Revista Digital de Biblioteconomia e Ciência da Informação DOI: http://dx.doi.org/10.20396/rdbci.v14i2.8645982 **ARTIGO**

UK WEB ARCHIVE PROGRAMME
A BRIEF HISTORY OF OPPORTUNITIES AND CHALLENGES

PROGRAMA DE ARQUIVO DE PÁGINAS WEB NO REINO UNIDO
UMA BREVE HISTÓRIA DE OPORTUNIDADES E DESAFIOS

PROGRAMA DE ARCHIVO DE PÁGINAS WEB EN REINO UNIDO
UNA BREVE HISTÓRIA DE OPORTUNIDADES Y DESAFÍOS

Aquiles Alencar-Brayner[1]

ABSTRACT
Webpages have been playing a key role in the creation and dissemination of information in recent decades. However, given their ephemeral nature, many Web pages published on the World Wide Web have had their content changed or have been permanently deleted without leaving any trace of their existence. In order to avoid the loss of this important material that represents our contemporary cultural heritage, various institutions have launched programmes to harvest and archive Web pages registered in specific national domains . Based on the example of development of the Web archive program in the UK, this article raises some key questions in relation to the technological obstacles and curatorial models adopted for the preservation and access to the content published on the Web.
KEYWORDS: Web page archive. Digital preservation. Curadoria digital.

RESUMO
Páginas Web vêm desempenhando um papel fundamental na criação e difusão de informação nas últimas décadas. No entanto, dado sua natureza efêmera, muitas das páginas Web publicadas na Word Wide Web têm seu conteúdo modificado ou são permanentemente deletadas da rede sem deixar rastros de sua existência. No intuito de evitar a perda deste importante material representativo do nosso acervo cultural atual, várias instituições vêm lançando programas de colheita e arquivamento de páginas Web registradas em seus domínios nacionais. Através do exemplo de desenvolvimento do programa de arquivos de páginas Web no Reino Unido, *UK Web Archive*, o artigo levanta algumas questões centrais em relação aos obstáculos tecnológicos e modelos curatoriais adotados para a preservação e acesso ao conteúdo publicado na Web.
PALAVRAS-CHAVE: Arquivo de páginas Web. Preservação digital. Digital curation.

RESUMEN
Páginas Web han tenido un papel clave en la creación y difusión de información en las últimas décadas . Sin embargo, dado su carácter efímero, muchas de las páginas web publicadas en la World Wide Web han cambiado su contenido o han sido eliminadas de forma permanente de la red sin dejar rastro de su existencia . Con el objetivo de evitar la pérdida de este importante material representativo de nuestro patrimonio cultural, diversas instituciones han puesto en marcha programas de cosecha y archivo de páginas Web registradas en sus dominios nacionales. A través del ejemplo de desarrollo del programa de archivos de páginas Web en Reino Unido, *UK Web Archive*, el artículo plantea algunas cuestiones fundamentales en relación a los obstáculos tecnológicos y modelos curatoriales adoptados para la preservación y el acceso a los contenidos publicados en la Web.
PALABRAS CLAVE: Archivo de páginas Web. Preservación digital. Curadoría digital.

[1] Curador digital de coleções Latino-americanas da British library; Mestre em História da Arte e Estudos Latinoamericanos pela Rijks Universiteit e em Ciência da Informação pela City University de Londres; Doutorado pelo King's College de Londres. Londres, Reino Unido. http://orcid.org/0000-0002-4405-8648. E-mail: aquiles.alencarbrayner@bl.uk
Filed in: 04/04/2016 - **Accepted in**: 24/05/2016

DOI: http://dx.doi.org/10.20396/rdbci.v14i2.8645982

INTRODUCTION

Much of what had been published in the early World Wide Web – presumably most of it – has been lost irretrievably. Since there is no general agreement from institutions and users on the value of the Web and of its contents, views seem to differ on whether attempts should be made to save some or all of Web page contents for the future and how much effort this warrants. Similar situations have arisen in the past for other media formats and documents that are now understood to be of considerable cultural value has been lost. The early films produced by the motion picture industry offer some significant examples on how content of importance for the world's cultural heritage can be permanently deleted. In its early days, motion pictures were considered ephemeral and/or irrelevant, and most were lost, often because film collections were simply recycled to retrieve their valuable silver content. As Peter Kobel explains, "[f]or decades the film industry saw its productions as having limited value: after their initial release, they were soon forgotten, or even destroyed for the few cents' worth of silver in the filmstrips' emulsion… It took a long time for people to realize the importance of preserving 'old' films" (2007, p. 275-6). In a report commissioned by the US congress, the Film Preservation Board came to the alarming evaluation that "fewer than 20% of the features of the 1920s survive in complete form; for features of the 1910s, the survival rate falls to slightly above 10%" (1993, n.p.). Today these few early silent films that were preserved for future generations are deemed to be invaluable cultural artifacts.

This is one of the many cases to illustrate what happens when a new technology or media channel appears for popular use. In general, the contents of these technological innovations are initially approached as ephemeral to become later appreciated as documents of high cultural significance. We are currently witnessing a similar stage in the history of Web pages. Although we all recognise the importance of Web pages in creating and disseminating information in our era, there is still a lack of awareness from the general public about the relevance of archiving this material and even research communities remain skeptical about the importance of preserving Web content. Cultural institutions, however, are become increasingly aware of the importance of Web archiving as an essential activity for preserving contemporary history and culture: "[w]hile many debates about the potential uses of web archives still remain at both a theoretical and practical level, web archiving is increasingly accepted by most cultural heritage institutions as an important complement to more traditional forms of collection development" (DOUGHERTY *et al.*, 2010, p. 9). The commercial sector has also started to devote attention to Web archival as IT companies like Webternity (http://Webternity.eu) and Mirror Web (https://www.mirror-web.com/) are offering Web and social media archiving services as a way to support business in preserving and managing their history of Web publications and interactions with users. The same trend can be found in the public sector where Archive-It (https://archive-it.org/), a Web archiving subscription service provided by the Internet Archive, offers support for academic institutions, research centers, NGO's, museums and libraries.

Valuable content is added to Websites not only by traditional publishers, but also increasingly by end users; and a vast proportion of information that appears on Web pages is not published in any other format. According to The National Archives in the UK, the majority of current government records are produced only in electronic format and the lack of a strategy for archival and preservation of this content will inevitably lead to the disappearance of important information for the future:

> Most government records are now created electronically as a result of the widespread introduction of electronic records management systems. Previous legislation meant that the bulk of records were not transferred until they were 30 years old. However, with the introduction of the Freedom of Information Act (FOI), 'closed until 30' disappeared in January 2005. We now needed to make arrangements to select and preserve such records as soon as possible after their creation since, unlike paper, they are highly vulnerable to corruption and loss (THE NATIONAL ARCHIVES, n.d.).

National libraries and archives recognise the value of capturing and preserving electronic information on the Web and in recent years a number of institutions have started harvesting selected Websites published in their specific countries according to registration in particular national domains. In 2003 six British institutions came together (The British Library, the National Archives, the National Library of Wales, the National Library of Scotland, the Joint Information Systems Committee [JISC] and the Wellcome Library) to form the UK Web Archiving Consortium, UKWAC. The Web archiving landscape has changed considerably since UKWAC's formation, notably resulting in the creation of a number of important collaborative projects and support for the development of Web archiving programmes in the UK. They have made considerable progress in harvesting and archiving Web pages, but the scale and effectiveness of their efforts is still limited by the continuing evolution of Web technologies.

1 DEVELOPMENTS IN WEB ARCHIVING

The archiving of Websites can be traced back to 1996 with the non-profit Internet Archive project in the US and the Preserving and Accessing Networked Documentary Resources of Australia (PANDORA) the Web archiving program launched by the Australian National Library. The Internet Archive started its activities (which have included the archiving of some UK websites), aimed at carrying out captures or 'snapshots' of the world Web with regular intervals, and providing free access to a great number of Web resources archived since 1996. This is the largest depository for archived Web pages: its collection, according to information provided on the Internet Archive Website, currently stands at about 484 billion web pages occupying something around 15 Petabytes (PB), or 15,000 Terabytes (TB) storage space, with an estimated growth of 20TB per month. It operates according to a variety of harvesting models: whole domain, thematic, and deposit. The Internet Archive has been able to build its large collection because, unlike UK institutions until 2013, it has not sought permission from website publishers before harvesting copies. It has harvested without attention to rights issues, operating instead a 'takedown' policy allowing Website owners to

request removal of a site from the archive. The National Library of Australia started harvesting Web pages also in 1996, developing some pioneering theoretical work in Web archiving in support of its PANDORA initiative. PANDORA has been harvesting Websites for around 20 years. Today, it is archiving at the rate of about 448 titles, or 1,493 instances, per month and has accumulated about 44,747 titles over 44,299 instances since the beginning of the project.[2]

The first Web archiving initiative in the UK was the UK Central Government Web Archive launched by The National Archives in 2003. The aim of the project was to harvest and archive government sites of interest to the British public, working in partnership with other Web archiving institutions such as the US Internet Archive and the European Archive programme. At the end of that same year, the UK Web Archiving Consortium – UKWAC was formed, establishing a shared platform for selecting, harvesting and granting public access to archived UK Web pages. The pilot project that run during the first two years of the consortium set up an integrated policy having in mind the different collection scope of each consortium member, identifying common interests and specific institutional strengths for the preservation of Web content.

In 2004 all UKWAC partners started to use PANDAS Web harvesting software, developed by the National Library of Australia, on a shared infrastructure that allowed them to store their collections in a single repository. Content harvested by UKWAC became publicly available in 2005 and in 2009 the consortium changed its name to UK Web archive after two of its founding members, The National Archives and the National Library of Scotland, decided to develop their own individual Web archiving policy according to their evolving needs, withdrawing from the consortium. The British Library offered to take on the service. It now hosts and provides the Web Curator Tool (WCT) harvesting service, and is responsible for the UK Web Archive repository infrastructure.

Currently, the UK Web Archive repository holds all instances previously harvested by UKWAC partners. The British Library, JISC, the National Library of Wales and the Wellcome library now use the WCT service developed by BL, and store their collections together in the UK Web Archive repository, which is managed under contract by the University of London Computer Centre (ULCC). The National Library of Scotland currently uses the Netarchive harvesting software developed by The Royal Library of Denmark, having its own repository. The National Archives now uses the services of the European Archive and stores its contents in the European Archive repository, with access provided through The National Archives' Website.

[2] An instance is a copy of a title harvested on one date. Copies of one single Web page title are added to the archive on different times in order to capture changes of content when the Web page has been updated. Data consulted on 15 March 2016 at http://pandora.nla.gov.au/statistics.html

DOI: http://dx.doi.org/10.20396/rdbci.v14i2.8645982

Between 2003 and 2013 the UK Web Archiving operated on a selective basis for the development of its Special Collection, archiving UK Web pages according to pre-defined groups referring to particular themes or subject (e.g. *Latin American Communities in the UK*) and relevant British events (e.g. The Olympic & Paralympic Games 2012). Each group contains between fifty and four hundred archived Websites. In this process UK Web Archiving partners had to identify relevant Web pages for archiving, contact Web page owners for granting archive permission and make the content openly available via the UK Web Archive portal. In April 2013, after the approval of the UK non-print legal deposit regulation, UK Web Archive became legally entitled to archive all Web pages published in the UK domain without seeking prior consent to the page owner. Contrary to the access policy for its special collection, UK Web archive is not allowed to make the content archived under non-print legal deposit openly available due to some other legal restrictions such as data protection act. In order to access the full content of UK Web Archive, users need to visit the British Library reading rooms where machines are connected to the electronic deposit holding the entire collection. The current operational and collection development programmes for the archive is set up as to capture Web content following three different harvesting approaches:

- A 'snapshot' of every website within scope, currently estimated at circa 4.8 million active sites, will be archived at least once a year
- Some 200 to 500 websites within scope will be archived on a more frequent basis such as quarterly, monthly, weekly or even daily, in order to ensure that rapidly changing or updated content is archived adequately. Such websites will be selected by the legal deposit libraries for their importance and research value, with the crawl frequency being adapted to the circumstances and nature of the content
- In addition, the legal deposit libraries envisage crawling other selected websites in order to develop 'special collections'. Perhaps four or five new collections will be developed each year for important events (which may involve crawling specific websites relatively frequently for a limited period) or important themes (which may involve crawling selected websites regularly over a longer period). (AGENCY FOR THE LEGAL DEPOSIT LIBRARIES, n.p.)

In a more global perspective, the International Internet Preservation Consortium (IIPC) works together with 37 member institutions across the world that are harvesting and archiving Web pages in large scale projects with the aim to "to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations" (NETPRESERVE, n.d.). Since its foundation in 2003, IIPC has been developing tools and carrying on research on Web archiving, collecting evidence on best practices and recommending policies for institutions interested in harvesting and preserving Web pages. It also supports study groups and discussion forums on specific areas of Web archiving such as content management, collection assessment, crawling software performance, digital preservation and public access to archived material. Despite the progress made in recent years on Web archive activities, global consent on Web archiving standards and approaches have not been reached. It was only in July 2009 that an important step was taken in the development of general standards for Web archiving, the WARC file format. Developed by the International Organization for

Standardization (ISO), WARC provides universal support for the harvesting, access and exchange needs of archiving organizations, and sharing secondary content such as metadata.

2 WEB ARCHIVING: APPROACHES AND POLICIES

The archiving of Websites is a complex task that involves harvesting, curating, storing, preserving and managing access to copies of Websites together with their associated digital objects. Web archiving extends to the information contained by sites, the appearance of the pages, separate information objects (text documents, video or audio files) referenced or rendered by the pages, and the behaviour of the sites in response to user interaction – all to the extent possible with archiving software. It follows that Web archiving is not solely about archiving electronic publications, such as reports or pamphlets published in PDF files that happen to be disseminated through the medium of the Web, but seeks more accurately to capture Web users' entire experience, so that this experience can be reproduced for future generations.

The Preserving Access to Digital Information (PADI) Website maintained by the National Library of Australia identifies four distinct approaches for Web page archiving, namely:

1. *Whole domain*: archiving of Web pages related to a specific national Web space. The national Web space is normally indicated by the Top Level Domain (TLD) of a Web address designated by the two final letters of its Universal Resource Identifier (URI), such as .uk; .fr; etc. which indicate the country in which a specific Web page is published. National libraries and archives usually adopt this approach for archiving Web pages.
2. *Selective*: archiving of pre-defined Websites, as chosen by curators using stipulated criteria such as collection scope or institutional services.
3. *Thematic*: a form of selective archiving, where the selection criteria relate to a theme or event.
4. *Deposit*: archiving of Websites deposited explicitly by their publishers and authors.

These different models of Web archiving are not mutually exclusive. In fact, as in the case of the UK Web Archive, many institutions operate on a combined approach policy, using multiple models to build up their Web collections. An essential step to be taken by an institution before setting up a policy for Web page archiving is the delimitation of a collection scope. This scope would differ in accordance to the nature of the archiving institution: a specific governmental organization, for instance, might decide to adopt a selective approach and archive Web pages that deal directly with the services provided by the institution. National libraries usually have a broader scope for Web archiving: their intent in most cases is to archive all Web pages produced by their constituent countries which are considered to be of research importance without restrictions on language, areas of

information or target audience, opting, in this way, for a whole domain approach. Some countries work on a combined approach to Web archiving, as it is the case for example in Australia, aiming to archive all Websites published in their national Web domain as well as including in their collections other Web pages related to the country's national interest despite belonging to other TLDs.

There are, however, two basic criteria that are normally taken into account by national institutions before setting up their selection policy for the archiving of Web pages: the size (micro and/ or macro archiving) and the maintenance of the collection. Web content can be archived with restriction to quantity (a limited number of Web pages), space (maximum storage capacity for each Web page archived), period (length of time to be considered for archival), subject areas and selection of media formats to be stored (inclusion of pages with audio and/or video contents). Institutions such as National Archives and Libraries usually carry out macro archiving project, establishing no restriction in terms of size, period, and subject or file content for the archiving of Web pages, sometimes also including personal blogs, videos and podcasts considered to be relevant for the national collection.

The complexity and costs of a Web archiving programme is reflected in the storage capacity and different media formats an institution aims to preserve. In face of the growing nature and changeability of Websites, the identification, selection and harvesting of Web pages produced in a country can be an expensive and time consuming activity not always producing satisfactory results in the archival of Web pages. PANDORA for example, includes files in different formats such as audio contents, streaming videos and PDF in its archiving selection. According to Crook (2009), one of the biggest challenges for librarians and archivists working on Web archiving is to develop strategic ways in which to assess the importance and/or quality of the Web pages that are being archived. Due to the high number of Web pages archived everyday by the harvesting software, it is impossible at the moment for professionals to be monitoring each individual title that is selected for the Web page repository. It is important to stress that whatever the selection policy for Web archiving might be, it must be accompanied by strategic planning to ensure continuity and consistency in the selection and maintenance of the pages archived.

Due to its limited scope, micro Web archiving is dependent on the decisions made by the archiving software. Web pages that have reached their storage limit, for example, might not have their contents updated in future changes. This problem is avoided in macro archiving which, without specifying the size of Web pages for archival, includes in its harvesting process generic Website domains (e.g. .com), national TLDs (e.g. .uk) and physical location (IP address) of Web servers, achieving, therefore, a more comprehensive yet still selective Web archiving activity. In the case of the UK, the country's national Web programme only archives Web pages within the .uk domain, limiting its collection scope to national sites. Consequently, this approach leaves aside many potential Web pages of particular interest to Britain which are published under other TLDs. Web pages such as those produced by the British settlers in Argentina and Uruguay (www.argbrit.arg), registered on

the ".ar" domain, are out-of-scope for the UK Web Archive consortium rendering, therefore, significant gaps in the archiving UK collection.

Despite working in partnership, each UK Web archiving institution follows its own insitutional collections policy when harvesting Wenb sited to be included in the programme's Special Collections. The National Library of Wales collect sites of interest to its respective region; JISC collects sites of projects funded by the institution; the Wellcome Library collects sites containing information about the history of medicine and the British Library collects sites selectively from the UK Web space, prioritising sites of research value and sites that are representative of British social history and cultural heritage. It also archives a small number of sites that demonstrate Web innovation (HOCKX-YU, 2008).

According to the initial archiving programme proposed by the UKWAC and carried on by the UK Web Archive partners, member institutions were requested to collect sites on matters of particular interest on a thematic basis such as swine flu, the London Olympic Games, and the European Parliament elections. Some of the collecting initiatives adopted by the UKWAC requested the collaboration of other non-member institutions as in the case for the archiving of Websites on the European Parliament elections, which involved the collaborative work of the British Library and seven other national libraries in continental Europe. Although working on collaborative basis, there have been numerous overlaps and potential duplication on archiving efforts, involving two or sometimes more UK Web archive institutions, though it is thought that this affects a small proportion of sites collected. A few examples are:

a. the British Library and The National Archives both have legal remits to collect central government information duplicating, in some cases, the archiving of official Web pages.

b. the British Library collection scope overlaps with that of other partner institutions such as the Welsh national library;

c. some sites that touch on medical research issues may be of interest to both JISC and the Wellcome Library.

These are some examples on how collecting policies for UK Web archive institutions are overlapping in scope. It is true that for libraries and archives collections a certain amount of overlap is acceptable, and it might even be beneficial for different institutions to hold copies of the same material in case of loss or deterioration of an item held in a particular collection. However, when dealing with Web archiving, overlap can cause user confusion because of discrepancies between the collected instances in the various repositories. However, from the perspective of end users, archiving institutions need to bear in mind that duplicated collection of Websites can also lead to difficulties in understanding and interpreting their content. As Hallgrimsson suggests: "[d]uplicate versions of the same document are a challenge because it can be very tedious and confusing for a user if he is presented with many identical documents during access" (2006, p. 139). By adopting different frequency in the harvesting of Web pages, some institutions can present conflicting

results in establishing when the content of a specific site has changed. They might also present inconsistencies in ascribing separate metadata for the pages harvested making it difficult for users to retrieve the material archived due to a lack of uniformity in the description of the Web page's content.

3 HARVESTING AND PRESERVATION OF WEB CONTENT

The fast development in Web technologies creates an urgency for Web pages to be archived. At every moment electronic content is being changed, deleted or becomes simply lost when Websites are redesigned. Websites disappear as their owners or Web servers go out of business or their pages might be removed by third party requests (legal suit, etc.). In the majority of cases, Web content becomes inaccessible as technology changes. The Internet has been characterized by an increasing number of users, rapid technological progress and faster growth of its volume: 40% of the world population has access to the internet which means more than 3 billion people are accessing and generating content on the Web.[3] As May 2016 there are an estimated 1 billion live Websites in the world,[4] yet it is only 20 years ago that the number of Web pages reached one thousand according to statistics provided by the Massachusetts Institute of Technology. This exponential growth is associated with even more rapid change, as new technologies and standards for the Web evolve, are accepted, to become later on, discharged. As Terry Kunny remarks in the very early years of the World Wide Web, "[i]nformation technologies are essentially obsolete every 18 months. This dynamic creates an unstable and unpredictable environment for the continuance of hardware and software over a long period of time and represents a greater challenge than the deterioration of the physical medium" (1997, p. 2). In fact, the underlying Internet technologies and standards are continuously changing, with Web designers constantly using state of the art features. These factors present a significant challenge to institutions charged with capturing the content of the Web. In practice, the capabilities of Web archiving will always lag behind the development of Websites, in exactly the same way as the capabilities of anti-virus software inevitably lags behind the development of new viruses.

Another problem related to Website preservation comes from the fact that digital information is a dynamic object or process which can be altered at any stage in its existence. Differently from printed sources, such as books for example, which are not subject to change of their content once they have been printed (different editions or print runs of the same title are indicated in the bibliographic records of books), Web pages are subject to constant changes during their lifespan without a clear indication when these changes have occured.

According to Brügger (2005), 80% of active Web pages are modified each year. This high rate of change - ranging from update of content to page restructuring to moving of provider to the complete deletion of a page from the internet - creates an awkward challenge

[3] Figure provided by *Internet Live Stats* (http://www.internetlivestats.com/internet-users/) accessed on 16 May 2016.
[4] *Ibid.*, (http://www.internetlivestats.com/total-number-of-websites/).

for preservation which needs to be addressed in the dynamic context of how electronic documents should be archived. To respond to this situation, procedures for Web archiving programmes are being set up so as to record the alterations to Web pages on an ongoing and consistent basis. The UK Web Archive stipulates a harvesting period of twice a year for most of the titles archived. This archiving period, however, is not applicable to Web pages that have their content changed on a more frequent basis such as Websites for news agencies and governmental information. In the majority of cases archiving frequency is decided by curators depending from specific cases. Some Web archiving bodies propose that the contents of a living Website must be archived at least every four months in order to efficiently capture the possible changes in the Web page. This quarterly archiving policy is adopted, among others, by the National Library of Australia, which sets up the exact dates when archived Web pages need to be recaptured for a consistent record of their possible alterations.[5]

The treatment of defunct Websites is another factor to be taken into consideration by Web archiving institutions. An early study carried out by the Digital Curation Centre (DCC) has reported that "the average lifespan of a Web page in 2003 was deemed to be 100 days and it is not unreasonable to suggest that it is even shorter today" (KELLY & PENNOCK 2006, p. 3). This ephemeral nature of Web pages also has implications for the way in which a Web archiving programme is set up. Once a Web page is reported dead, the archiving institution needs to re-access the page within a period of 4 to 8 weeks after the notification of closure of the Website to guarantee that its contents can be considered 'static' which means that no alterations have been made in the page since it has been reported inactive. Once a Web page is considered 'static' by automatic decision, it no longer will be archived by the harvesting software. Web curators have to dedicate a fair amount of time identifying Web pages which become active after a long period of inactivity. As Crook reports: "to successfully create collections takes far more curatorial time than was initially envisioned. Selection of which web sites to crawl is an often misunderstood activity and can take up surprisingly large amounts of time" (2009, p. 833).

It is important to highlight here that even today not all the content of a Web page can be harvested for archival. Web archiving technology still faces limitations which prevent it from operating as a complete and optimal archiving procedure. The list below refers to the most common shortcomings of Web archiving software:

a. Web page content that requires a log-in process is not captured by Web archiving software even when passwords and usernames are provided to access the stored data.

b. Contents of a Web page which use an absolute path or are stored in a different root URL (as is the case with Web pages that store their images on Flickr) are in most of cases not retrieved by Web archiving programmes due to the software's inability to relate the content of a specific Web page to other third party Websites.

c. Extension languages like JavaScript are not possible to be accessed by Web archiving software thereby restricting the harvesting of Web page contents that use such

[5] NLA set dates for harvesting instances of Web pages are: 1 January, 1 April, 1 July and 1 September.

scripts. The same rule applies to any other form of interactive parts in Websites based on exchange of information between client and server.

In order to deal with the shortcomings of Web archiving software in harvesting external content embedded into a specific Web page, archiving institutions are initiating dialogue with content-sharing Websites such as Flickr, MySpace and YouTube seeking permission to collect and archive their material. This would enable institutions to offer a more comprehensive collection of the various elements that constitute Websites granting public access to this material in the future. Due to the limitations of existing software's capability to archive files encoded in scripting languages (JavaScript, graphic user interface, etc) as well as multimedia extension files such as ShockWave and Flash, Web archiving institutions are converting these different languages and extension files into simpler formats, such as Jpeg or Mpeg, thereby making files available for archiving. Although this offers a quick solution for Web archiving, file conversion has proved to be a labour intensive process that requires "a fair amount of technical skills to recode the archived pages with the changes we had to make" (CROOK, 2009, p. 834). While the number of Web sites that use extension files is increasing rapidly, only a small fraction of harvested pages have had their extension files converted into archiving formats. Consequently many archived Web pages that use extension files are still missing important parts of their content.

Most of the times, when a Web page contains a link to an external Website (i.e. one that is not explicitly being harvested) the software captures the link and also a snapshot of part of the external Website, so as to preserve the user experience for someone browsing the archived site. In practice, the software often captures only the home page of the external site. In some circumstances this results in the solitary home page being indexed and listed as if it were an instance in the index for the external site, along with the full instances. These are referred to as 'artifacts' of the software. This happens frequently for some sites. So some index entries represent complete instances, but most represent only an artifact, or home page; and users have no way to determine which is which.

4 THE EVOLUTION OF WEB 2.0 APPLICATIONS: NEW CHALLENGES FOR WEB ARCHIVE

In the early days of the Web, each Website's was mostly ascribed to a single owner, and most of its content was static. The roles of 'publisher' and 'reader,' and the position of Web pages as 'publications' were fairly clear, similar to those in the print world. This soon changed, through a variety of mechanisms, and the print paradigm no longer applies for Web pages. A key reason for this is the trend towards interactive, or 'Web 2.0' applications and user-created content (UCC). Static, traditionally published content still exists on the Web, but Web 2.0 and UCC blur or make impossible to identify the distinction between publisher and reader.

Some of the early Web content is relatively easy to preserve. For example, static documents published on Websites can be preserved as PDF files regardless of the existence or non-existence of Web archiving. In practice, this is somewhat moot – experience of Web archiving has already found several instances in which such documents are not adequately maintained or preserved elsewhere. However, the same does not apply for most UCC and Web 2.0 interactive content: the majority of this kind of information exists only on Web pages and if the content on live Web is not preserved or secure it will be lost forever.

Unlike static documents, Web 2.0 content is not, and in many cases could not be, preserved in any way other than by the capture of Websites. There is no firm deadline that dictates a need for preservation actions. However, several factors combine to make the need for preservation increasingly important. In most of the cases, the continuing evolution of Web browsers that allows Web 2.0 content to be accessed make old Website technology obsolete. In order to make their Websites compatible with new browsers, authors need to migrate the content of their Web pages to new formats, loosing the touch-and-feel of their original pages when these were created.

It is difficult to define Web 2.0 applications precisely. Web 2.0 is generally understood to refer to Web technologies that allow interaction of some sort, bringing constant modifications to Website content by the site's users. According to a report commissioned by JISC almost ten years ago but still very relevant when describing the current state of Website platforms (Anderson, 2007), Web 2.0 technologies can be divided into the following categories:

a. blogs;
b. wikis;
c. tagging and social bookmarking;
d. multimedia sharing;
e. audio blogging and podcasting;
f. RSS and syndication;
g. newer Web 2.0 services and applications.

The inclusion of a category for "newer Web 2.0 services and applications" is indicative of the speed of change. As the report suggests: "[i]n recent months, however, there has been an explosion of new ideas, applications and start-up companies working on ways to extend existing services. Some of these are likely to become more important than others, and some are certainly more likely to be more relevant to education than others" (2007, p. 12).

Web 2.0 technologies continue to be adopted in all sectors – syndication feeds on government run sites, wikis in the non-profit sector, UCC sites for consumers, marketing blogs in the commercial sector – applications are uncountable. Significantly, take-up of Web 2.0 continues. Blogs are one of the most pervasive Web 2.0 applications and their use in Web page contents is rapidly increasing and their widespread use illustrates the scale and nature of the challenge presented to Web archivists. Figures on blog creation and interaction prove how

much this resource is significant for today's society. Reports provided by Statista (2011), a major statistics portal on digital markets, estimate a exponential growth in the number of blogs created in a timespan of 5 years: from 35.8 million in 2006 to 173 million in 2011. Susan Gunelius (2014) from ACI, a scholarly blog index provider, points out to the fact that WordPress, one of the most popular blogging platforms, registered 14 million new blogs in 2013. For libraries and archives, the archiving of blogs is an important issue that needs to be addressed in terms of cultural heritage preservation.

Turning to a different Web 2.0 technology, the growing popularity of online videos shared in Web platform is an important factor to take into consideration when dealing with external content being embedded or broadcast via life stream in Web pages. It is estimated that users upload 500 hours of videos every minute on YouTube channel (RELSEO, n.d.) and that the majority of the videos accessed on the Web are from UCC sites (YouTube, Megavideo.com, etc) rather than from broadcasters like BBC, ITV and others in the UK, as this means that the content available through UCC sites is unlikely to be preserved in any way other than by the preservation of the Websites in which they are embedded.

CONCLUSION

Web pages are today an essential medium for publication, management and dissemination of information and their importance continues to increase at a fast rate. Many cultural institutions are developing programmes for preserving relevant Websites published in their countries, addressing effectively the vast challenge that Web platforms development impose on to the task mainly by the rapid change in technology and interactive nature of today's Web sphere as described in this article. At this stage and based in the examples here discussed of programmes supported by national governments to secure efficient actions in Web page archive, it is imperative for countries like Brazil that haven't yet addressed the importance of archiving and preserving their rich digital patrimony at a national level, to implement policies for Web archival. Data provided by Registro.br, the regulation agency that provides Web addresses in the .br domain, shows that there are to date 3.839.319 active Web addresses in the country, representing a staggering annual growth of Blogs and Web pages since 1996. Without A Web archiving initiatives, most of this content will be permanently lost within a few years. A delay in moving forwards with a comprehensive national Web archiving initiative in Brazil will lead to the long term and irretrievable loss of valuable cultural content.

REFERENCES

AGENCY FOR THE LEGAL DEPOSIT LIBRARIES (ALDL). Electronic legal deposit: legal deposit libraries' collecting plans for 2013-2014. Available at <http://www.legaldeposit.org.uk/electronic/2013-2014-collecting-plans.html>. Accessed on 15 March 2016.

ANDERSON, Paul. **What is Web 2.0? Ideas,Technologies and Implications for Education**. London: JISC, 2007. Available at: <http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanniing/hs0701.aspx. Accessed on 9 April 2016.

BRUGGER, Niels. **Archiving Websites: General Considerations and Strategies**. Aarhus: The Centre for the Internet Research, 2005.

CROOK, Edgar. Web Archiving in a Web 2.0 World. **The Electronic Library**: **the international journal for the application of technology in information environments**, Bingley, v. 27, n. 5, p.831-836, 2009.

DEPARTMENT FOR CULTURE, MEDIA & SPORT. **Guidance on the Legal Deposit Libraries (Non-Print Works) Regulations 2013**. Available at <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/182339/NPLD_Guidance_April_2013.pdf>. Accessed on 20 March 2016.

DOUGHERTY, M. et al. **Researcher Engagement with Web Archives State of Art**. London: JISC. 2010 Available at: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997>. Accessed on 5 December 2015.

FILM PRESERVATION BOARD. **Film Preservation Study**. Washington D.C.: Government Office Public Hearing, 1993. Available at: <https://www.loc.gov/programs/national-film-preservation-board/preservation-research/film-preservation-study/overview/>. Accessed on 3 November 2015.

GUNELIUS, Sandra. The state of blogging in 2014. Available at <http://aci.info/2014/12/27/the-state-of-blogging-in-2014/>. Accessed on 15 March 2016.

HALLGRIMSON, Theo. Access and Finding Aids. **Web Archiving**. Berlin: Springer, p. 131-151, 2006.

HOCKX-YU, HELEN. Web Archiving at the British Library. Available at: <http://blogs.loc.gov/digitalpreservation/2011/08/helen-hockx-yu-web-archiving-at-the-british-library/>. Accessed on 12 March 2016.

INTERNET ARCHIVE. The way back machine. Available at: <https://archive.org/index.php>. Accessed on 10 April 2016.

INTERNET LIVE STATS. Internet usage & social media statistics. Available at <http://www.internetlivestats.com/>. Accessed on 16 May 2016.

KELLY, Brian and PENNOCK, Maureen. (2006) Archiving Web Site Resources: a Records Management View. Available at: < http://opus.bath.ac.uk/424/1/wwwpp100-pennock.pdf>. Accessed 10 March 2016.

KOBEL, Peter. **Silent Movies: the Birth of Film and the Triumph of Movie Culture**. Boston and London: Little Brown, 2007.

KUNNY, Terry. A Digital Dark Ages? Challenges in the Preservation of Electronic Information. **63rd IFLA Council and General Conference, 4th September**. Copenhagen, p. 1-12,1997. Available at: <http://archive.ifla.org/IV/ifla63/63kuny1.pdf>. Accessed on 10 May 2016.

MASSACHUTES INSTITUTE OF TECHNOLOGY. Web Growth Summary. Available at: <http://www.mit.edu/people/mkgray/net/web-growth-summary.html>. Accessed on 15 October 2015.

NATIONAL LIBRARY OF AUSTRALIA. Preserving and Acessing Networked Documentary Resources (PANDORA). <http://pandora.nla.gov.au/about.html>. Accessed on 2 May 2016.

NATIONAL LIBRARY OF AUSTRALIA. Preserving Access to Digital Information (PADI): Guidelines. Available at: <http://www.nla.gov.au/padi/topics/92.html>. Accessed on 8 October 2015.

NETPRESERVE. International Internet Preservation Consortium – IIPC. Available at: <http://netpreserve.org/about-us>. Accessed on 15 September 2015.

Preserving Access to Digital Information (n.d.) PADI: Guidelines. [online] Available at: <http://www.nla.gov.au/padi/topics/92.html> [Accessed 8 October 2010].

REELSEO, "Youtube creator rankings leaderboard – most popular YouTube channels March 2016." Availabe at <http://www.reelseo.com/top-youtube-channels/>. Accessed on 16 May 2016.

REGISTRO.BR. Web pages registradas no domínio *.br* desde 1996. Available at: <http://registro.br/estatisticas.html >. Accessed on 12 May 2016.

SOURCEFORGE. The WARC (Web ARChive) format, Version 9.0. Available at: <http://archive-access.sourceforge.net/warc/warc_file_format-0.9.html>. Accessed 15 November 2015.

STATISTA. Number of blogs worldwide from 2006 to 2011. Available at: <http://www.statista.com/statistics/278527/number-of-blogs-worldwide/>. Accessed on 2 May 2016.

THE NATIONAL ARCHIVES. Digital Repository Transfer System. Available at: <http://www.nationalarchives.gov.uk/information-management/our-services/digital-transfer-system.htm>. Accessed on 15 August 2011.

UK WEB ARCHIVE (UKWAC). Preserving UK Websites. Available at: <http://www.webarchive.org.uk/ukwa/info/about#what_uk_archive >. Accessed on 5 March 2016.

ISSN 1678-765X