hurricane Katrina and Rita that occurred in 2005. This collection was generated by the Internet Archive along with many individual contributors and it is publicly available. It includes diverse information published on the web by news agencies, governmental agencies, rescue organizations and individuals that survived the catastrophe.

Figure 1. Collection of web-archived documents about the effects of hurricanes Katrina and Rita (2005). Source: Internet Archive, archived between September 4 and November 8, 2005.

Figure 2 presents a collection of archived blogs, Twitter feeds, media coverage and other sites related to the Egyptian 'revolution' that occurred on the January 25th, 2011. This was the first political revolution mainly planned, organized, and executed through the web. It would have been impossible to document exhaustively this historical event without web archiving. Notice that the post-revolution changes are also being published on the web and archived for future historical research.

4.2 Regional events

The web provides historical information with diverse granularities of geographical scopes. Figure 3 presents the official governmental site about the results of Portuguese elections that was archived in April 2003. This site emphasizes the results of the latest results of the 2002 elections. However, by following its links, the web archive users can also access results for different levels of administrative elections since 1997. This kind of content is of extreme historical relevance for future research.

Some web pages are by themselves historical events. Figure 4 presents the first Portuguese web page created in the early 1990's. Although this page is relatively recent it already demonstrates significant historical changes. The map

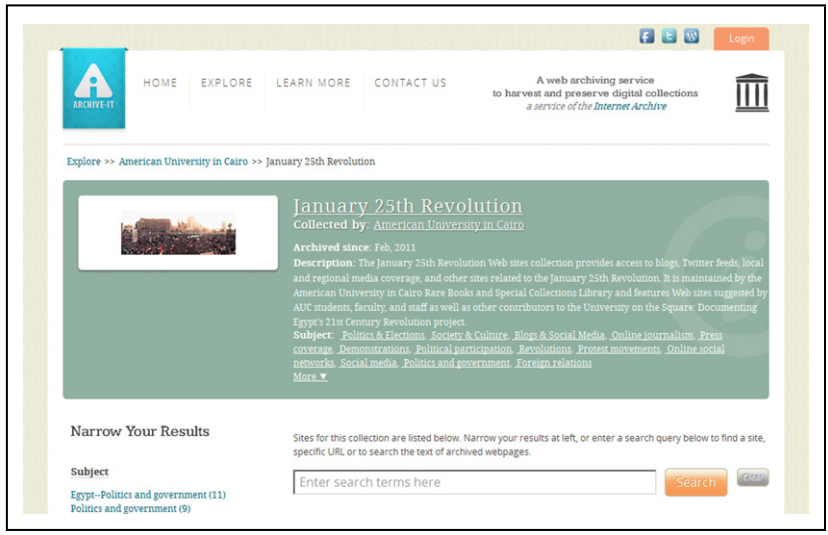

Figure 2. January 25th Egyptian 'revolution' (2011): a web-based revolution, a web-archived revolution. Source: Internet Archive/American University in Cairo, archived since February 2011.

Figure 3. Official results of the 2002 Portuguese elections. Source: Portuguese Web Archive, http://www.eleicoes.mj.pt archived on April 20, 2003.

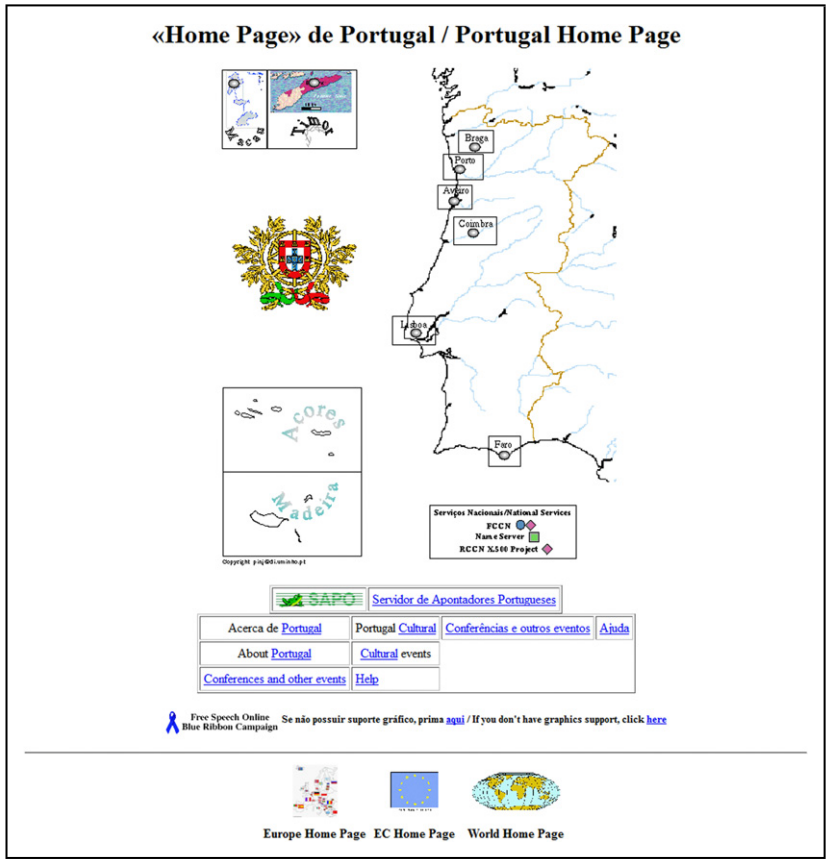

Figure 4. First Portuguese Web page. Source: Portuguese Web Archive, http://s700.uminho.pt/homepage-pt.html archived on October 13, 1996.

of Portugal included the islands of Macau which administration was returned to China in 1999 and East Timor that became independent in 2002. The page footer presents links to the 'Europe Home Page', 'EC Home Page' and 'World Home Page'. The web was so young and experimental that there were single Home Pages for these large organizations.

4.3 Personal events

All information has the potential of becoming historically relevant. It just depends on the context and event that it is being researched. For instance, a free advertisement with a picture selling a kids bike in second-hand may seem

irrelevant for preservation at first sight. However, twenty years from now this could be the first bike of a famous cycling champion or it could be interesting for a researcher studying the design and technological evolution of bicycles across time. Nonetheless, twenty years from now this ordinary advertisement may have a personal and sentimental value to the person that owned it. Traditionally, archivists have to decide which printed documents are worth preserving for later access. Narrow selection criteria had to be applied to suite the available resources. Thus, the chance of ordinary citizens having an important personal event, such as their child births or professional achievements, preserved by cultural heritage organizations for later access has been very low. Web archiving preserves documents with personal relevance granting individuals the possibility of accessing their own History. Notice that the amount of individuals that use the web as primary means of communication has been growing. For instance, people take digital photos and directly share them on the web. However, they are not aware of the importance of preserving their digital data and the most elementary preservation concerns, such as creating backup copies on several disks, are not undertaken. Although everyone became a mass publisher of digital data, fifty years from now most people will not have access to any registries of their memories, such as photos of their loved-ones. Memories play a crucial role in human behaviour and losing access to them may have unexpected impact on modern societies. Web archives will probably be the only source of personal memories to many people.

4.4 Preservation of non-born digital content

In the digital era there is information meant to be printed. However, most of it ends being also published online. Newspapers make their print versions available on PDF format on their sites, individuals or news agencies digitize printed information, such as magazines front pages, publish it online to increase its visibility and obtain revenues. Hence, information initially destined to be printed turns being collected from the web and preserved by web archives. Web archives also contribute to preserve and disseminate contents generated before the digital era that are already valuable for historical research. For instance, the site http://olisipo.no.sapo.pt, that was created by a small group of individuals, presented several old pictures with manifest historical relevance.

Figure 5 shows a military parade that took place in Lisbon on the 10th of June 1966. This date is a national holiday that celebrates the National Day of Portugal and the Portuguese communities. This picture holds high historical symbolism because it was taken on the Day of Portugal during the colonial war on a place that represented the centre of the Portuguese empire. Historical documents published online are collaboratively annotated and enriched.

Figure 5. Military parade during the colonial war on June 10, 1966 (Day of Portugal), Source: Portuguese Web Archive http://olisipo.no.sapo.pt/lisboa/fotos/index20.html archived on June 1, 2010.

Figure 6 presents a blog that publishes digitized historical documents related to the Portuguese first republic with additional annotations generated by the author of the blog, such as transcriptions, citation information and related documents. This example shows a digitized page of the Portuguese republican journal *A Corja*, originally printed on August 16, 1915. This article was enriched on the blog with the information about the journal editorial board, context in which the article was published, highlights on texts related to relevant historical events or entities and related documents. Web archives preserve enriched historical documents because they also include post-publication annotations.

Notice that the presented contributions were voluntarily made by individual citizens. Preserving history is no longer a burden exclusively carried by cultural heritage organizations, but a collaborative endeavour supported through web publication services and web archives. Cultural heritage professionals are specialists on preserving information for later access. However, they cannot master all areas of human knowledge to interpret and enrich the preserved historical artefacts. Personal or domain-specific knowledge of individuals that is voluntarily added to the preserved artefacts provides additional contributions to support research, that otherwise would hardly be obtained. This fact contributes to a more effective preservation of artefacts and to improve the services provided by cultural heritage organizations to support research without requiring additional investments.

Figure 6. Blog with an excerpt of the Portuguese republican journal 'A Corja' originally printed on August 16, 1915. The printed article was digitised and published online by the Library of University of Coimbra. Source: Portuguese Web Archive, http://arepublicano.blogspot.com archived on August 10, 2010.

5. Web Archive Services for Historical Research

Web archive collections are composed by huge volumes of data that make it difficult to interact and take advantage of them. We will precede by presenting and discussing some free tools that can aid researchers which are not web archiving specialists.

5.1 Data mining and search

The automatic extraction of knowledge from large amounts of data is known as data mining in computer science. Data mining functionalities are required

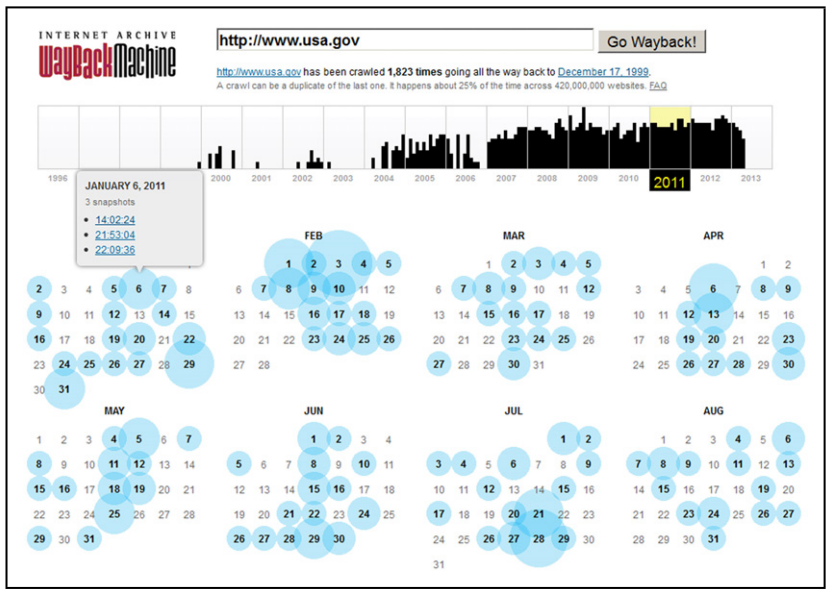

Figure 7. *Internet Archive Wayback Machine*'s view of all archived versions of the U.S. Government's Official Web Portal (http://www.usa.gov/).

for web archives.[28] However, they are not supported by most web archives. The *Portuguese Web Archive* provides software and services to facilitate mining its archived data.[29] Publicly, it provides a service that enables automatic search and processing over one billion files via the *OpenSearch* protocol and tools to convert saved web files to the format used by web archives (ARC format). For research purposes, the *Portuguese Web Archive* provides logs of the crawled web data, a test collection to support research on web archive information retrieval and a computing platform to process its archived information. The *UK Web Archive* provides an N-gram Search service that displays a graph showing how the search phrases occurred in this web archive over time.[30]

Searching gives users the ability to quickly explore through vast amounts of unstructured text, powered by sophisticated ranking tools that order results based on how well they match users' queries. Most web archives provide web address (URL) search as a way for users to explore archived contents. This type of search returns a list of chronologically ordered versions archived from that URL.

Figure 7 presents the result of a search for the U.S. Government's Official Web Portal (http://www.usa.gov/) performed with the Internet Archive URL search service, known as the *Wayback Machine*.[31] Each spotted date means that the address was archived at least once on that day. The size of the spots is proportional to the number of versions collected on that day. For instance, on

January 6, 2011, the URL was archived four times, while on January 8th it was archived just once. These archived versions can be compared with tools such as the *Diff-IE* Add-on for *Internet Explorer* that highlights the differences among them.[32] This kind of features can be used, for instance, to analyse the professional evolution of people through the several versions of their curricula vitae or track the price changes of a product along time.

URL search does not fulfil many of the users' needs because it forces them to know the exact URL where the desired information was published in the past. Hence, some web archives also support catalogue search, where users can filter results by meta-data, such as title or collection. An example of catalogue search is provided by the Library of Congress.[33]

Catalogue search is supported by high quality meta-data generated by archivists or librarians. Thus, it can only be supported over relatively small collections of web-archived documents.

Full-text search has become the dominant form of information access. Most users submit short queries with only one or two terms and expect that the most relevant result will be at the top of the results list. This simplicity of usage makes full-text search the most desired and used functionality for web archives.[34] A scholar can search by the political positions of worldwide leaders before a country invasion or by the feelings of citizens during an economic crisis. Full-text search is supported by web archives, such as the *Australian PANDORA Archive*.[35]

5.2 Browser add-ons for access and self-archiving

Web browsers can be enhanced with add-ons to fulfil the professional needs of specific users. The *Resurrect Pages* add-on can be installed on the *Firefox* browser to facilitate looking for missing information on the web (Figure 8). Users browsing the live web frequently reach pages that are unavailable. This add-on enables them to look for alternative copies or previous versions of the page hosted on web archives, search engines or content delivery networks, with a single click.

The *MementoFox* add-on installs a browser toolbar with a timeline that enables users to visualize previous archived versions of the page they are browsing on the live web. This add-on is a contribution from the *Memento* research project. This project is funded by the Library of Congress and proposes to change the main communication protocol that supports the web (HTTP) by adding a temporal dimension to it.[36]

Web archiving is a complex task that raises significant technological challenges but humanities researchers can manage their own web archives through paid online services, such as the *Archiveit.org* or hire experts to build their web archive of selected sites.[37] However, some information is available on

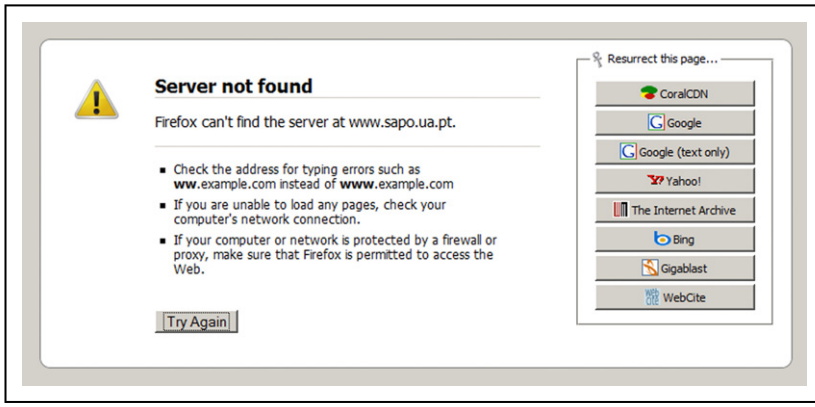

Figure 8. *Resurrect pages* add-on https://addons.mozilla.org/en-US/firefox/addon/resurrect-pages/.

the web for a very short period of time, such as controversial political content, and may need to be immediately archived before it vanishes. *WARCreate* is a free extension for the *Google Chrome* browser that enables users to download a web page and store it in the standard file format used by web archives.[38] This way, any humanities researcher can autonomously select and archive relevant information from the web and then supply it to be integrated in web archives.

Web archives typically gather information that is publicly available on the web and restricted access information such as the one published on social networks is not archived. However, people use social networks to publish and share important information that they may want to archive and preserve for later access. Figure 9 presents the *ArchiveFacebook* add-on for *Firefox* that allows saving content from a *Facebook* account directly to a hard drive, such as photos, messages, activity stream, friends list, notes, events or groups[39]. The *Facebook* wall can be later accessed as it was on the day it was saved.

6. CONCLUSIONS

The web is becoming the main record for our current days. However, its information becomes unavailable after a short period of time, typically less than one year. These two aspects represent challenges for current and future historians. The need for preservation and digital curation is now felt by archivists concerned about the difficulty in finding methods and tools that achieve these objectives, or the uncertainty of the viability and long-term maintenance of these same methods, in a world where technologies and computer languages, likewise tend to become quickly obsolete. But this is also a concern for historians, on

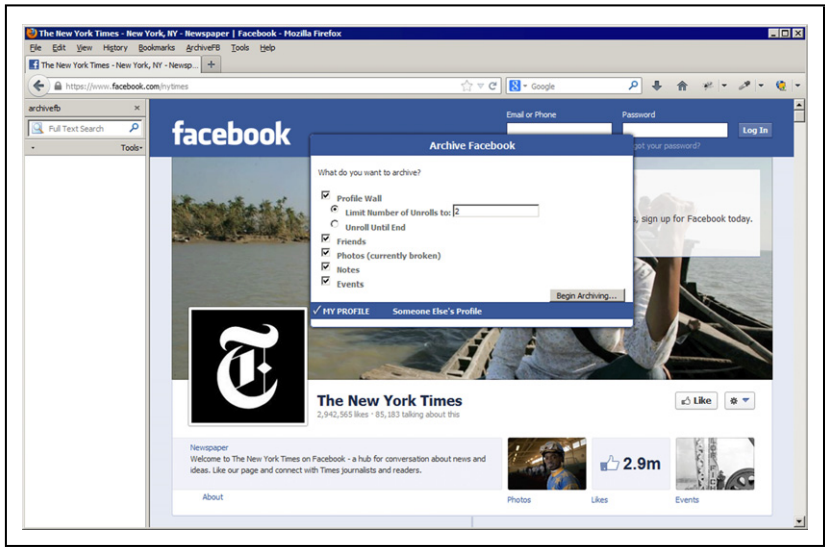

Figure 9. *ArchiveFacebook* add-on for the *Firefox* browser (available at https://addons.mozilla.org/en-us/firefox/addon/archivefacebook/). It enables to archive the information from the owner of a *Facebook* account.

the one hand, because the volume of information produced places them, perhaps for the first time in history, on the edge of data overabundance, and with the difficulty in making choices about what analyze and what to consider relevant to research. On the other hand, the fluidity and volatility of digital may represent a paradox, because the excess of information can be transformed in a huge data paucity, over time, making it very difficult to write the history of the present time which largely is being recorded precisely in digital format[40].

To this extent the digital can make the task of the historians a very complex one, and they must be aware of and involved in the possible solutions to these two hypotheses. Their role should be increasingly active, encouraging measures for digital preservation, as the web archives can clearly be seen, and collaborating in the definition of measures and regulations for setting comprehensive standards of digital curation, in some sense, contributing with their knowledge to a better definition of what to preserve and how to preserve it. But they must also be very aware of these new archives and of the research tools that allow to explore them.

Web archives collect, preserve and provide access to historical web data. They enable numerous use cases to support historical research about global, local or personal events and even contribute to the preservation of non-born digital content. Humanities researchers, and specially historians can benefit from exploring web-archived information but they can also significantly contribute to

web archiving. There are already services and tools that facilitate the exploitation of web archives and even self-archiving of information from the live web. As the web widespread worldwide, web archives will become crucial tools to support historical research.

END NOTES

[1] D. Gomes, J. Miranda and M. Costa, 'A survey on web archiving initiatives', in *International Conference on Theory and Practice of Digital Libraries 2011* (Berlin, 2011).

[2] T. Berners-Lee, 'Cool URIs don't change', W3C, 1998, http://www.w3.org/Provider/Style/URI.html, last accessed 15 March 2013.

[3] UNESCO, 'Charter on the Preservation of Digital Heritage', 2003, http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en. pdf.

[4] Gomes, Miranda and Costa, 'A survey on web archiving initiatives'; Wikipedia, 'List of Web archiving initiatives -Wikipedia', http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives, last accessed 15 April 2013.

[5] M. Ras and S. van Bussel, 'Web Archiving User Survey', Technical report, National Library of the Netherlands (2007); Internet Memory Foundation, *Web Archiving in Europe*, 2010, http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf.

[6] Internet Archive, *About the Internet Archive*, http://archive.org/about/, last accessed 15 March 2013.

[7] E. T. Meyer, A. Thomas and R. Schroeder, *Web Archives: The Future(s)*, 2011, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1830025.

[8] K. R. Murray, *Findings of the Web Archive Survey of Federal Depository Libraries*, 2011, http://research.library.unt.edu/eotcd/w/images/2/29/fdlp_survey_report_krm_14dec2010.pdf.

[9] S. G. Ainsworth, A. AlSum, H. SalahEldeen, M. C. Weigle and M. L. Nelson, 'How much of the Web is Archived', in *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (New York, 2011), 133–136.

[10] A. Jatowt, Y. Kawai, H. Ohshima and K. Tanaka, 'What can history tell us?: towards different models of interaction with document histories', in *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia* (New York, 2008), 5–14.

[11] M. Costa and M. J. Silva, 'Understanding the Information Needs of Web Archive Users', in *Proceedings of the 10th International Web Archiving Workshop* (Vienna, 2010), 9–16; K. R. Murray and I. K. Hsieh, 'Archiving web-published materials: A needs assessment of librarians, researchers, and content providers', *Government Information Quarterly* 25, 1 (2008), 66–89; J. Niu, 'Functionalities of Web Archives', *D-Lib Magazine* 18, 3–4 (2012); Ras and van Bussel, 'Web Archiving User Survey'.

[12] M. Dougherty, E. T. 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 (2010); P. Stirling, P. Chevallier and G. Illien, 'Web archives for researchers: Representations, expectations and potential uses', *D-Lib Magazine* 18, 3–4 (2012).

[13] W. Arms, D. Huttenlocher, J. Kleinberg, M. Macy and D. Strang, 'From Wayback Machine to Yesternet: new opportunities for Social Science', in *Proceedings of the 2nd International Conference on e-Social Science* (Manchester, 2006); W. Arms, S. Aya, P. Dmitriev, B. Kot, R. Mitchell and L. Walle, 'A Research Library Based on the Historical Collections of the Internet Archive', *D-Lib Magazine* 12, 2 (2006).

[14] P. Wouters, 'The Virtual knowledge studio for the humanities and social sciences', in *Proceedings of the 1st International Conference on e-Social Science* (Manchester, 2005).

[15] M. Kitsuregawa, T. Tamura, M. Toyoda and N. Kaji, 'Socio-Sense: A system for analysing the societal behavior from long term Web archive', in *Proceedings. of the 10th Asia-Pacific Web Conference on Progress in WWW Research and Development* (Shenyang, 2008), 1–8.

[16] M. Toyoda and M. Kitsuregawa, 'Extracting evolution of web communities from a series of web archives', in *HYPERTEXT '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia* (Nottingham, 2003), 28–37.

[17] K. Foot, S. M. Schneider, M. Dougherty, M. Xenos and E. Larsen, 'Analyzing linking practices: Candidate sites in the 2002 US electoral Web sphere', *Journal of Computer-Mediated Communication* 8, 4 (2003).

[18] M. Thelwall and L. Vaughan, 'A fair history of the Web? Examining country balance in the Internet Archive', *Library and Information Science Research* 26, 2 (2004), 162–176.

[19] M. Dougherty, K. A. Foot and S. M. Schneider, 'Ethics in/of Web Archiving', in *Computer Supported Cooperative Work Pre-conference on Revisiting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research* (Savannah, 2010).

[20] M. Bragg, E. A. Fox, M. Hedstrom and C. A. Lee, 'Moving Web Archiving into the Classroom', *Proceedings of DigCCurr2009. Digital Curation: Practice, Promise and Prospects* (North Carolina, 2009).

[21] Miniwatts Marketing Group, *World Internet Users Statistics Usage and World Population Stats*, http://www.internetworldstats.com/stats.htm, last accessed 15 March 2013.

[22] The Economist, *Archiving the web: Born digital*, http://www.economist.com/node/17306104, last accessed 15 March 2013.

[23] G. Mohr, M. Kimpton, M. Stack and I. Ranitovic, 'Introduction to Heritrix, an archival quality web crawler', in *Proceedings of the 4th International Web Archiving Workshop* (Bath, 2004).

[24] D. Gomes, J. Miranda and D. Cruz, 'Recommendations for authors to enable web archiving', *FCCN*, 2010, http://arquivo.pt/recommendations, last accessed 15 March 2013.

[25] PADICAT – Patrimoni Digital de Catalunya, *Propose a web site*, http://www.padicat.cat/en/collaborate-and-participate/propose-web-site/send-us-your-proposal, last accessed 15 March 2013.

[26] M. Raymond, *How 'Big' Is the Library of Congress?*, 2009, http://blogs.loc.gov/loc/2009/02/how-big-is-the-library-of-congress/.

[27] Gomes, Miranda and Costa, 'A survey on web archiving initiatives'; Wikipedia, 'List of Web archiving initiatives -Wikipedia'.

[28] Niu, 'Functionalities of Web Archives'.

[29] FCCN – Foundation for National Scientific Computing, *Portuguese Web Archive: Tools and open-source projects*, 2012, http://sobre.arquivo.pt/tools, last accessed 15 March 2013.

[30] British Library, *UK Web Archive N-gram Search*, http://www.webarchive.org.uk/ukwa/ngram/, last accessed 15 March 2013.

[31] Internet Archive, 'About the Internet Archive'.

[32] J. Teevan, S. T. Dumais, D. J. Liebling and R. L. Hughes, 'Changing how people view changes on the web', in *Proceedings of the 22nd annual ACM symposium on User interface software and technology* (Victoria, 2009), 237–246.

[33] Library of Congress, *Search Across Collections (Library of Congress Web Archives)*, http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-search.html, last accessed 15 March 2013.

[34] M. Costa and M. J. Silva, 'Characterizing Search Behavior in Web Archives', in *Proceedings of the 1st International Temporal Web Analytics Workshop* (Hyderabad, 2011); Ras and van Bussel, 'Web Archiving User Survey'.

[35] National Library of Australia, *About | Australia's Web Archives*, http://blogs.nla.gov.au/australias-web-archives/about/, last accessed 15 March 2013.

[36] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth and H. Shankar, 'Memento: Time travel for the web', *arXiv preprint arXiv:0911.1112* (2009).

[37] P. Stirling, P. Chevallier and G. Illien, 'Web archives for researchers: Representations, expectations and potential uses', *D-Lib Magazine* 18, 3–4 (2012).

[38] M. Kelly, C. Northern, H. SalahEldeen, M. Nelson and F. McCown, *WARCreate – Create WARC files from any webpage!*, http://matkelly.com/warcreate/, last accessed 15 March 2013.

[39] M. Kelly, *ArchiveFacebook :: Add-ons for Firefox*, https://addons.mozilla.org/en-us/firefox/addon/archivefacebook/, last accessed 15 March 2013.

[40] B. M. Rogers, 'The Historical Community and the Digital Future', *James A. Rawley Graduate Conference in the Humanities* 26 (2008), http://digitalcommons.unl.edu/historyrawleyconference/26/; L. Roland and D. Bawden, 'The Future of History: Investigating the Preservation of Information in the Digital Age', *Library & Information History* 28, 3 (2012), 220–236.