Moisés Rockembach:

Inequalities in digital memory: ethical and geographical aspects of web archiving

Abstract:

This paper approaches web archiving as preservation of digital memory and as a dynamic informational environment with complex problems of harvest, use, access and preservation. It uses a qualitative and exploratory-descriptive approach, identifying web archiving initiatives and promoting a reflection on the ways of defining web information collection, geographical gaps in web archiving and problems regarding uses and rights of this information. Whereas initiatives such as Internet Archive harvest a lot of information from across the web, an imbalance of digital memory exists where many countries do not possess their own web archiving initiatives, and therefore, coverage of information is unequally produced.

Agenda:

Author:

Prof. Dr. Moisés Rockembach:

- Faculdade de Biblioteconomia e Comunicação (Fabico) - Universidade Federal do Rio Grande do Sul (UFRGS), Rua Ramiro Barcelos, 2705, Porto Alegre, Brazil
- ☎ + 55 - 51 - 33085434 , ✉ moises.rockembach@ufrgs.br, 🖥 www.rockembach.com
- Professor of Information Science and Communication Studies (Master Degree / PhD – PPGCOM / UFRGS) and Professor of Archival Science and Information Science - Department of Information Science (UFRGS)

Introduction

Information as financial capital and as a fundamental element of the functioning of organizations and society as a whole around this value are characteristics grounded by Castells (2000), exemplified in the perspective of information-based production, in contrast with the traditional industry-centered model and the detention of the means of production. The concepts and issues regarding the power of the network, automation, intensive use of technologies, obsolescence related to technological innovation and the global economy, even with local resistance, are factors that lead to technological and ethical reflections.

There is an ongoing shift in the relationship between information and society, due to the billions of information entities being constantly produced and disseminated through the Internet, across platforms, from the web to social networks. To place this phenomenon in context, we look at two concepts from the work of Luciano Floridi (2010). Floridi refers to the Zetabytte Age, a concept of magnitude that defines how much digital information has been produced in recent years. The second is what Floridi calls the infosphere, an environment that includes analog supports, potentialized by digital media.

How does one retrieve, use and preserve retrospective information produced on the Web? Objectively, web archiving is defined as a process that includes collecting, storing and making available retrospective information from the World Wide Web to future researchers. This process involves several initiatives around the world, some of them global-wide approaches and others geographically localized in their respective countries. This attribute is identified by the electronic domain or by identifying the producer of the information and the context in which it is embedded.

The ethical aspects considered are the means used to perform the web content harvest, the geographical gaps in harvest and preservation of information and the problems regarding the use and rights of the harvested and stored information.

Using a qualitative and exploratory-descriptive methodology, we explore research in web archiving, conceptually and theoretically, to verify the initiatives in several countries and to identify ethical questions about the use and preservation of retrospective web information processes.

Web as a contemporary digital environment of information

The world-wide web was born in 1989, a result of Bernes-Lee's proposal to support scientists in sharing information, with CERN (Conseil Européen pour la Recherche Nucléaire) as the initial context of application. The initial proposal of the Web had the aim of organizing, giving access to information and avoiding the loss of important details from the projects developed at CERN (Bernes-Lee, 1989).

The world's first networked web page was published on December 20, 1990 and can still be accessed online. The maintenance of this web address, as well as the first links (hyperlinks) has been preserved by a project developed at CERN. The World Wide Web Consortium (W3C), founded in 1994, discusses and establishes standards and guidelines to ensure long-term web growth based on an open and collaborative web.

By 2014, the number of online networked web pages surpassed the 1 billion mark, stabilizing in 2016 (Internet Live Stats, 2017). As many web pages have become inactive with the current number of inactive webpages sitting at approximately 1.2 billion web pages, though the number is floating. The graph below shows the evolution of web page production.
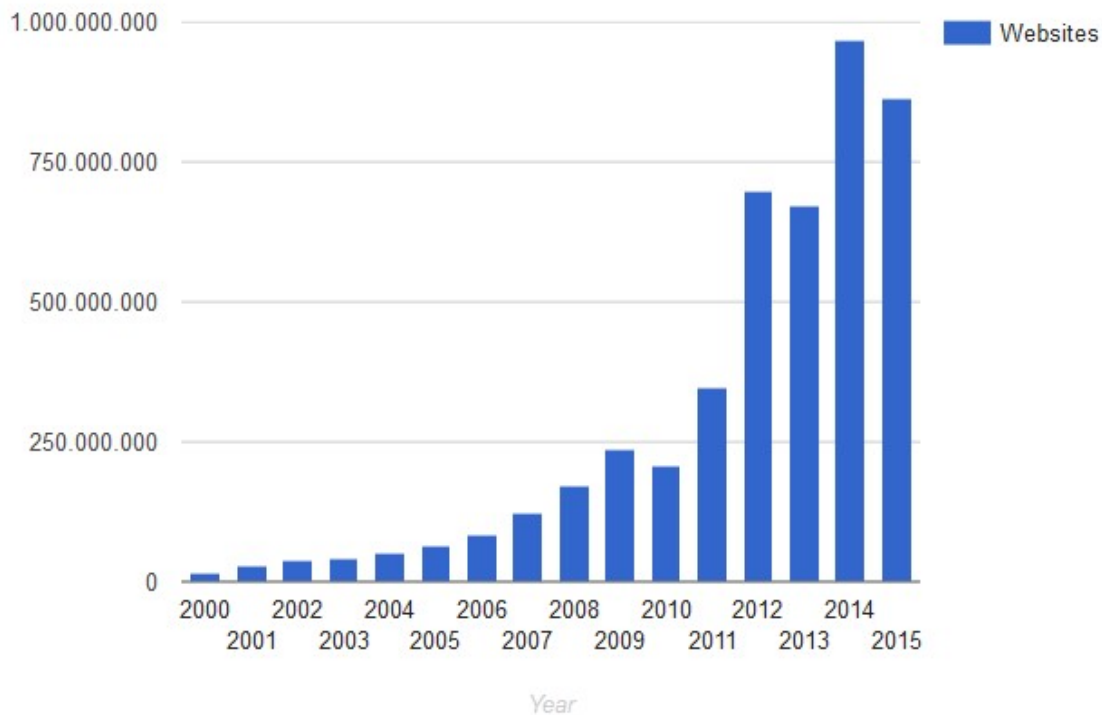
www.i-r-i-e.net   139
    ISSN 1614-1687

Figure 1 - Total number of Websites in the world with unique hostname
Source: Internet Live Stats, 2017

The web environment has become one of the main ways of production and flow of information and communication, generating interactions and new possibilities of articulation among network users. However, several studies describe that "80 % of webpages are not available in their original form after 1 year; 13 % of web references in scholarly articles disappear after 27 months; and 11 % of social media resources, such as Twitter posts, are lost after 1 year" (Costa, et al., 2016). Web Archiving is and will be an important technology of a digital research infrastructure for Science and Society, it can be used to preserve institutional sites, as well as to support special research collections or building new collections for research.

Web archiving initiatives

Masanés (2006) explains the concept of the Web Archive and the background involving technologies and platforms, with contributions from several authors, regarding web archiving approaches. The collecting or 'harvesting' of web content is made in several ways, two of those ways being *wide method*, where everything is collected; and *selective strategies* or events, for example, political and sporting events. There are specific harvesting methods, with themes that interest specific communities or researchers of specific scientific fields. The access to web archive contents, in some countries, can be achieved online; in others, it is necessary to go to the Archive / Library to access it locally.

Certain pioneers of the field stand out as particularly relevant to the history of web archiving, such as the Internet Archive initiative (http://archive.org), based in the United States, Australian National Library 'PANDORA' (Preserving and Accessing Networked Documentary Resources of Australia) project and the Kulturarw3 from Sweden, initiated in 1996. The work of the International Internet Preservation Consortium (IIPC), formed in 2003 and dedicated to the development of standards and tools that assist in the web archiving process, must also be acknowledged.

In research performed by Gomes, Miranda and Costa in 2010, 42 web archiving initiatives around the world were identified in 26 countries (Gomes, Miranda, Costa, 2011). This study was also used for the production of a Wikipedia page, which mapped web archiving initiatives around the world.

Figure 2 - Web archiving initiatives around the world
Source: Gomes, Miranda, Costa, 2011

The IIPC website (netpreserve.org) recently made available a tool on the map mundi model, powered by Google Maps, updated with the 51 members of the consortium, shown in the figure below:

Figure 3 - Members of the International Internet Preservation Consortium (IIPC) in 2017
Source: IIPC (http://netpreserve.org/)

In South America, Chile recently became a member of the IIPC. However, as highlighted in the above map, it is the only member country in Latin America. Likewise, Africa boasts only Egypt as a member. Many countries do not have their own web archiving initiatives, which makes them dependent on what other international institutions preserve.

In Brazil, for example, there are approximately 4 million websites registered under the .br domain, according to the institution NIC.BR, responsible for the activities of domain registration in Brazil (NIC.BR, 2017), and there is not a systemic web archiving instituted in the country. Similarly, other countries do not have a systematic collection of their web pages, a fact that may be detrimental in the formation of their own digital memory.

Ethical aspects regarding Web Archiving

Based on the above analysis, a number of questions should be raised about the uses of technology towards a clarification of the issues, thus encouraging new questions and reflections, all towards the greater analysis. The primary questions concerning web archiving centre around the phases of harvest, access/use, and preservation:

Harvest

  I.   **Regarding harvest:** What parameters are used to evaluate and select the information to be archived digitally? What is the frequency of harvest and the influence of this on digital memory? Does this collection occur equally in all regions? If the relationship is unequal, why is it this way? How can the right to privacy be protected?

Web archiving initiatives have several ways of evaluating and selecting data harvest. The scope is usually defined by the type of organization that does the harvest and its objectives in harvest and preservation. According to Masanès (2005), "the quality of a Web archive can be defined by (a) the completeness of material (linked files) archived within a designated perimeter and (b) being able to render the original form of the site, particularly regarding navigation and interaction with the user".

Internet Archive has the widest and most exhaustive harvest of all initiatives in the world and, as akin to prerogative of the Ancient Library of Alexandria, it intends to keep all human knowledge ever produced (Barabási, 2003) in the form of webpages. From a geographical context, the National Archives and National Libraries are generally the institutions that make the harvest, from the internet domain that corresponds to the country, or a selective method, by events or by topics, an example being the work developed by The National Archives (United Kingdom) or The National Library (France).

Worldwide, the Internet Archive is a model of web archive, the largest initiative to preserve web pages and other file types, like books, movies, software, music and games. The methods of harvest used by the Internet Archive are identified by national domain, regional domain, bulk, selective, event and thematic. For each method, specific parameters are applied, such as the configuration of crawlers per domain, or manually, if the crawler has limitations (Niu, 2012). Unfortunately, this harvest may not occur in a geographically balanced way, as it shows in the following example.

As exemplified by Leetaru (2015) with the harvest of news outlets, a source of information with high rates of updating, and monitored by the GDELT Project (The GDELT Project, 2017), there is a strong centralization of the web archiving harvest resources among a relatively small number of countries, noted in the comparison of the maps below.
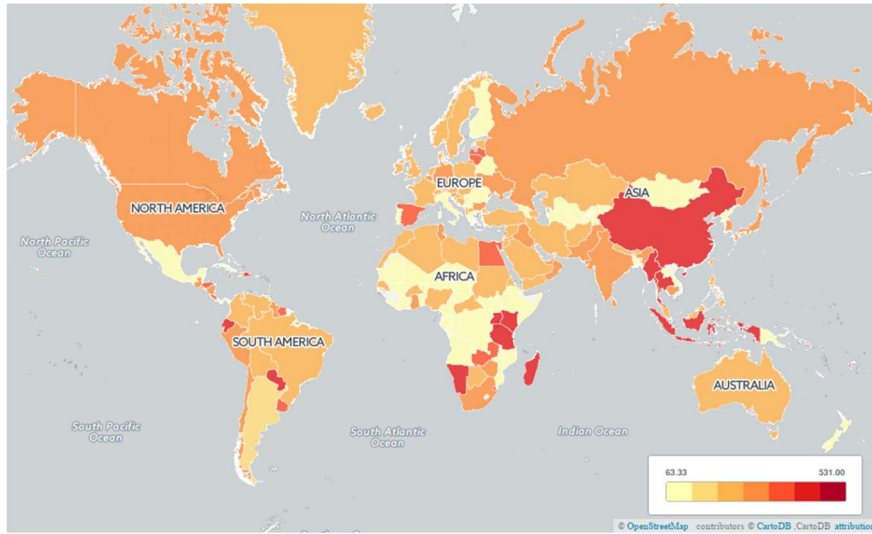
Figure 4 - Average number of hours with at least one snapshot by outlet for online news outlets by country (2013)
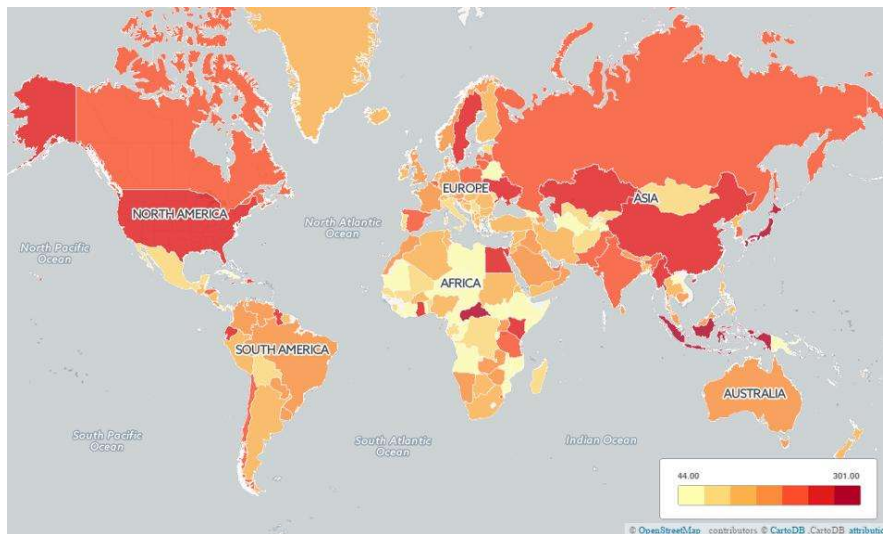**Source: Leetaru, 2015**

Figure 5 - Average number of hours with at least one snapshot by outlet for online news outlets by country (2014)
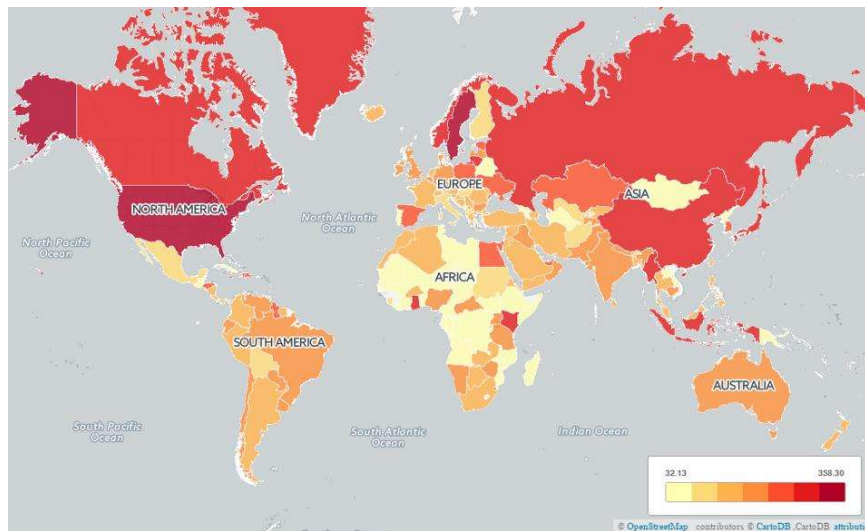**Source: Leetaru, 2015**

Figure 6 - Average number of hours with at least one snapshot by outlet for online news outlets by country (2015)
Source: Leetaru, 2015

Leetaru (2015) wonders how the web pages are archived, how to select what is harvested and how exhaustive it is done, because the platform declares that it has 445 billion pages archived, but still there is not a decent explanation of what is archived, as well as transparent algorithms that determine what was captured, when and how often, in addition to a master inventory of what was captured. According to Leetaru (2015), "The significant geographic change over time adds further evidence that the behaviour of the archive's crawlers is constantly changing in profound and undocumented ways". This emphasizes the need for local initiatives that address the contents of different countries, as a complement, and also that the way web pages are crawled and collected must be transparent and publicly available.

The United Nations (UN) establishes a set of fundamentals called 'Internet Rights and Principles', which states in its principle number 5, regarding privacy and data protection: "Everyone has the right to privacy online. This includes freedom from surveillance, the right to use encryption, and the right to online anonymity. Everyone also has the right to data protection, including control over personal data collection, retention, processing, disposal and disclosure" (Franklin, Bodle, Hawtin, 2015).

As such, everyone has the right to not have their information collected and, if the information is collected without authorization, the right to have the information removed, because the protection concerns the entire information life cycle.

These aspects add technical and social problems concerning the guarantee of data protection and also the right to be forgotten, a new social phenomenon, related to avoid stigmatization by past events that are potentialized by web publishing. Bunn (2015) considers the case 'Google Spain vs. Agencia Española de Protección de Datos and Mario Costeja González' as the founding act of the right to be forgotten. However, even if the information is removed or hidden from the Google index, it is possible to make others copies, without Google control. Web Archiving is one of these copies, therefore archiving also needs to be regulated.

It is necessary to highlight that, even though the harvested sites contain personal information, this information was disclosed publicly by the author or the holder. In this case, the access and use of this archived information becomes authentic, as it would be the access to this information if still available on the web, for research goals. If the information is personal with non- or unauthorized publication, as the use of the Right to be Forgotten, it entails the same criteria, for both the web available in the network and the archived web, removing the information availability from the search engines.

As in the example from Hiiragi, Sakaguchi and Sugimoto (2009) in Institutional Web Archiving System (IWAS), instead of removing all the archived content, it is possible to hide some parts of the archived sites that hold sensitive information.

Access and use

> **II.   Regarding access and use:** What questions must be addressed around how information is used and manipulated? How can copyright be addressed in terms of the use of that information?

How archived websites can be used and manipulated is an issue that can be analyzed from a technological and ethical-legal point of view. As far as access and use are concerned, from a technological perspective, indexing, retrieval and visualization are important factors to be developed. In order to search for content in different ways, it is necessary that the contents are indexed and are retrievable not only by the original electronic address, but also by commercial search engine through the use of keywords.

If the web archive is being used for scientific purposes, it becomes even more critical to understand the harvest criteria that led to the collection of the data. Just as it is necessary to establish clear criteria in the collection of scientific data, whether by questionnaire, interview, observation, among other instruments, care must be taken when using existing data, available online. Towards this end, Jones and Johnson (2006) identify web uses and studies such as Content Analysis, Surveys, Rhetorical Analysis, Discourse Analysis, Visual Analysis, Ethnography and Network Analysis. Highlighting the pertinent ethical considerations, they note that, "some material found on the Web and being archived is confidential, inadvertently made available and then stored by search engines like Google", adding that "it is important to note that the thoroughness with which search engines scour the Web can lead to the archiving of material that users had neither intended to make public nor to have it archived".

Glanville (2010) points to political, legal, and ethical factors in web archiving programs in Australia and The Netherlands where the ethical and legal aspects important to the study are copyright and privacy. Web archiving initiatives are caught in a copyright dilemma between a public responsibility to collect, providing access to the archived web, and concerns for the privacy, use and access to this data.

On ethical-legal issues involving archiving the web, Day (2003) notes problems around copyright and other responsibilities for the content made available. The use and reuse of information is also at stake with regard to permissions, which are not always explicit in the web page, as is the case of the application of Creative Commons and its formats of concessions. This implies, not only legal, but also ethical problems, because it involves the use of information under copyright, applied to international contexts, under different jurisdictions, and it also implies reflection on informational ethics when using such content, such as data protection and privacy issues on the network.

One of the legal and ethical alternatives to this collection and use of web archiving, even though it does not have copyrights on the sites, is fair use, a term used in the American legislation which is related to the tradition of Common-law, applied to the use of content under copyright, in certain situations, such as pedagogical / educational use, as news, or as scientific research. These issues are specifically discussed by Minow (2003), who argues that because the web is becoming an increasingly important resource for production and dissemination of information, there is an interest in preserving parts of its content.

Understanding that most sites use copyright poses a dilemma to the harvest and use of their data. However, solutions such as fair use, where laws and jurisprudence allow open access to information using tools that flag abuse through monitoring the control of access permissions, are a start. Using a robots.txt file – or better yet, sending a consent form to the information owner with an express request to remove some captured content offers a feasible common ground.

This legislation and the right to fair use does not apply to all countries, and it must be assessed on a case-by-case basis. Swedish law, for example, restricts online access to web archiving, since it only allows the user to

access the site in the place responsible for archiving the Swedish web, such as the National Library of Sweden and the Kulturarw3 initiative. The same goes for accessing the web pages archived by the National Library of France (BnF) which are only available to access in the library reading rooms.

The imbalanced representation of the harvest is also reflected in access and use, where unfortunately some countries or geographic communities may not have archived their memory. If it supposes that Historiography will depend on current records and that much information is produced only on the web, users will be limited by the harvest made. According to Brügger (2011), "archiving of the web enables us to write web history, which is a necessary condition for the understanding of the Internet of the present as well as of new, emerging Internet forms". Milligan (2016) also defends the contemporary use of web archives in History, with ethical attention in the analysis of personal information and data protection, using as example a study of the Geocities platform. Despite the difficulties in preserving the web, "a nation's web domain can be considered a relevant object of study for historiography as well as a genuine and valuable historical source" (Brügger, 2017). In addition, research from the web historiography approach depends on the resource or source available, with some challenges when using web archives, namely related to content preserved, which will determine the choices and bias of information sources (Brügger, 2012).

Preservation

> **III.   Regarding preservation:** Who is responsible for preserving the web? How long will information be preserved? What determines such parameters?

In digital preservation, recurring themes are related to the technological problems regarding obsolescence, data migration and in the case of Web Archiving, the standardized file format (ISO 28500:2009). Furthermore, one of the most important issues in the preservation aspect is the responsibility of the preserver. As mentioned, copyright is a key piece in the study of web archiving. In order for preservation to avoid legal risks, one possibility is for the organization that preserves to obtain the right to store, as determined by legal deposit. According to Glanville (2010), "legal deposit exists to preserve the nation's cultural heritage, a function that it is not fulfilling in its current form". This means that organizations, whether National Archives or National Libraries, are unique organizations, their prerogative being the objective of preserving memory, nationally or locally. Thus, there is a legal requirement for the expansion of their scope, which includes the digital realm as an information environment to be preserved. However, as exposed by Glanville, such advances still need to be regulated in some countries.

The IIPC (2017) has a list of members who support their country's copyright deposit legislation: Austria, Canada, Croatia, Denmark, Estonia, Finland, France, Germany, Iceland, Israel, Japan, New Zealand, Norway, Slovenia, Spain, Sweden and United Kingdom. The countries awaiting legal deposit legislation or that use permission, fair use or opt-out based systems to collect web archives, include Australia, Canada (Quebec), Spain (Catalonia), China, Czech Republic, Egypt, Netherlands, Poland, Singapore, South Korea, Switzerland and the United States.

Glanville (2010), quoting Murray and Hsieh (2008) says that "while ideally the onus should be on the content's creator to preserve their own material, an unawareness of the fragility of digital materials and unwillingness to take responsibility for the preservation of these materials means that libraries feel a responsibility to preserve material of value".

What determines how long the records are kept is traditionally connected to the uses and users of the information in question. Despite the digital environment allowance to keep a lot of data, a data curation, or specifically a web curation, can be an important preservation strategy, in order to consider ways of covering web contents more consistently.

However, problems in web curation can arise from several fronts and web preservation also implies a process of *selection* of content that will inevitably involve issues of bias. This becomes relevant since web curation becomes an integral part of the digital memory of culture and society. In Singapore, for example, some bloggers

have been intimidated and monitored due to their web archiving activities (Cadavid, Basha, Kaleeswaran, 2014). Hence, web preservation initiatives can be undermined for any number of reasons. When people feel under surveillance, or also censored in the production of web contents, the monitored, controlled and/or biased results will remain for posterity.

As a relevant guide, the UNESCO (United Nations Educational, Scientific and Cultural Organization), published in 2003 the Charter on the Preservation of Digital Heritage, where in its article 1, they state that "the digital heritage consists of unique resources of human knowledge and expression. It embraces cultural, educational, scientific and administrative resources, as well as technical, legal, medical and other kinds of information created digitally, or converted into digital form from existing analogue resources. Where resources are "born digital", there is no other format but the digital object" (UNESCO, 2003).

The important aspect to emphasize is that, among the digital objects to be preserved, including images, software and digital audio, are also included web pages, that, "have lasting value and significance, and therefore constitute a heritage that should be protected and preserved for current and future generations. This ever-growing heritage may exist in any language, in any part of the world, and in any area of human knowledge or expression" (UNESCO, 2003).

As the Internet Archive and other platforms do not have a legal responsibility to preserve all websites of countries that do not have their own web archiving initiatives, even if they collect a vast collection of information, there is no guarantee of what information, or how long this information will be preserved. As such, it is important that there exist local web archiving initiatives, focused on local events and contents.

Concluding remarks

Content that is preserved is also content that can be retrieved. What has not been systematically collected can only be found if a platform, such as Internet Archive, has collected it by individual demands, from users who have actively collected the information, or collaterally, in an unsystematic way. These three aspects, harvest, access/use and preservation, will be the basis for establishing History and for the rescue of current events. It is important to emphasize that the web is an environment marked by information dynamism and a lack of control over its flows. What is available and accessible online today, may not be available tomorrow.

Finally, one must ask what the limits of web archiving consist of. The intent of archiving is to preserve all the knowledge of the world, globally or locally, but this can only be done in a jointly concerted effort, but must also remain decentralized, as per the default nature of the internet. These multiple initiatives converge towards the same objective: impacting society's digital cultural heritage.

In terms of personal information contained in web pages, harvested and archived, it is necessary to evaluate on a case by case basis. Unfortunately, not every country has a personal data protection act or a regulation regimentation established to defend its citizens. Even so, respect for Internet Rights and Principles of the United Nations (2015) must be followed so that there can be standard digital ethical principles to meet expectations of archive and access.

At the same time, even though many countries do not have their own web archiving initiatives, it will be possible to take advantage of the documented experiences of the last 20 years, considering that the first web archiving initiatives began in 1996, and implement the most appropriate technological requirements and models. It must be emphasized that there are continents and large geographic extensions with very few web archiving initiatives, which implies loss of significant digital memory for people or nations. Web archiving is a pertinent solution, but the ethical problems listed need to be broadly discussed.

References

Barabási, A. L. (2003). Linked: The new science of networks. Basic Books.

Berners-Lee, Tim (1989). "Information Management: A Proposal." Word Journal Of The International Linguistic Association

Bunn, Anna (2015). The curious case of the right to be forgotten. Computer Law & Security Review, 31(3), 336-350.

Brügger, Niels (2011). Web archiving—Between past, present, and future. In: The handbook of Internet studies, 24-42.

Brügger, Niels (2013). Web historiography and Internet Studies: Challenges and perspectives. New media & society, 15(5), 752-764.

Brügger, Niels (2017). Probing a nation's web domain: A new approach to web history and a new kind of historical source. In: Routledge Companion To Global Internet Histories (pp. 61-73). Routledge.

Cadavid, J. A. P., Basha, J. S., & Kaleeswaran, G. (2014). Legal and technical difficulties of web archival in Singapore. Rev. Prop. Inmaterial, 18, 35.

Castells, Manuel (2000). The rise of the network society. Blackwell Publishers

Costa, Miguel; Gomes, Daniel; Silva, Mário J.(2016) The evolution of web archiving. In: International Journal on Digital Libraries, p. 1-15.

Day, Michael. (2003) Preserving the fabric of our lives: A survey of web preservation initiatives . In: International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg

Floridi, Luciano.(2010) Information: A very short introduction. Oxford University Press. Oxford.

Franklin, Marianne; Bodle, Robert; Hawtin, Dixie (Ed.) (2015) The Charter of Human Rights and Principles for the Internet. Retrieved Jun 20, 2017 from http://internetrightsandprinciples.org/site/wp-content/uploads/2017/03/IRPC_english_4thedition-1.pdf

Glanville, Lachlan (2010) Web archiving: ethical and legal issues affecting programmes in Australia and the Netherlands, The Australian Library Journal, 59:3, 128-134

Gomes, Daniel; Miranda, João; Costa, Miguel.(2011). A survey on web archiving initiatives. In: International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 2011. p. 408-420.

Hiiragi, W., Sakaguchi, T., & Sugimoto, S. (2009). A policy-based institutional Web archiving system with adjustable exposure of archived resources. In The 9th International Web Archiving Workshop (IWAW 2009) Retrieved Jun 22, 2017, from http://lekythos.library.ucy.ac.cy/handle/10797/14074

International Internet Preservation Consortium - IIPC (2017) Legal Deposit. Retrieved Jun 10, 2017, from http://netpreserve.org/web-archiving/legal-deposit/

Internet Live Stats. (2017) Total Number of Websites. Retrieved Jun 20, 2017, from http://www.internetlivestats.com/total-number-of-websites/.

International Organization for Standardization (2009). ISO 28500:2009 - Information and documentation -- WARC file format

Jones, Steve; Johnson, Camille. Web Use and Web Studies. In: Masanès, Julien (Ed.) (2006). Web Archiving. Springer Science & Business Media

Leetaru, Kalev.(2015) How Much Of The Internet Does The Wayback Machine Really Archive? Forbes. Retrieved Jun 10, 2017, from https://www.forbes.com/sites/kalevleetaru/2015/11/16/how-much-of-the-internet-does-the-wayback-machine-really-archive/#38f51b369446

Masanès, Julien (2005). Web archiving methods and approaches: A comparative study.Library Trends

Masanès, Julien (Ed.) (2006). Web Archiving. Springer Science & Business Media.

Milligan, I. (2016). Lost in the infinite archive: the promise and pitfalls of web archives. International Journal of Humanities and Arts Computing, 10(1), 78-94.

Minow, Mary.(2003) Digital Preservation and Copyright by Peter Hirtle. Retrieved Jun 06, 2017, from http://fairuse.stanford.edu/2003/11/10/digital_preservation_and_copyr/

Murray, Kathleen R.; Hsieh, Inga K.(2008).Archiving Web published materials: A needs assessment of librarians, researchers, and content providers. Government Information Quarterly 25, no.1

NIC.BR - Núcleo de Informação e Coordenação do Ponto BR (2017). Retrieved Jun 18, 2017, from https://registro.br/estatisticas.html

Niu, Jinfang (2012). Functionalities of web archives. D-Lib Magazine, 18(3/4).

Pabón Cadavid, J. A., Sathik Basha, J., & Kaleeswaran, G. (2013). Legal and technical difficulties of web archival in Singapore. IFLA World Library and Information Congress 2013, Singapore. Retrieved Jun 08, 2017, from http://library.ifla.org/217/1/198-cadavid-en.pdf

The GDELT Project (2017). Retrieved Jun 21, 2017, from http://gdeltproject.org/

UNESCO (2003). Charter on the Preservation of Digital Heritage. Retrieved Jun 25, 2017, from http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html