ARTIGO

RDBCI: Revista Digital Biblioteconomia e Ciência da Informação
RDBCI : Digital Journal of Library and Information Science

DOI 10.20396/rdbci.v16i1.8648747

WEB ARCHIVING: INTERNATIONAL CASE STUDIES AND THE BRAZILIAN CASE

ARQUIVAMENTO DA WEB: ESTUDOS DE CASO INTERNACIONAIS E O CASO BRASILEIRO AUTORES

ARCHIVO DE LA WEB: ESTUDIOS DE CASO INTERNACIONALES Y EL CASO BRASILEÑO

[1]Moisés Rockembach
[1]Universidade Federal do Rio Grande do Sul

Correspondence to Author

[1]Moisés Rockembach
Universidade Federal do Rio Grande do Sul
Porto Alegre, RS - Brazil.
Email: moises.rockembach@ufrgs.br
ORCID: http://orcid.org/0000-0001-9057-0602

Submitted: 01/06/2017
Accepted: 29/08/2017
Published: 12/09/2017

JITA: HC. Archival materials.

RESUMO: O objetivo deste estudo foi delimitar conceitualmente e teoricamente o tema arquivamento da web, além de verificar estudos de caso internacionais e a situação brasileira sobre este tema. Pesquisa de natureza exploratória-descritiva, utiliza abordagem qualitativa, com a aplicação de metodologia de seleção e análise de estudos de caso internacionais para exemplificar o funcionamento do arquivamento da web em vários países, além da análise do contexto brasileiro. São abordados os modelos de processo e ciclo de vida do arquivamento da web, bem como aspectos legais. Dos exemplos encontrados, foram selecionados seis estudos de caso internacionais, contemplando uma variedade de contextos (organizações sem fins lucrativos, arquivos nacionais, bibliotecas regionais, bibliotecas nacionais, universidades, provedores de serviços), não encontrando nenhum sistema específico de arquivamento da web brasileira documentado na literatura, com somente alguns assuntos arquivados de forma esparsa. A investigação diferencia-se por trazer um tópico de pesquisa ainda incipiente no país, mas com muitas possibilidades de estudo. Conclui que é preciso fomentar a discussão no panorama nacional e sugere que iniciativas sejam desenvolvidas para o arquivamento da web brasileira.

PALAVRAS-CHAVE: Arquivamento da web. Tecnologias de arquivamento. Memória digital. Casos internacionais. Caso brasileiro.

ABSTRACT: The objective of this study was to delimit, conceptually and theoretically, the webarchiving, besides, verifying international case studies and the Brazilian situationon this subject. Research with exploratory-descriptive nature, using qualitative approach, with the application of methodology of selection and analysis of international case studies to exemplify the operation of web archiving in several countries, in addition to the analysis of the Brazilian context. It approached the process and life cycle model of web archiving, as well as legal aspects. From the examples found, it has selected six international case studies, covering a variety of contexts (non-profit organizations, national archives, regional libraries, national libraries, universities, service providers), not finding any specific web archiving initiative in Brazil, documented in the literature, with only a few issues archived in a dispersed way. The research differs by bringing a topic of research still incipient in Brazil, but with many possibilities of study. It concludes that it is necessary to foment the discussion in the national panorama and suggests the development of initiatives for the Brazilian web archiving.

KEYWORDS: Web archiving. Archiving technologies. Digital memory. International cases. Brazilian case.

RESUMEN: El objetivo de este estudio fue delimitar, conceptual y teóricamente, el tema archivamiento de la web, además de verificar estudios de caso internacionales y la situación brasilera sobre este tema. Investigación de naturaleza exploratória-descriptiva, utilizando abordaje cualitativa, con la aplicación de metodología de selección y análisis de estudios de caso internacionales para ejemplificar el funcionamiento del archivamiento de la web en varios países, además del análisis del contexto brasilero. Son abordados los modelos de proceso y ciclo de vida del archivamiento de la web, así como aspectos legales. De los ejemplos encontrados, fueron seleccionados seis estudios de caso internacionales, contemplando una variedad de contextos (organizaciones sin fines de lucro, archivos nacionales, bibliotecas regionales, bibliotecas nacionales, universidades, proveedores de servicios), no encontrando cualquier sistema específico de archivamiento de la web brasilera documentado en la literatura, con solamente algunos asuntos archivados de forma dispersa. La investigación se diferencia por traer un tópico de investigación todavía incipiente en el país, pero con muchas posibilidades de estudio. Concluye que es preciso fomentar la discusión en el panorama nacional y sugiere que iniciativas sean desarrolladas para el archivamiento de la web brasilera.

PALAVRAS CLAVE: Archivamiento de la web. Tecnologías de archivamiento. Memoria digital. Casos internacionales. Caso brasileño.

1 INTRODUCTION

In a straight way, web archiving can be defined as a process that comprehends harvesting, to store and provide the retrospective information from the *World Wide Web* for future researchers. This process encompasses worldwide initiatives, some with global approaches and others geographically located, focused in their respective countries. This attribute can be identified by the electronic address domain or through the validation of the information's producer and the context.

However, due to the large production of digital information on the web environment, it has become crucial to analyse how this archiving is being developed in the international scene, as well as identifying studies that show the national situation. Based on the literature and published studies, this investigation converges to two questions: what is web archiving and what are its uses? What researches are being developed concerning the web archiving topic in the international level and which is the Brazilian situation nowadays?

In a first approach about the topic, it was identified the difficulty in finding studies on the Brazilian scenario, which characterizes the originality of the research. The importance and relevance of the subject are justified by being a recent topic of research on the national scope that needs further investigation. Besides, because it concerns the memory of everything that was and it is being produced and diffused in the Brazilian Web and the perspective of future access to those information.

For an adequate framework of the research, it was systematized with the description of the methodological procedures, followed by the presentation and discussion of results, which encompass the results retrieved from scientific journals databases, the fundamental concepts related to the Web archiving and possible uses of the information retrieved by this process, as well as the final considerations about the research.

2 METHODOLOGY

The research was designed in a qualitative, exploratory-descriptive approach, which tried to identify international and national case studies in Web archiving, through the search of scientific literature, papers published in scientific journals and Grey literature, specifically documents, reports and white papers published on the Internet, that establish concepts, theories and politics of Web archiving. Besides, it searched an epistemological framework about the papers found, identifying the concepts and uses assigned to the Web archiving, as well as the selection of international study cases and the Brazilian case.

In the beginning of the research, it was identified the scientific production about Web archiving, because, since it is a new topic on the national context and with an exploratory research approach, it is fundamental a survey of what has been produced more recently on the field, as well as the classic literature about the subject.

The research was conducted at the Scopus database, with the terms "web archiving", limiting the research to the last 15 years (2002-2016).

Since, beyond the exploratory characteristic, the work has also a descriptive approach, the analysis continues with the establishing of concepts and uses of the Web archiving, outlining what it was found on the research at the Scopus database and in the Websites that discuss about Web archiving, for a better theoretic framework.

Still, it was possible to select 6 international case studies about Web archiving, surveying the main characteristics, and the Brazilian context on the subject.

3 WHAT IS WEB ARCHIVING AND HOW DOES IT WORK?

For a better understanding, it was elaborated an answer for this question from the concepts identified on the international literature, as well as the description of its functioning.

Through the research performed at the Scopus database with the delimitations specified in the methodology, 210 results have been recovered. From the three major document types found, most part of it is of conference papers, a total of 114 results, followed by journal papers, 72 results, and books' chapters, 10 results. The reading and analysis of these materials, along with the use of the information available in the websites that execute Web archiving, allowed to comprehend, explain and exemplify this area of study.

It is important to delimit what is going tobe called Web, beyond everything inserted on this sphere, to also limit the object of study on the Web archiving area. It is also necessary to understand what the digital archiving of this material means, how it works, in which phases it is divided and how these information are recovered; to determine the process they pass, from harvest to recovery.

If the question "what is the world wide web, or its abbreviated term Web" is asked, the answer needs to point directly to Tim Bernes-Lee and his concept, which turned out to be the web known nowadays. The world wide web was created in 1989, result of Bernes-Lee proposal to supply the need of scientists to share information, having the European Organization for Nuclear Research - CERN (*ConseilEuropéen pour la RechercheNucléaire*), as initial context of implementation.

The initial proposal of the Web emergence brought, in its general view, the idea of organizing, providing access to information and avoiding to miss important details of the projects developed at CERN, given the high complexity of the process and the several related documents. The *hyperlink* system, term coined by Ted Nelson in the fifties, has the function of linking information, in a different way from the hierarchical structure in trees or with key words (BERNERS-LEE, 1989). The establishing of a manner to identify the objects inside of the Web, through the *Uniform Resource Identifier* (URI), the use of the *Hypertext Markup Language* (HTML), the use of a data transfer protocol, *Hypertext Transfer Protocol* (HTTP) and the use of a browser – at first the Worldwideweb, subsequently renamed as Nexus, and afterwards the Mosaic, as the first web graphic browser – were essential elements for the Web to become the informational and communicational environment that it is known nowadays.

The first web page of the world was not inserted in the network until December 20, 1990 and it can still be accessed online[1]. The preservation of this web address, as well as of the first links (*hiperlinks*) has been done by a project developed by CERN[2]. The World Wide Web Consortium[3] (W3C), founded in 1994, discusses and establishes patterns and guidelines to ensure the web's growth in the long term, based on an open and collaborative web.

One of the biggest worries nowadays is the speed in which the access to information produced and available in the web will be lost. A paper published in 2006, has unveiled that 80% of the Web pages are not available originally after one year, 13% of web references in academic papers disappear after 27 months and 11% of the social media resources, as posted in Twitter, are lost after one year (GOMES, SILVA, 2016).

For that reason, the comprehension of how the Web archiving works becomes an important point to be highlighted. According to Gomes (2010), the three stages of Web archiving encompass gathering information, indexing, and providing services of search and access, so that the first stage subdivides in harvesting the file, store it, extract the addresses to other files from the hyperlink and insert the new addresses detected for the collection. He also highlights the possibility of research by periods or intervals of time in the archived Web and that the need of digital content preservation becomes essential, given the dynamics and inaccessibility of the prior web, even with the few years of informational production.

[1] Available in <http://info.cern.ch/hypertext/WWW/TheProject.html>. retrieved Mar 22. 2017
[2] Restoring the first website - A project to restore info.cern.ch - the world's first website. Available in <http://first-website.web.cern.ch/>. retrieved Mar 30. 2017
[3] Available in <https://www.w3.org/>. retrieved Mar 20. 2017

The image below illustrates the functioning of the Web archiving process:

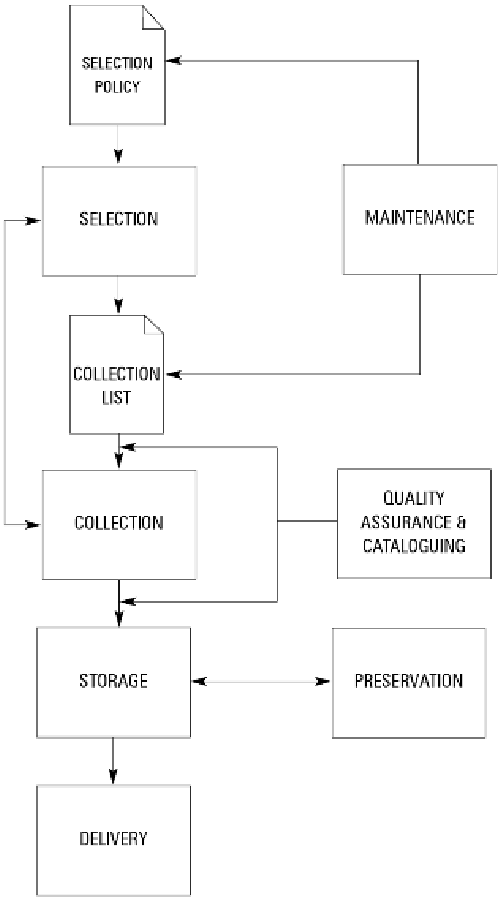

Figure 1. Web archiving process
Source: BROWN, 2006

The process of selection and harvesting is performed continuously and may take in consideration a number of factors, such as content to be harvested, if the external links to the harvested site will be collected as well and to what extent this is done, the equivalent to the harvest extension, or yet its frequency.

Ensuring the harvesting quality is also a fundamental factor for a better archiving and recovery of this information. As put by Hockx-Yu (2012), this would mean trying to store identically to what is seen when accessing a Website directly, however, because of a number of facts, such as dynamic scripts, media streaming, social network structures and database based content, it has become necessary to ensure the quality by four aspects also framed by Hockx-Yu (2012), that emphasize more the content than the graphic visual:

I. If the intended content was fully harvested;

II. If the intellectual content, in opposition to the style and layout, can be reproduced on the access tool;

III. If the harvested copy can be reproduced, including the behavior showed in the live Website, as the capacity of browsing interactively among links;

IV. If the site appearance is maintained.

Still, it has to be stressed that the harvest of web information can be made in two ways, one in which the information's author or owner send it to archiving and other in which the harvest is performed actively by the institution responsible for the archiving. As Websites harvesting approaches, Day (2003) classify it as automatic, selective and by storage harvest, or a combination of these approaches. Some factors cited by Gomes (2010) point to an advantage in the active collection by the archive, such as the harvest automation – which is fundamental given the accelerated growth of published data and that in a short period of time may not be accessible anymore.

Another major point concerns the life cycle of Web archiving. A work group from Archive-It, archiving service connected to the Internet Archive initiative, developed, in 2013, a *White Paper* about the Web archiving life cycle, where the diverse stage and dimension involved on this cycle can be observed.

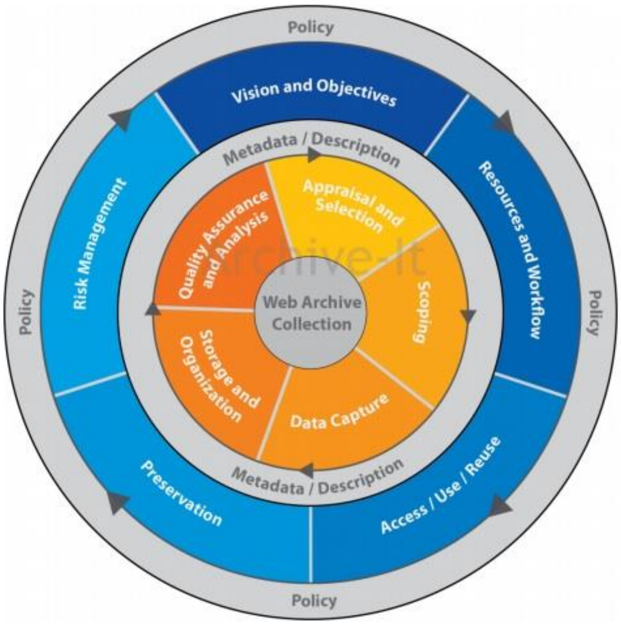

Figure 2. Web archiving life cicle model
Source: DONOVAN, HUKILL, PETERSON, 201

The 'Politics' sphere comprehends the whole Web archiving life cycle. On the blue circle are represented the high level decisions: the definition of the vision and archiving objectives; the available resources and work flow; to monitor how the data access, use and reuse happen; how preservation happens; and how proceeds the risk management, that encompass issues of copyrights, permissions and modalities of access.

The orange sphere, encompassed by definitions of metadata and descriptions, concerns the Web archiving operational tasks, such as evaluating and selecting which Websites will be harvested; the harvest scope; the choice of type and frequency of the data collection; the definition of short or long term storage of the web archives and their right organization; and finally the quality assurance and analysis linked to the objects defined for the web archiving.

Another aspect found that relates to the implementation of web archiving systems is the ISO 28500:2009, that specifies the *Web ARChive* or WARC format as the standard file to be used, replacing the old format ARC (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2009).

About the uses of Web archiving, according to Gomes (2010), some examples that can be listed as users profiles go from a journalist searching for old information, a manager of a website recovering a lost version of the page, an historian studying digital documents, an user searching a broken link on its ‘favorites’, to a jurist obtaining evidence for a case.

Recovering the use of the web retrospectively directed to its scientific side can also become an interesting using approach. The advent of the web goes back to the need of information sharing among scientists. The application in longitudinal studies on the most diverse knowledge areas may lead to breakthroughs, not only in the investigating process and the use of distinguished methodological procedures, but also in the research results.

Beyond a paradigm of custody and preservation (SILVA, et al., 1999) about this retrospective web information, it is important to understand the potential uses of this technology and of the storage content for many purposes, with a perspective about the informational and communicational phenomena involved on the archiving and recovery of personal and organizational memories .

The recovery form of the archived information, the analysis about the amount or type of information harvested, lead to quantitative and qualitative approaches, can be considered relevant fields of study in Information Science and other subjects related to it, such as Archive Science, Librarianship and Museology.

On the legal issues about web archiving, Day (2003), highlights some legal problems, for instance, the author rights and further responsibilities for the available content. The use and reuse of this information is also at stake concerning the permissions, which are not always explicit at the web page, as in the case of the *Creative Commons* application and its concession formats. This implies not only legal problems, but ethical ones, because it involves the use of information under copyrights in international panoramas, under different jurisdictions, and also implies a reflection about informational ethic when using such contents, like the issues of data protection and network privacy.

One of the legal and ethic alternatives for the harvest and use of web archiving, even with no copyright over the websites, is by means of *fair use*, term used on the US legislation and linked to the Common-law tradition applied to the use of content under copyrights in specific situations, such as, pedagogical/educational, news, or yet, research. These issues are specifically discussed by Minow (2003), who defends that the interest in preserving parts of the web content arises due to the importance of it in the production and diffusion of information. And since the majority of websites are under copyrights, there is a dilemma in harvesting these data, which could be solved with the combined use of fair use, of a tool that indicates the non-harvesting intent from the search engines or the control of access permissions – with use of a file robots.txt – and yet, with the express request of taking some captured content.

This legislation and the right of fair use do not apply to every country, therefore, it is necessary to evaluate each case. The Swedish Law, for example, restricts the on line web archiving and the user is allowed to access the content only in the place responsible for the Swedish archiving[4], like the National Library of Sweden and the Kulturarw3 initiative. The same thing happens with the web pages archived by the National Library of France (BnF), which can only be accessed in the reading rooms of the library[5].

4 INTERNATIONAL AND NATIONAL PANORAMA ABOUT WEB ARCHIVING

Through the analysis of the researched literature and the current members of the International Internet Preservation Consortium, six case studies about web archiving and its main characteristics were selected. Also, the Brazilian context was analyzed, from what was found on the bibliographic research.

4.1 International study cases

In terms of web preservation initiatives, some pioneers can be highlighted, such as the Internet Archive Initiative, the Australian National Library project ‘PANDORA’ (*Preserving and Accessing Networked Documentary Resources of Australia*), as well as the Kulturarw3 from Sweden, started in 1996. It is important to emphasize the work of the International Internet Preservation Consortium (IIPC) created in 2003, which dedicates to develop standards and tools that help in the web archiving process.

[4] Restrictions found on the National Library of Sweden web archiving initiative page, “*You can only search the Web Archives in person at the National Library, where there are special computers for the purpose. According to Swedish law, the Web Archives may only be displayed inside the library*”. Available in <http://www.kb.se/english/find/internet/websites/> Retrieved Apr 09. 2017

[5] Restrictions found on the National Library of France, “The Web archives are accessible to authorized users of the BnF, in the reading rooms of the Research Library only. This restriction is the same as that which applies to all legal deposit collections”. Available in <http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html> Retrieved Apr 09. 2017

In 2003, Day published a research about web preservation initiatives, highlighting the relevance of some Organizations, which are listed in descending order of stored information: Internet Archive (based on the United States but with international coverage), Kulturarw3 (Sweden), BibliothèqueNationale de France (France), AOLA (Austria), PANDORA (Australia), Helsinki University Library (Finland), Britain on the Web (UK) e MINERVA (United States).

In research performed by Gomes, Miranda and Costa in 2011, 42 web archiving initiatives around the world were raised and analysed (GOMES, MIRANDA, COSTA, 2011). This study was also the baseline for the information which grounded the production of a Wikipedia[6] page, from where the map of web archiving initiatives around the world is brought.

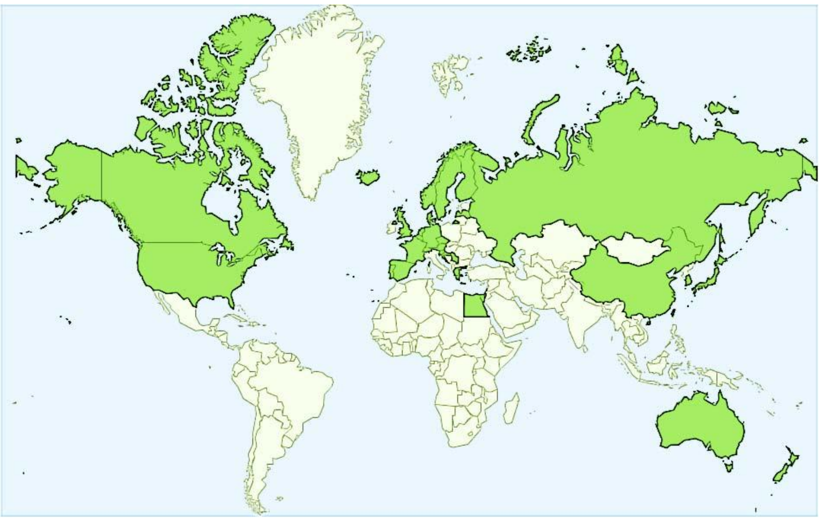

Figure 3. Web archiving initiatives around the world
Source: GOMES, MIRANDA, COSTA (2011) and Wikipedia.org

Six initiatives were selected from those listed at Gomes, Miranda e Costa (2011) work, the selection criteria was based on the analysis of the available literature about these cases in the International Internet Preservation Consortium[7] and in the search for describing a variety of distinct contexts (non profit organizations, national archives, regional libraries, universities, service providers), illustrating with information related to the organization type, the web harvesting forms and in how the access to the stored content is granted.

[6] Wikipedia page "*List of Web archiving initiatives*". Available in <https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives>. Retrieved Feb 10. 2017
[7] IIPC members. Available in < http://netpreserve.org/about-us/members >. Retrieved Apr 2. 2017

DOI 10.20396/rdbci.v16i1.8646067

CHART 1. International Initiatives.

Name/organization type	**Organization type/country**	**Harvest**	**Access**
Internet Archive	Non profit organization - USA	Exhaustive, by domains, selective, by events, thematic, mass harvest, questionnaire, end of life and specifics websites	By the website https://www.archive.org/ and *Wayback Machine* tool
The National Archives	National Archive - United Kingdom	Selective, by event, thematic	By the website http://nationalarchives.gov.uk/webarchive
Library of Catalunya	Regional Library - Spain	'.cat' domain, websites recommended by the public of PADICAT project and institutions that possess a cooperation agreement.	By the websites http://www.bnc.cat/ and http://www.padicat.cat/
Bibliothèque Nationale de France	National Library - France	Significant samples based on two strategies: sample collection in partnership with other institutions and tracking focused in 20.000 selected websites.	Authorized users, only in reading rooms of the Library.
Harvard Library	University - USA	Selective	Web Archive Collection Service (WAX) By the web site http://wax.lib.harvard.edu/collections/home.do
Hanzo Archives Limited	Service provider - United Kingdom	Under demand of companies and organizations, that goes from the corporative website harvest to social network.	Restrict to clients

Source: the author

Beyond the presented data, the bigger initiative to be described is the Internet archive, a non profit foundation, considered one of the oldest web archiving initiatives, dating back to 1996. In these 21 years, the Internet Archive harvested and archived nearly 286 billions pages, from more than 361 millions websites (INTERNET ARCHIVE, 2017). The Internet Archive is also one of the founding members of the International Internet Preservation Consortium.

The Internet Archive initiative has a very popular tool called *wayback machine*[8], which allows to access a website the way it was in a particular time (*snapshot*). For that, it is necessary to insert the domain to be researched and then, with the research results, to select some specific date from the captured data.

Among the harvesting methods used, it can be identified the exhaustive web data harvesting, first level domain harvesting, national and regional, selective, by event, thematic, mass harvesting, survey, sites that will be closed (end of life) and specific sites (IIPC, 2017). The harvesting can be made actively by the user in the Wayback Machine tool page, where it is possible to insert the electronic address to be captured and save the page at the Internet Archive database, therefore enabling the user to quote the reference originally in the future. However, this tool is available only for sites that allow the use of web tracking (*crawlers*) for their indexing, which is the same technology used by search engines.

Beyond websites, the Internet Archive also archives books, texts, movies, software, music and images. Also, are archived collections of many thematics and many of those are displayed in the Internet Archive’s home page, showing the total of items in each collection.

However, in initiatives as the Regional Library of Catalunha and the National Library of France, are archived specifics domains. In the first, the “.cat” and in the second the domains “.fr” and from French territories, as “.re” (Reunião), “.nc” (Nova Caledônia), “.gf” (Guiana Francesa) among others, as well as domains related to France or French culture (.bzh, .alsace, .paris) and other common domains (.com, .org, etc.), created at French territory or with content from France.

The partnerships between institutions are also frequent on the cases studied. The United Kingdom has a Web archiving formed by contributions of several institutions, leaded by the British Library, the *UK Web Archiveou UKWA*[9], which aims to preserve the United Kingdom websites and has the collaboration of the National Library of Walesand the National Library of Scotland. The UKWA already participated in the National Archive of United Kingdom,JISC (Joint Information Systems Committee) and The Wellcome Library.

It is worth noting that in Portugal there is a specific structure concerning the web archiving, the Portuguese Web Archive (Arquivo da Web Portuguesa - AWP)[10]. Some important points raised by the Portuguese researchers are described below, one connected to recommendations on how the sites could be better archived, and another connected to a Web collaborative preservation project.

[8] Wayback Machine. Available in <https://archive.org/web/> Retrieved Apr 05. 2017
[9] UK Web Archive (UKWA). Available in <https://www.webarchive.org.uk/ukwa/> Retrieved 10 Apr. 2017
[10] Portuguese Web Archive – Available in <http://arquivo.pt/> Retrieved Feb 19. 2017

The Portuguese Web Archive (2015) published guidelines so that the web pages can be better archived. These guidelines are classified in 'website organization', 'content of each web page' and 'General', with fundamental and/or advisable guidelines each. In terms of the website organization, the fundamental guidelines are that, links to the address of each content and to the main or home page, should have a friendly form for the web trackers (*crawlers*); and the advisable guidelines are to keep the same address for content over time and the use of the Robots Exclusion Protocol (REP), through the file robots.txt, which will allow the page's author to distinguish what must be collected or not.In terms of the content of each web page, as fundamental guidelines, it is recommended the use of HTML links in the published pages, texts published in text forms and content type (*MIME Type*), and encoding of characters properly identified; while as advisable are, the existence of metadata about the content, the use of file format standards (validated by the W3C), the date of identified publication and the use of appropriate formats for preservation. At last, as advisable guideline in the general scope, the respect for the guidelines of usability and accessibility for people with disability, which are worth to search in the W3C guidelines about accessibility.

From 2007 to 2011 it was developed a project that aimed the Web collaborative preservation, called rARC (archive replicator ARC), where Internet users gave part of the storage space from their hard disks for the archiving, through the installation of a program to perform this archiving automatically. The project had as contribution a Master thesis, titled "Web preservation through replication distributed on a large scale" (NOGUEIRA, 2008). The collaborative preservation project has been discontinued, but being an open source, its source code is available for use in others web archiving initiatives[11].

In United States, in the attempt to integrate the research in several web archives, the project Memento arises, funded by the *National Digital Information Infrastructure and Preservation Program (NDIIPP)* and the *Time Travel tool*, which allows the research in web archives that follow the protocol Memento Time Travel for web, RFC 7089[12].

4.2 The Brazilian case

Some global initiatives, like the Internet Archive, try to harvest and analyze the whole worldwide web. However, as shown at figure 3, there are no web archiving initiatives found in Brazil, and the same happens to the other countries in Latin America. Chile is not at the image yet, but in 2014 it has entered as a member of the International Internet Preservation Consortium, through the National Library of Chile[13].

[11] Collaborative preservation project rARC – Available in <https://code.google.com/archive/p/rarc/>. Retrieved Feb 16. 2017

[12] About the *Time Travel* service. Available in <http://timetravel.mementoweb.org/about/>. Retrieved Apr. 2017

[13] Available in < http://www.netpreserve.org/member-organizations/biblioteca-nacional-de-chile-national-library-chile >

The Permanent Program of Preservation and Access to Digital Archival Documents from the National Archive of Brazil (AN Digital), started in 2010, disclosed in its web page the document Digital Preservation Policy, with a version from 2012 and an updated one from 2016. In both documents, it is highlighted the need to preserve the digital documents, yet, it is exposed that the preservation may be imposed to the documental types "formatted structured text, matrix image, vectorial image, audio, audiovisual, electronic mail message, presentation (slides), spreadsheet, and relational database" (ARQUIVO NACIONAL, 2016, p.11), and that " in the future, more complex types of documents in digital format, like multimedia and web pages, will be contemplated as well" (ARQUIVO NACIONAL, 2016, p.11).

There is an initiative related to the Brazilian web called Latin American Web Archiving Project, hosted in an electronic address from the University of Texas in Austin[14],the focus of this initiative is the governmental and of politic expression documents. The beginning of the harvest was in 2005, but not all the collections are still updated. At the page Latin American Web Archiving Project, 4 collections have been found: Latin American Government Documents Archive (LAGDA)[15]; Mexico 2010[16]; Archive of Venezuelan Political Discourse (ARVEPODIS)[17]; e Archive of Political Parties and Elections in Latin America (APPELA)[18].

Another specific initiative was from the *Content Development Working Group (CDG)*[19] linked to the International Internet Preservation Consortium, which harvested sites, papers, news, blogs and social media (Twitter, Facebook) about the 2016 Rio Olympic Games. The hashtag '#RIO2016WA' on Twitter was also a way to mark and track information about the 2016 Olympics archiving, besides connecting people willing to contribute with the web archiving process of the event.

Dantas' PhD thesis (2015) addressed issues about the web archiving with a theoretic-practical approach, it discussed cultural aspects and about what society considers digital memory as well as technical aspects. Those issues led to the identification of international web archiving initiatives, but also to the finding that there are no web pages collections in Brazilian institutions, and to empirical experiences, of showing how the archiving process is performed and the formation of a collection related to search tools in Brazil.

After the research at the Internet Archive site, it was possible to find some collections related to content produced in Brazil, but very scattered, with no defined plan or established selection and archiving policies. For instance, old magazines collections like

[14] Available in <http://lanic.utexas.edu/project/archives/>. Retrieved Feb 16. 2017
[15] Available in <http://lanic.utexas.edu/project/archives/lagda/>. Retrieved Feb 17. 2017
[16] Available in <http://lanic.utexas.edu/project/archives/mexico2010/>. Retrieved Feb 17. 2017
[17] Available in <http://lanic.utexas.edu/project/archives/arvepodis/>. Retrieved Feb 17. 2017
[18] Available in<http://lanic.utexas.edu/project/archives/appela/>. Retrieved Feb 17. 2017
[19] Available in < http://www.netpreserve.org/working-groups/content-development-working-group>

"GeraçãoPrológica (Brazil)"[20], that contains editions from 1984, 1985 e 1986; or the collection "Brazilian Web Engines (1997-2013)"[21], developed in research project at Federal University of Rio de Janeiro (Unirio), which tried to store *snapshots* of the Brazilian search engines, linked to Dantas research (2015).

As raised by Brayner (2016), countries like Brazil, which have not yet concerned about storing and preserving the web on national level, should develop web archiving policies in order to safeguard their digital patrimony.

5 FINAL CONSIDERATIONS

The research shows that the attribution or initiative of developing the web archiving is not attached only to an specific type of organization. In the examples analyzed, there are institutions public and memorial such as National and Regional Archives and Libraries, as well as private organizations; there are also the non profited ones and some linked to research, as in the universities.

Those organizations use different technologies and establish distinct forms of data harvesting, not only according to the geographic region or electronic domain, but, as verified, also in the selection policies, linked to events, thematics, some extensively, trying to encompass the whole web, and others more focused in certain contexts.

Today the web is configured as an environment containing information over a vast number of knowledge fields and with diverse file formats. The production and use of web content has transformed the way in which people communicate nowadays. Many studies arise from the interactions on the digital environment, and many others studies about the uses of web retrospective information are arising. As far as this archiving technology is applied to more contexts and countries, more researches considering the local characteristics might be started.

If the countries, regions or the governmental and research institutions do not take responsibilities in trying to recover what was produced in this informational environment, much of what is being generated digitally will be lost. Besides, all the possibilities of research over these contents will not be seized by future researchers from the most diverse knowledge fields.

[20] Collection GeraçãoPrológica. Available in <https://archive.org/details/geracoprologica>. Retrieved Apr 02. 2017

[21] Collection*Brazilian Web Engines*. Available in <https://archive.org/details/ArchiveIt-Collection-4266>. Retrieved Apr 02. 2017

REFERENCES

ARQUIVO DA WEB PORTUGUESA. **Recomendações para a criação de conteúdos preserváveis ao longo do tempo.** Disponível em: < http://arquivo-web.fccn.pt/colaboracoes/recomendacoes-para-autores-de-sitios-web >, 2015.

ARQUIVO NACIONAL DO BRASIL. **Política de preservação digital.** Versão 2. Dezembro de 2016. Disponível em < http://www.siga.arquivonacional.gov.br/images/an_digital/and_politica_preservacao_digital_v2.pdf >. Acesso em: 05 abr. 2017.
BERNERS-LEE, Tim. "Information management: a proposal." **Word Journal Of The International Linguistic Association**, 1989.

ALENCAR BRAYNER, Aquiles. Programa de arquivo de páginas web no reino unido: Uma breve história de oportunidades e desafios. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação,** Campinas, SP, v. 14, n. 2, p. 318-333, maio/ago. 2016. ISSN 1678-765X. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8645982>. Acesso em: 10 fev. 2017. doi:http://dx.doi.org/10.20396/rdbci.v14i2.8645982.

BROWN, Adrian. **Archiving Websites**: a practical guide for information management professionals. Facet publishing, London, 2006.

COSTA, Miguel; GOMES, Daniel; SILVA, Mário J. The evolution of web archiving. **International Journal on Digital Libraries**, p. 1-15, 2016.

DANTAS, Camila Guimarães. **Criptografias da memória**: um estudo teórico-prático sobre o arquivamento da web no Brasil. 2015. Tese (Doutorado em Memória Social) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2015.

DAY, Michael. Preserving the fabric of our lives: a survey of web preservation initiatives. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 7.: 2003: Berlin. [**Proceedings of…**]. Berlin: Springer-Verlag, 2003.

DONOVAN, Lori; HUKILL, Graham; PETERSON, Anna. **The web archiving life cycle model**. 2013. Disponível em: < http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf >. Acesso em: 10 fev. 2017.

GOMES, Daniel. Preservar a Web: um desafio ao alcance de todos. In: CONGRESSO NACIONAL DE BIBLIOTECÁRIOS, ARQUIVISTAS E DOCUMENTALISTAS. 2010. **Actas do...** [S.l.] : [s.n.], 2010.

GOMES, Daniel; MIRANDA, João; COSTA, Miguel. A survey on web archiving initiatives. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 15.: 2011: Berlin. [**Proceedings of…**]. Berlin: Springer-Verlag, 2011. p. 408-420.

HOCKX-YU, Helen. **How good is good enough?** – quality assurance of harvested web resources, 2012. Disponível em: <http://blogs.bl.uk/webarchive/2012/10/how-good-is-good-enough-quality-assurance-of-harvested-web-resources.html> Acesso em: 07 abr. 2017

INTERNET ARCHIVE. Disponível em: <https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/>. Acesso em: 16 fev. 2017

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 28500:2009. Information and documentation - WARC file format, 2009.

MINOW, Mary. **Digital preservation and copyright by Peter Hirtle**, 2003. Disponível em: <http://fairuse.stanford.edu/2003/11/10/digital_preservation_and_copyr/>. Acesso em: 06 abr. 2017.

NOGUEIRA, André Ricardo Lopes. **Preservação da web através de replicação distribuída em larga escala**. 2008. Dissertação (Mestrado) **-** Universidade Nova de Lisboa, 2008.

SILVA, Armando Malheiro da et al. **Arquivística**: teoria e prática de uma ciência da informação. Porto: Edições Afrontamento, 1999.

DOI 10.20396/rdbci.v16i1.8646067

ISSN 1678-765X
9 771678 765041 17 >