CITY UNIVERSITY LONDON

Web Archiving in the UK: Current Developments and Reflections for the Future

Aquiles Ratti Alencar Brayner

January 2011

Submitted in partial fulfillment of the requirements for the degree of MSc in Library Science

Supervisor: Lyn Robinson

Abstract

This work presents a brief overview on the history of Web archiving projects in some English speaking countries, paying particular attention to the development and main problems faced by the UK Web Archive Consortium (UKWAC) and UK Web Archive partnership in Britain. It highlights, particularly, the changeable nature of Web pages through constant content removal and/or alteration and the evolving technological innovations brought recently by Web 2.0 applications, discussing how these factors have an impact on Web archiving projects. It also examines different collecting approaches, harvesting software limitations and how the current copyright and deposit regulations in the UK covering digital contents are failing to support Web archive projects in the country. From the perspective of users' access, this dissertation offers an analysis of UK Web archive interfaces identifying their main drawbacks and suggesting how these could be further improved in order to better respond to users' information needs and access to archived Web content.

Table of Contents

Tables and images

Acknowledgements

I would like to thank many people who helped me to conclude this dissertation. First I want to thank Dr. Lyn Robinson and Prof. David Bawden for their stimulating classes, extensive knowledge on Library Science and academic support throughout these last two years.

Second, I have to thank the British Library, especially Carole Holden, Elfrida Roberts and the BL Web archiving team for supporting me during my study – I would not be able to carry out this research without their assistance.

Last but not least, I would like to thank my partner, Christopher Connolly, for his patience and understanding, especially during this last semester when I had to spend many evenings and weekends working on this dissertation.

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 content that is now understood to be of considerable cultural value has been lost. The early films produced by the motion picture industry is one of the most 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 (2007) 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" (p. 275-6). In a report commissioned by the US congress in 1993, the National 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%" (Film Preservation Board, 1993). Today these few early silent films which were preserved for future generations are deemed to be invaluable cultural artifacts.

This is one of the many examples to 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. Today, we are at a similar stage in the history of Web pages. Many appreciate their present and future significance, while others remain skeptical about the importance of preserving Web content. As it has been commented recently on a Web archiving report: "While 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*., 2010a, p. 9).

Web pages are today an essential medium for publication, management and dissemination of information and their importance continues to increase at a fast rate. 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, 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. b).

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 in several countries and 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 and by the absence of copyright and deposit legislation permitting cost-effective large-scale harvesting as I will discuss in more detail later in this work.

The aim of this dissertation is two fold. In the first part I will provide a general overview on the development of Web archiving initiatives worldwide, paying particular attention to the development of the UK Web archiving consortium and reflecting on the progress and problems faced by UK Web archive institutions when implementing their Web archive programmes. In the second part of this work I will discuss, from a user's perspective, how current UK Web archive interfaces could be improved as to facilitate user navigation and information retrieval.

Part I: Current situation of Web archives

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 - Web archiving program launched by the National Library of Australia. 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 150 billion web pages occupying 2 Petabytes (PB), or 2,000 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, it has not sought permission from website publishers before harvesting copies. It has harvested without attention to rights issues, operating instead a 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 14 years. Today, it is archiving at the rate of about 170 titles, or 760 instances,[1] per month and has accumulated about 26,500 titles over 59,781 instances since the beginning of the project.

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 National Archives together with the British Library, the national libraries of Wales and

[1] 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.

Scotland, JISC and the Wellcome Trust embarked on a joint project, the UK Web Archiving Consortium - UKWAC, 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 the same Web harvesting software - PANDAS developed by the National Library of Australia - on a shared infrastructure, storing their collections in a single repository. This was in parallel with The National Archives' pre-UKWAC harvesting, which was transferred to the European Archive in 2005. UKWAC became publicly available in 2005 and remained active until 2009 when 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. In the next sections of this work I will discuss in more details the work developed by UKWAC and the current status of UK Web Archive under the leadership of the British Library.

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. UKWAC was founded by institutions that had a commitment to Web

archiving at the time of its foundation. As a matter of policy, the membership has not been expanded since the consortium was formed. Consequently UKWAC did not include three of the six legal deposit libraries (The Bodleian Library, Oxford; Cambridge University Library and the Library of Trinity College, Dublin).

Some important public sector institutions' Websites which were not captured by UKWAC's archival crawls because they fell beyond the consortium remits, are archiving their own pages. Examples include the BBC archive (although the British Library already has an agreement for the harvesting of the BBC News Website), the United Kingdom and Scottish Parliaments, and the Royal Household. According to Netcraft (n.d.), the BBC Website is one the most-visited Website in the UK. Its content is widely perceived as particularly valuable: the site is exceptionally complex, unusually large, and technologically advanced. The BBC started an in-house Website archiving initiative about six years ago. However, this initiative did not keep pace with the technological development of the Website; it cannot deal with interactive content of any kind, and is no longer considered to be fit for purpose. A new initiative is now under way to archive the Website including its interactive content. This will involve capturing, storing and preserving all the content components of the Website individually, along with the software that both combines them into viewable pages and provides interactivity. It is believed that the new BBC archiving solution may be operational within approximately two years, at least in part.

The approach chosen by the BBC is to archive all of its online content including messages on life chat forums, message boards and other interactive media features used on the BBC site. This archiving approach will become necessary everywhere, as more and more Websites develop interactive features that cannot be harvested by the 'traditional' currently established methods. This may represent a direction for the future, but it is far from clear that such an approach would be sustainable on a national scale. An important implication however is that it is unclear whether this new BBC archive can be integrated with the UK Web Archive partners' collections. The BBC has initiated discussions with the British Library

regarding the way forward (Bünz, 2009).

The United Kingdom Parliamentary Archive and the Public Record Office of Northern Ireland (PRONI) are both planning to archive their Websites. The former has been in discussion with UKWAC Partners, and is arranging to use the same European Archive service as is used by The National Archives, initially for a trial period. The latter has recently held discussions with the British Library and The National Archives concerning coordination of activities. PRONI is contemplating making use of The National Archives' contract with the European Archive to harvest government sites of interest in Northern Ireland (Digital Preservation Coalition, 2010).

By comparison with the Internet Archives and PANDORA holdings, UK Web Archive partners' collections are relatively small. Today about 5,300 titles have been harvested to the UK Web Archive collection. There are some 17,000 instances and the collection is growing at the rate of 400 - 500 instances per month. This takes up just under 4TB of storage. The National Archives is harvesting approximately 1,500 titles regularly and growing at roughly 250 - 300 instances per month and the National Library of Scotland is harvesting approximately 120 titles on a regular basis. Altogether, the UK Web Archive collections amount to about 10TB, equivalent to around 0.5% of the Internet Archive's holdings. By comparison, the ".uk" Web domain, containing approximately 5.5 million sites, is equivalent to about 2% of the 255 million Websites worldwide as reported by Netcraft (2010).

In a more global perspective, the International Internet Preservation Consortium has 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.) The International Internet Preservation Consortium has been also developing tools and carrying on research on Web archiving best practices and policies. 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 (Sourceforge, n.d.).

2. Web archiving: approaches and models

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 PADI (Preserving Access to Digital Information) 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. 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, 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 the national library aims to preserve. In the 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 contents. 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 which 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. Harvesting is selective, mostly depending on permissions granted explicitly by Website publishers, as I will discuss later in this work.

Despite working in partnership, each UK Web archiving institution follows its own collections policy. The National Library of Wales and the National Library of Scotland collect sites of interest to their respective nations; The National Archives aims to collect UK central government sites; JISC collects sites of projects funded by JISC; 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. In addition, the British Library archives a selection of sites that are representative of British social history and cultural heritage. It also archives a small number of sites which 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 (Highways Agency, 2008). 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 and Scottish national libraries;

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. As Hallgrimsson (2006) 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” (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 Internet has been characterized by rapid technological progress and even faster growth of its content. As of December 2010 there are an estimated 255 million Websites in the world, yet it is only 15 years ago that the number of Web pages reached one thousand (Massachusetts Institute of Technology, n.d.). This rapid 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, "[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.

The fast development in Web technologies creates an urgency for Web pages to be archived. At every moment electronic content are being changed, deleted or become simply lost when Websites are redesigned. Websites disappear as their owners or Web servers go out of business or they content might be removed by third party requests (legal suit, etc). In the majority of cases, Web content becomes inaccessible as technology changes. The remainder of this section presents some information to illustrate the scale of the challenge faced by Web archiving institutions towards the fast evolving Web technology, together with a short analysis of some of its implications.

3.1 The UK Web space

According to statistics from Nominet (n.d.), the agency that runs the registration of Websites for the .uk domain, there are approximately 8.96 million registered

domains in the .uk TLD; the number is growing at 10.9% per year, with much of the growth being attributable to the increase in blogging. This figure, however, needs to be taken into consideration since the actual number of Websites is normally lower than the number of domains. This happens, for example, when Websites move to a new domain. The old domain is maintained valid but users are redirected to a new Web address from which they can access the site content. Other common example of multiple domains leading to a single Web page title occurs in redirections by synonym. Visitors typing two different Web addresses such as http://www.worldpress.com and http://wn.com, for example, can access one same page for the World News site. Although the URI http://www.worldpress.com has a valid domain, it has no Website as visitors to it are redirected automatically to http://wn.com.

The rough approximations used by the Legal Deposit Advisory Panel is that 65% of current Web pages in the .uk domain needs to be archived and at least other 50,000 Web pages with a domain other than .uk are of interest for archiving institutions. According to the latest report produced by the panel, "approximately 5.4 million websites were potentially in scope for harvesting as at June 2010, rising to perhaps 14.2 million by 2020" (2010, p. 49). These statistics reinforce the urgent need for the implementation of a legal depository law in the UK, compelling authors and publishers to submit their page contents to Web archiving institutions reducing, therefore, the time spent by these institutions in contacting each Web page owner for asking permission to archive their pages. I will discuss this in more details in chapter 4 when dealing with deposit legislation for electronic content.

3.2 The changeable nature of Web pages

One of the main problems of Website preservation comes from the recognition 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 description

of the 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 for preservation which needs to be addressed in the dynamic context of how electronic documents should be archived. It is in the face of this situation that rules 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.[2]

The treatment of defunct Websites is another factor to be taken into consideration by Web archiving institutions. A study carried out by the Digital Curation Centre (DCC) in 2006 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

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

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 archived. According to the guidelines for Web archiving produced by the Central Office of Information (2009) and the International Internet Preservation Consortium (2006), 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 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 archiving external links to a specific Web page, archiving institutions are now seeking permission from content-sharing Websites such as Flickr, MySpace and YouTube

to archive their material. This would enable institutions to archive parts of Web pages which are linked to external content that refer to these content-sharing 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 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 (Rilling, 2006).

3.3 The evolution of Web 2.0 applications

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 (Garrett and Waters, 1996). However, the same does not apply for most UCC and Web 2.0 interactive content. The majority of this 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 “What is Web 2.0?” (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. Forrester (2008), an industry analyst, predicted in a recent report that business spending on Web 2.0 technologies will grow at an annual rate of 43%, reaching $4.6 billion globally by 2013.

Blogs are one of the most pervasive Web 2.0 applications. The use of blogs in Web page contents is rapidly increasing and their widespread use illustrates the scale and nature of the challenge presented to Web archivists. For libraries and archives, the archiving of blogs is an important issue that needs to be addressed in terms of cultural heritage preservation. Statistics on blog accesses and interaction prove how much this resource is significant for today’s society. Figures produced by Technorati (2010), a major blog search engine, reports that blog readership last year was in the range between 77 million to 94 million worldwide, which means that 77% of active Internet users visit or write in blogs. For the UK, blog usage statistics are less readily available; one kind of blogging, namely Twitter’s micro-blogging

service, is reported to have grown in the UK at a rate of 974% during last year ranking high amongst the most visited Websites in the country which makes the UK population the second biggest user of Twitter after the USA (Marketing Gum, 2010).

Turning to a different Web 2.0 technology, ComScore (2010) reports the number of videos accessed online in the UK reached an average of 5.5 billion each month in 2009. It is significant that the majority of the videos accessed on the Web are from UCC sites (YouTube, Megavideo.com, etc) rather than from broadcasters (BBC, ITV and others), 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.

In sum, as we have been discussing here, Website content, including Web 2.0 tools, plays an important role in how British people are generating and accessing information – key figures such as 5.5 billion video viewings per month and 974% annual growth in micro-blogging attest to this. Without Web archiving, most of this content – and some static Web content also – will within a few years be permanently lost. A delay in moving forwards with comprehensive Web archiving initiatives will lead to the long term and irretrievable loss of valuable cultural content.

4. Legislation

One of the major hindrances that limits the archiving of Web content is the lack of a strong legislation from the part of some national governments. There is an urgency for countries to implement a compulsory depository law for electronic materials, pressing publishers and authors to deposit publications in electronic format to their respective national libraries and granting unrestricted access to their contents. In the current legal framework, most of Web archiving institutions in the UK, except for The National Archives which deals only with government sites, have to obtain permission from a Website's publisher before they can harvest the site. The process of obtaining permissions is a major pre-occupation for Web archivists in these institutions, and it has a major impact on the nature of today's collections.

In order to comply with copyright laws, archiving institutions need to identify the intellectual property holder of a Web page and request by a contractual document permission to proceed with the archiving of its content. Taking into consideration the huge dimension of the UK Web space of approximately 5.5 million pages and with an annual growth estimate on 11%, we can easily see how labour intensive this process is, being most of the times more inefficient than effective. Requests for archiving permission in any significant volume requires software support and this was one of the key requirements that led the British Library to develop the Web Curator Tool (WCT) software (Sourceforge, 2010). Accordingly, WCT implements a database to record permissions requests and their status, and a workflow to manage the progress and its outcome. Figure 1 shows this workflow as outlined in a deployment flowchart from the British Library.

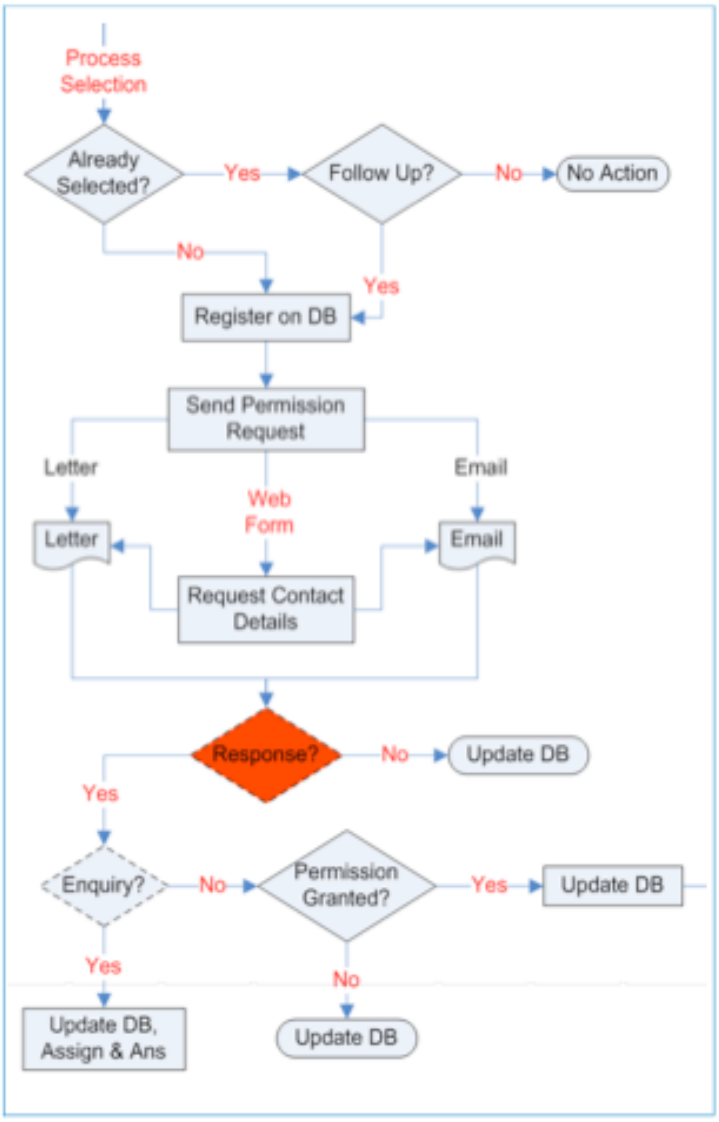

Figure 1: British Library permissions process

The labour content of the process is compounded by the low success rate of permission requests. On average only 25% - 40% of requests result in the granting of permission (Hockx-Yu, 2008). The low rate results from several factors, which include Website publishing organisations failing to respond, the difficulty of identifying the most appropriate individual to address in an organisation, and organisations feeling unable to grant permission because their Website includes material for which they do not own intellectual property rights (e.g. photographs licensed from a third party). The effort required to obtain permissions severely limits the number of titles that can be added to the collections. The low success rate results in collections that are incomplete. By contrast, The National Archives

and JISC do not require permissions for harvesting Web content within the scope of their respective institution archiving policies, and thus have been able to build stronger and more comprehensive collections.

4.1 Legal deposit

A number of memory institutions have been active in the UK in seeking legislation that will permit harvesting of all Websites relevant to the country. Primary legislation for the depository of electronic publications considered to be part of the UK cultural and intellectual heritage was passed in 2003, but the regulations required for it to take effect are still waiting to be put into practice. The 2003 Legal Deposit Libraries Act offers a revision of the Copyright Act implemented in 1911 which determined that copies of any material published in print format in the UK should be deposited in specific national and academic libraries, by including a new section on the deposit of non-print publications such as CD-ROMs and online publications. The added section on legal deposit fails, however, to address the specific regulations for publishers and authors to deposit their electronic content. As Adrian Brown (2006) remarks,

> the procedures governing the deposit of electronic publications will need to be defined in a series of regulations to be brought forward under the Act. Until such time as these regulations come into effect, the British Library is operating an interim arrangement… working closely with the UK publishing community to establish and encourage arrangements for the voluntary deposit of both offline and online electronic publications (p.158).

In the specific area of Web archive, depository laws are still failing to address and implement the necessary requirements for harvesting, archiving, maintaining and granting public access to archived versions of Web contents, obstructing in many ways the full development of Web archiving initiatives. Legislation to enable capturing of Websites is still waiting to be enacted as a follow up of the Legal Deposit Libraries Act of 2003. The Act will provide a legal basis for the harvesting

of Websites by deposit libraries without the need for permissions. However, as far as Websites are concerned, the Act will take effect only after appropriate regulations have been passed. As highlighted on the JISC report on legal issues concerning the archiving of Web pages (Charlesworth, 2003), the definition of the term publication as it appears referred in the Legal Deposit Libraries Act of 2003 is fairly ambiguous and it fails to mention documents available through the Internet as published works. As a consequence,

> the downloading and storage by one of the copyright depositories of material from a Web site, whether that site was based in the UK or elsewhere, would appear to be a straightforward infringement of copyright, in that such downloading and storage would inevitably involve the creation of unlicensed copies of the works that went to make up the webpage. In such circumstances, unless the agreement of the copyright owner was obtained in advance, web archiving in the UK without explicit permission from rightsholders would seem to place the budding archivist at risk of legal action (Charlesworth, 2003, p. 7).

Since 2005 the Legal Deposit Advisory Panel (LDAP), supported by the Joint Committee on Legal Deposit, has been working to draft more specific regulations concerning Web archiving by deposit institutions (Tuck, 2006). At present there is no certainty about the timing or content of such regulations and the exact nature of Web archiving by deposit libraries is still unclear. Responses to this situation have been negative, usually leading to a high level of frustration amongst users and institutions. A researcher's survey commissioned by the British Library on their expectations for the preservation of electronic material and the urgency to implement more precise regulations for the access to their contents shows that current copyright legislations are not only unfit for the digital age but, more importantly, are helping to create a "digital black hole" in the nation's cultural heritage. According to Lynne Brindley, the British Library's chief executive, "[t]here is a supreme irony that just as technology is allowing greater access to books and

other creative works than ever before for education and research, new restrictions threaten to lock away digital content in a way we would never countenance for printed material" (British Library, 2010).

Regulations concerning the archiving of Websites were expected to be implemented by 2010 (Tuck and Milne, 2008), but up to now there has been no change in the copyright law as to include new clauses on digital deposit. Once in force, the regulations will result in the British Library undertaking routine UK whole domain crawls. The Websites harvested by the whole domain crawls will be stored in a repository and access to content will be available to members of the public only in some deposit library reading rooms. There are some concerns that whole domain crawls will be shallow collecting and that a greater proportion of pages will contain more unresolved links than the current selective harvesting. Due to the vast number of pages and instances archived by the system in whole domain crawls, there are also some issues on quality control over the archived material including the frequency of Web page harvests and the way this material will be curated. The example provided by other national libraries which have attained copyright permission from their government to harvest Web sites from their national domain shows the limitations of whole domain crawling. As Web archivists from the National Library of France remark,

> [t]he goal of a large domain crawl is therefore to collect a representative sample of the national domain and to illustrate the French production at the time of the harvest. This sample is often designated as a snapshot – a way to record and to freeze a moving space. As it is not possible to harvest everything, we prefer harvesting few documents on every website rather than collecting entirely few websites, at the expense of others (Lasfarge et al 2008, p. 3).

The assumptions that whole domain crawls will be shallow and/or infrequent are especially significant, baring in mind that "[o]ur ability to use preserved materials in

the future, and the cost and quality associated with that use, are affected by what we do today" (Blue Ribbon Task Force, 2010, p. 28). As mentioned previously in this work, the estimate size of the UK national Web space is about 5.5 million Websites. This statistic, though approximate, is sufficiently accurate to establish that any ambition to archive the entire UK national Web space must require whole domain crawling. No combination of selective, thematic and deposit approaches could reasonably handle millions of sites. However, there is a widely held view that whole domain harvesting alone is insufficient to support a national Web collection, assuming it will be shallow and because whole domain harvests have reduced curatorial participation (i.e. quality assurance checking restricted to a small portion of the total pages and little metadata input). Some Web archivists believe that a shallow whole domain crawl will not capture enough of the national Web space to be useful, and that deeper, more curated, selective harvesting therefore needs to continue. The basic premise for defending more curatorial input to Web collections is that (Pymm and Wallis, 2008):

a. if whole domain harvesting is shallow, it will need to be supplemented by curated selective harvesting;

b. if whole domain crawling is not shallow, parts of the harvested content will need to be curated, in the way that they would have been if they had been harvested selectively.

It is worth emphasising that these concerns regarding whole domain archiving represents expectations of the UK Web archiving community (Thomson, 2010), not established fact. It is possible that by the time whole domain crawling starts technologies and budgets will have improved so much as to allow them to be as deep and as frequent as selective harvests. In any event, extension of copyright regulations including the deposit of electronic content in depository libraries should transform the UK Web archiving scene, for the better. They will allow all Websites to be harvested, regardless of whether the harvesting is shallow or deep (or indeed both); and will do away with the need to expend effort on seeking permissions. This

will liberate manpower resources currently assigned to permissions-seeking, potentially allowing these resources to be re-assigned to more beneficial curating. For the first time, it will be possible not only to collect comprehensive records of the Web in the UK, but also to build thematic, curated collections that are complete.

4.2 Other legal issues

As I have briefly discussed in chapter 3, what characterizes the behaviour of Web pages today is their dynamic nature which allows a myriad of possibilities for the inclusion of electronic documents and other multi-media contents such as audio and video into a single Website. Web 2.0 technologies offer users the option to post comments in Web pages visited, upload material and share content with other users dealing, very often, with material under copyright restrictions without obtaining formal permission from the content rights holder to do so. This situation heightens the risk for Web archiving institutions to incur into potential legal infringement, especially in the areas of Intellectual Property rights (IP), and content liability.

Copyright protection is a specific type of IP rights which establishes that any piece of creative work, no matter in which format it has been created, cannot be reproduced and/or disseminated without formal consent from its creator or rights holder. The concept of creative work is all encompassing, “including past school papers, letters and emails to friends and family” (Joint Information Systems Committee, 2008, p. 81). According to conventions approved by The Copyright (Librarians and Archivists) (Copying of Copyright Material) Regulations 1989 (DCMS), library and archives are given the statutory rights to reproduce copyrighted material without seeking formal consent from its rights holder as long as copies of the original material are made for preservation reasons only (i.e. migration to a new format) and that the archived material is not used for commercial purposes. Likewise the Legal Deposit Libraries Act of 2003, the lack of a straightforward copyright regulation that includes the preservation and use of electronic materials leaves margin for different interpretation on how libraries and archives should deal with electronic content in order to fulfill with copyright requirements. Format migration of documents from analogue to digital might be

considered an ‘adaptation’ of the original work and not necessarily treated as a preservation measure and which, according to copyright regulations, should not be carried out by archiving institutions without the right holder consent (Brown 2006, p. 149). In the specific case of Web archiving, copyright issues become much more complex. As explained before, current Web pages are dynamic in nature facilitating a high level of interaction among users. Visitors can post their personal comments or even upload or add links from third party material (photos, audio, etc) on a visited page making it impossible for archiving institutions to identify and seek permission from each page visitor to archive their contribution to the page before archiving the whole content of the Website.

In addition to seeking permission to collecting, preserving and granting access to copyrighted material, archiving institutions must also take account any liability regarding the archival of illegal content such as defamatory statements, obscene material and the promotion of illegal activities that might appear in the content of an archived Webpage. In order to minimize the risks of incurring in charges of archiving unlawful material, Web archiving institutions need to set up a programme of quality assurance aimed to check the archived instances on a periodical basis and make sure that the material harvested does not include any sensitive content which would put the institution on the risk of potential liability (Charlesworth, 2003). It is worth noting that existing Web page collection of some tens of thousands of instances is large; but not so large that quality issues cannot be addressed. However, the size of the collections will soon grow to a point that makes manual attention to quality impractical. In real terms quality maintenance would be impossible to be implemented for each and every harvesting made in a Web archiving collection, so institutions will need to rely on sample examples to carry out quality check.

When notified about the archiving of improper material or that a specific content is being disputed under defamatory law, Web archiving institutions are requested to immediately remove the material from public access and notify legal authorities about the content archived, bearing in mind that the “[d]estruction of the material should, however, be left to the authorities, as immediate destruction by the

archivist might hinder criminal investigations against the original supplier" (Charlesworth 2003, p. 21).

Even when the Legal Deposit Libraries Act becomes fully implemented granting national archiving institutions rights to harvest and preserve Web pages in the UK domain without the need to ask for explicit permission from their publishers or right holders, there is an assumption that Web content will be accessible solely in deposit library reading rooms (Field, 2010). This situation will be difficult to be reconciled with public interest; since access restriction is at odds with the ethos of making collections available as widely as possible. It is feasible that in the future archived Websites could be made openly available without restrictions to access but this could only be achieved by specific changes in current copyright regulations.

Part II: User access

1. Users

The collecting of digital content by libraries and archives is a recent endeavor: it was only in the second half of the 1980s that material recorded on digital media such as databases on CD-ROMs or floppy disks started to make part of library collections and became available for users. Digital material, although differing from other analogue resources such as printed books and recorded tapes, was usually treated as physical objects within a specific collection. It was catalogued following the same principles of traditional bibliographic records from which users were able to identify digital content in a physical location within the institution (Miller, 2000). Today, material published through the World Wide Web (a system of interlinked hypertext documents accessed via the Internet) raises new issues regarding how digital content becomes created, stored, disseminated and used. Archiving, organizing and granting access to digital content published on archived Web pages requires that libraries and archives adopt new approaches to deal with a multitude of hyperlinks and a variety of file formats in order to render users the same experience that they have when accessing live content on the Web. A recent report on the challenges faced by Web archiving institutions in delivering services that fulfill the expectations of today's Web users states that:

> [t]he introduction of new content formats (such as multi-media and dynamically executable content) has been accompanied by the evolution of completely new paradigms for content building and interaction, loosely grouped under the rubric Web 2.0, and particularly including user (and multi-user) generated content and the new social media platforms. All of these developments pose significant challenges to a web archive community which is still struggling to cope with Web 1.0 (that of largely static content) (Dougherty et al, 2010b, p. 5).

After years of discussing and outlining best practices for the harvesting and preservation of Websites, Web archiving institutions are just starting to turn their

attention to issues concerning public access for the material archived. According to the International Internet Preservation Consortium,

> [a]rchives of material published on the Internet are still in their infancy; the oldest archive (www.archive.org) is a mere 10 years old. It is conceivable that just as the advent of the Internet has forever changed the way people publish and exchange information, the availability of archives of Internet material will result in new and innovative usages of archives. Much of the current effort related to Internet archives focuses on collecting and preserving the information [as opposed to providing access to it] (2006, p. 1).

This situation, according to Dougherty (2010a), has not changed since it was first detected in 2006. It seems that Web archiving institutions are still concentrating their efforts in establishing rules for the harvesting and preservation of Web content, spending little time in liaising with users in order to better understand their needs and expectations: “large libraries and archives continue with their efforts to build large multi-purpose web archives, while researchers – either on their own, or partnering with archivists – develop their own archives for use in their research” (p. 8).

There is indeed - in comparison to the number of studies published on Web archiving methods, tools and preservation - a substantial lack of research on how users evaluate and engage with different Web archiving collections. One of the first studies on the importance of user’s survey in determining the relevance of archived pages, resource tools and types of services that Web archiving institutions should be providing was carried out by the National Library of the Netherlands in 2007. Looking at the activities of 16 Web archives around the world at that time, the study concluded that Web archiving institutions were working on an isolated basis and all their efforts to establish a Web archiving programme and access policies would be doomed to failure if they would not take into account the suggestions for service, access and archiving priorities required by potential Web archiving users (Ras and Van Brussel, 2007).

As these studies have suggested, different kinds of Web archiving users look for very different kinds of information in different ways. Some are looking for informational content and/or downloadable documents, others need to see the Web pages as they might have appeared in the past, while yet others are interested in the development of Websites over time. The use cases report published by the International Internet Preservation Consortium (2006) identify 17 possible kinds of Web archiving users according to their specific needs for retrieving information of Web pages that have had their content changed or removed from the Internet. These would range from arbitrary users (those who will use the archive as a tool to determine whether a page or a document has changed since the previous harvest or request an alert whenever a page is detected to have changed) to IT researchers (those seeking information on specified technologies, using archive to compare coverage of news items in different media Websites or looking for information on file formats, for preservation purposes). The report also rates the importance of Web archiving between the proposed user groups, claiming that legal institutions and researchers would benefit the most from the content archived in Web page repositories. Many organisations that have a regulatory role, or an interest in evidential issues, could have a stake in Web archiving, especially if whole domain crawling is introduced. This could include:

a. The Information Commissioner (for example, where incorrect information about an individual is published on a Website then corrected after the incorrect information has been archived);

b. The Internet Watch Foundation (for example where Websites have been archived that include child sexual abuse content, criminally obscene content, or incitement to racial hatred behavior and attitudes);

c. Police forces (when carrying out criminal investigations);

d. Courts (if archived content is put forward as evidence in litigation).

These organisations' stakes, suggests the report, are likely to increase significantly if whole domain harvesting is introduced.

Researcher's group would also benefit from Web archiving programmes in several ways. According to Webcite (n.d.), 13% of Internet references that are cited in academic articles disappear after 27 months of their publication, and there is a tendency for this rate to increase in the future if Web archiving institutions do not take any action for the preservation of academic content published on the Web. Since the number of online academic publications is growing, there is certainly an urgency for archiving its content before it disappears from the Internet. As Hallgrimsson (2006) suggests, "[f]or the scholarly community it will be important to preserve Web documents that have been cited or referred to. If they disappear the citations and references become meaningless and the original paper loses some of its value (p. 132). It is also important to note, as Hallgrimsson (2006) highlights, that information behaviour and research trends change over time and that in a feasible future Web archive "will be a resource on par with today's research resources like journals, letters, books, and media like radio, television and movies" (p. 132). Although still hard to predict how researchers will use the information accessed on Web archiving repositories and which topic and subjects will be relevant for their research, the International Internet Preservation Consortium report (2006) suggests that in the future research communities will benefit widely from Web archiving collections not only for providing access to a copy of an original electronic information removed from the Internet but also because they will indicate with accuracy how and when this information has been altered in a specific Website.

The use cases described in the International Internet Preservation Consortium report are hypothetical and varied. Naturally they all involve scenarios in which information can be obtained only, or best, from archived copies of Websites. A strength of the report is that it illustrates well the enormous variety of possible uses for a Web archive but, for the purposes of any deeper user survey, it may be

important to think of use cases as needing to take into account user questionnaires that approach at least three other classes of requirements, namely:

a. How and where will users be able to access Web archive content? This question should take into account the current limitations imposed by copyright restrictions. If, as the British Library (2010b) forecasts, “[m]obile devices will soon overtake personal computers as the most common web access tools worldwide” (p. 6), and given the current copyright legislation that prohibit remote access to archived contents, there is a need to investigate how Web archive users will respond to access restrictions and how effective Web archiving services will be for remote users.

b. How will users find what they need? This question should consider, among other things, how users evaluate current search engine tools and how relevant is the information provided by Web archive catalogues (i.e. metadata, classification schemes by subject area and/or thematic fields, etc) for the retrieval of the information needed.

c. How the archived material should be presented to the user? Closely related to the other previous questions, user’s surveys should look into current Web archive interfaces and get a feedback from users on how more efficient interfaces and access tools could be constructed as to respond to the demands of different user groups and in line with the latest development in Web 2.0 technologies. Users might like, for instance, to access Web archive in the same way that they use social network platforms. By logging into an account, users could save archived pages that they have previously accessed, share Web content with other users and even be able to make annotations on the content saved.

Perhaps a new comprehensive survey trying to map out user needs along these lines could help to distinguish other potential group of users and foresee use cases which were not identified in the 2006 International Internet Preservation Consortium report.

2. UK Web archive user statistics

Public knowledge of the existence of web archives appears to be very low, and knowledge of the UKWAC and UK Web Archive collections lower yet. This is consistent with the status of, and relatively low publicity for the archives, which characterizes the reduced use of Web archiving collections. It is reasonable to assume that the use of Web archiving will increase dramatically when Websites start to be harvested on a larger scale (i.e. if the need for archiving permissions is removed) and Web archiving initiatives become better publicised. It is important to stress that, because much of the use of Web page archives is likely to be in the distant future, the majority of prospective users of such archives have probably not yet been born. As pointed out before, Web archiving programmes are still relatively young, and because of being on a pioneering phase, issues of public access seem to attract little interest from the part of Web archiving institutions. However, the relatively low usage of the archives today must not be taken as evidence that Web archiving is not valuable and that it is not going to be more popular in the future.

The analysis of Web access statistics is notoriously complex, depending as it does on many definitions and on data gathered automatically by Web analytics software. This software often makes decisions which are not controlled by humans, and which cannot consistently measure all attributes of Web archive use. Accordingly, these figures must be seen as approximate. Traffic measurement is complicated further by a unique feature of The National Archives' collection, namely that visitors to government Websites can be redirected automatically to archived copies of pages of documents in certain situations (if the necessary software has been installed on the Website hosting server). In effect, these visitors become users of the National Archive Web collection, without having made a conscious request to do so, by redirection from a live site where the page no longer exists. Though the archived pages are labelled as such, users may also not realise that they are accessing documents from a Web archive collection. Figures for redirected access are shown separately in the table below, in the column related to 'automated

redirection' traffic, providing statistics on Web archive usage in the UK for the first semester of 2009:

	UK Web Archive (beta site)	The National Archives	
		The National Archives collection	Automated redirection
Unique visitors per day	180	704	694
Percentage of visitors who visit more than once per month	13%	13%	10%
Pages viewed per visit	3.02	5.5	2.2
Documents downloaded per day	11	313	391
Average duration of visit	1 min 13 sec	2 min 56 sec	2 min 19 sec

Figure 2: Traffic statistics for the UK Web archive and The National Archive (The National Archives, 2009)

Unfortunately UK Web Archiving institutions do not provide statistics on the number of visitors accessing the collection on a periodical basis. The latest published figures showing public visits to the collections were gathered between February and June 2009, the period when UK the Web archived launched its beta site after the dissolution of the UK Web Archive Consortium. Figures for the collections of The UK Web Archive and The National Archives are shown above, indicating the statistics of automated redirection from live Websites to archived content in the National Archives Web collection.

An early study carried out by UKWAC partners in September and October 2005 by the Digital Preservation Consortium reported about 8,000 unique visits per month. The above statistics for 2009 represent about 26,500 visitors per month excluding automated redirection visits (i.e. 180 + 704 per day for 30 days per month), indicating a growth of over 300% in under four years. From this we can conclude that traffic volumes are growing, but remain low for a national Internet resource. These results, however, seems to be consistent with the relative youth of the web archive collections and the limited publicity they have received.

2.1 Other Archives

Arguably the most relevant comparator for UK Web archiving institutions is PANDORA, the Australian national Web archive programme, relevant because of its greater wealth of content. The statistics available for user traffic in PANDORA for June 2009, the last period reported for UK Web archive access, shows that the site had approximately 1,000 visitors on a daily basis (excluding robots software), accessing an average of 3.63 pages per day. It is interesting to note that PANDORA's traffic is higher than the traffic to the UK collections (it represents roughly 3½ times as many accesses per citizen if automated redirection is ignored); and that the number of pages viewed per visit is close to the number viewed in the UK Web Archive collection, leading to the conclusion that different PADORA users tend to access mostly the same pages, at least for the specific month used here. Apart from making available user statistics on a monthly basis, PANDORA also publishes separate lists of popular search terms. A brief examination shows that among the meaningful search terms input by users, information of local importance figures highly. Examples of the top key terms used by visitors include "Tatiana Grigorieva" (a Russian athlete who took Australian citizenship in 1997), "first families", "Australian citizenship" and "2000 Olympics." It is fair to point out that a relatively high proportion of the search terms reflect curiosity about PANDORA itself rather than genuine research on its contents; roughly half of the top 40 search terms, accounting for one quarter of all search terms, include various combinations of PANDORA, Web archive, etc (The National Library of Australia, 2009). Another interesting information provided by PANDORA is the breaking down of the total number of visitor attributed to different countries. The UK appears in the statistics provided for June 2009 as the fifth largest source of visits after Australia, Japan, France, and the Netherlands. The US Internet Archive does not publicly report user's statistics.

Unfortunately, public information on the number of visitors for Web archiving collections in the English speaking countries is not provided by their respective institution or consortia, except for PANDORA which offers comprehensive data on the number of visitors, the countries from where the platform is accessed and the

main search terms adopt by users. By withholding statistic data on use and access, many Web archiving institutions are missing the opportunity to offer evidence to their respective governments on the relevance of Web archive for the general community failing, therefore, to obtaining more visibility and public support for their projects. By the same token, they will also fail to provide relevant information for researchers (historians, sociologists, informational professionals, etc) interested in analysing, now and in the future, the development of Web archiving initiatives.

3. Access paths

UK Web collections can be accessed in several ways. A public application programming interface (API) is available to support access to the UK Web Archives collections. The collections and parts of collections stored in the UK Web Archive are available through the main UK Web Archive website http://www.webarchive.org.uk, providing Web content access through a three level scheme: an alphabetical listing Web page titles, a browse option by subject of interest or through a list of 26 thematic collections. Users can also access content by a search facility that lists titles, and by direct entry of URIs. The interface also provides a visualisation function by which users can browse titles through snapshots of the main pages or cloud tags, but this only applies for the special collections feature and not for the whole content of the archive.

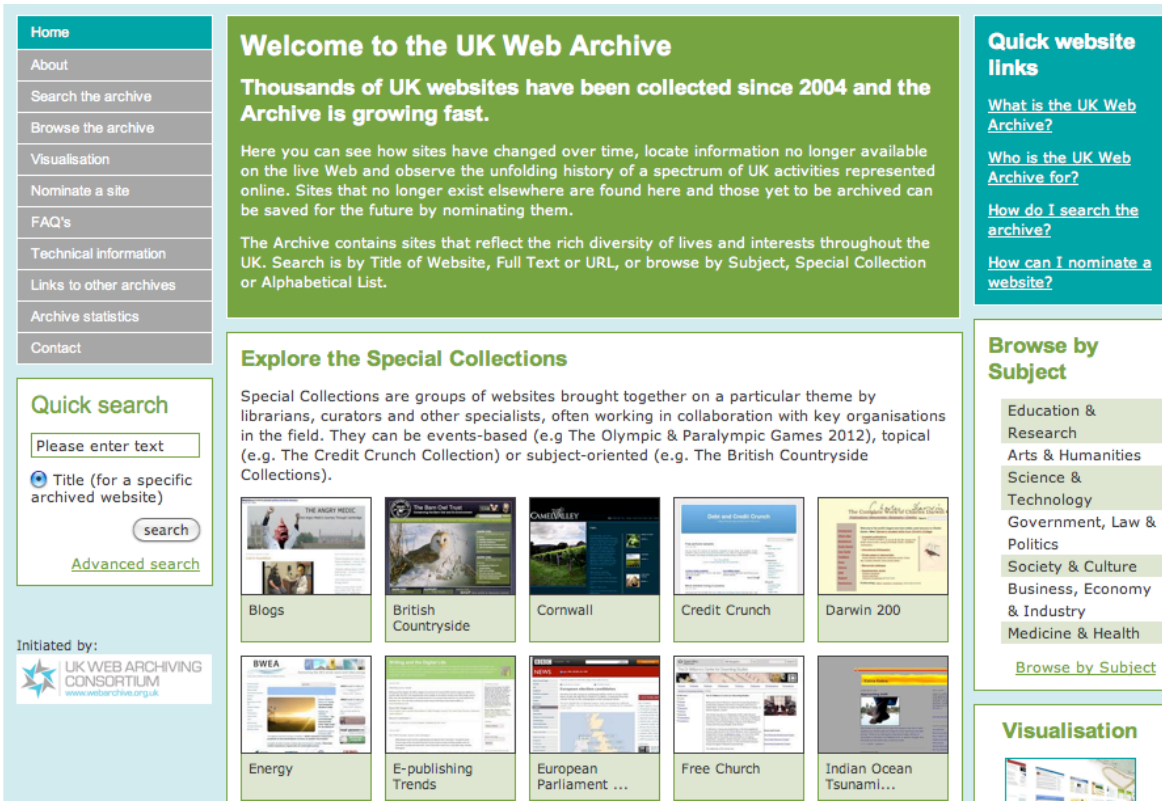

Figure 3: Interface for UK Web archive

The British Library provides an additional access path through a page on the Modern British section of its Website, http://www.bl.uk/reshelp/bldept/modbrit/webcoll.html. This provides a direct link to the UK Web Archive list of 26 thematic collections. A small number of Website collections are also indexed in the British Library's searchable integrated catalogue. The Wellcome Library uses its existing Online Public Access Catalogue (OPAC) http://catalogue.wellcome.ac.uk/ to provide access to its Website collection. The OPAC is scalable and offers powerful search facilities through its Web interface. Here the archived Websites are indexed with comprehensive metadata in the same way as other holdings such as books, periodicals etc., with comprehensive search facilities. The interface allows users to search either for archived Websites alone, or the entire collection that includes Websites, books, periodicals, etc.

Figure 4: Wellcome Library online catalogue – list of archived Webpages

Access to the National Archives titles other than those archived during the UKWAC partnership, are available through The National Archives' Web page http://www.nationalarchives.gov.uk/webarchive. This too offers a simple two level subject classification scheme, through an alphabetical listing by title name, by a search facility for title and direct entry of URIs. Two subsets of The National Archives' collections ("UKGOV weekly Web Archive" and "UKGOV Six Monthly Web Archive") are also available through the European Archives' home page http://www.europarchive.org/. Recently-collected holdings of the National Library of Scotland (NLS) are not publicly accessible, although there are plans to provide public access to this collection in the near future (The National Library of Scotland, n.d.). Users trying to access the NLS Web archive collections are redirected to the early pages crawled under the UKWAC partnership.

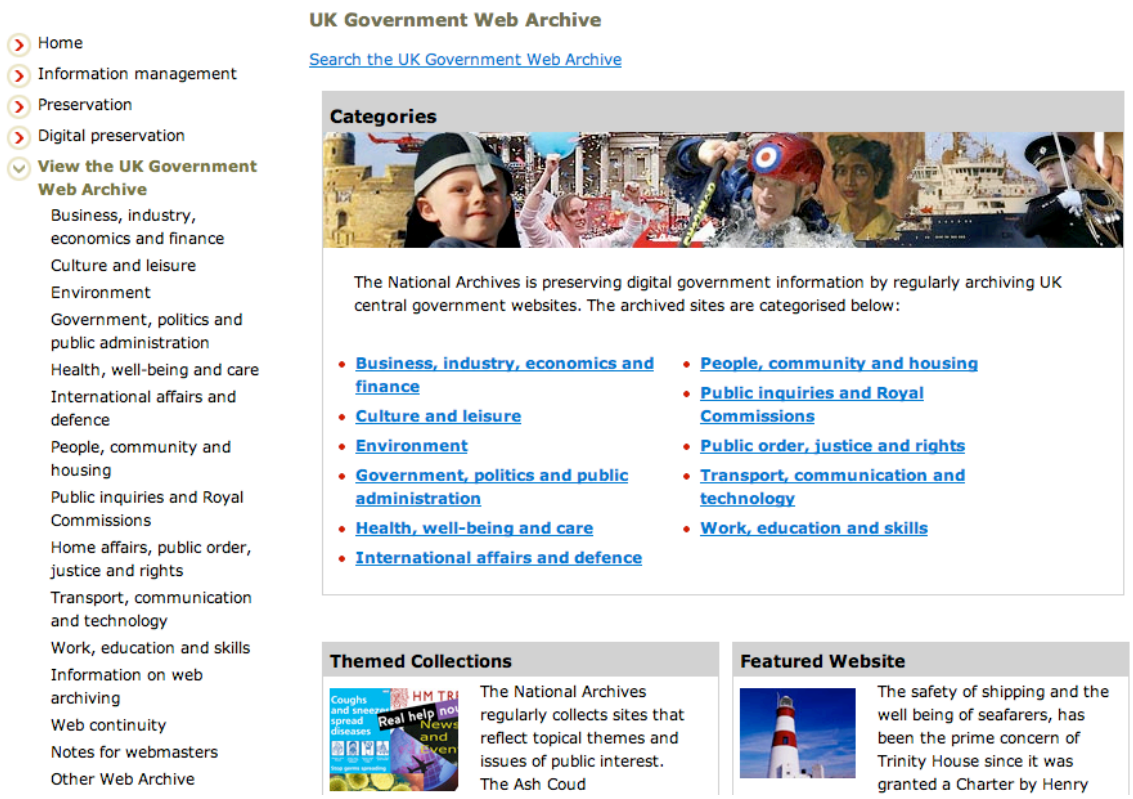

Figure 5: Web archive interface of The National Archives

In general, public access paths for UK Web archive collections offer limited search features that could certainly be improved, facilitating navigation and retrieval of information by users. The UK Web archive search features, despite failing to offer narrower scopes for search options, are more powerful than those provided by the Internet Archive and PANDORA. These do not offer any option for advanced search, restricting users to browse the collections through two options: search by Web page title or subject. There are, nevertheless, some limitations in the current search facilities offered by the UK Web Archive interface, and these could be improved in several areas. The examples below show some of the main problems encountered by users when trying to locate archived pages:

a. The lowest level of sub-categories such as "'Conditions and Diseases" from the classification heading "Medicine & Health" leads to a flat alphabetical list of 265 titles organised on 19 screens. There is no way to navigate the 265 titles except to page through the 19 screens. In order to provide more powerful retrieval feature,

UK Web Archive interface could offer a search facility within each sub-category listed on its main page.

b. The search through alphabetical lists offers very poor search functionality. If a user accesses, for instance, the heading “M” in the list of Website titles, he will be lead to a flat alphabetical list of 413 titles organised on 21 screens. Likewise, there is no other navigation option other than paging through the various screens.

c. The advanced title search is limited to combining searching for titles, URIs, subject and special collection offering, in this way, no other search parameter for the retrieval of more filtered results. Users would certainly benefit from filtering options for the retrieval of an archived version of a Web page that appears in a specific period or for obtaining information on the harvesting dates of a Website.

d. Users would benefit enormously from a broader search facility that not only retrieves information by Web page title but that would go deeper into the documents themselves looking for matching sentences and words as they appear in each single page archived. This feature would help users to identify information that deals if one same theme but has been reproduced in different pages which are not necessarily grouped together under the same subject classification. In this case, for instance, a user looking for information on ‘HIV transmission and treatment among drug users’ would be able to retrieve documents from various subject headings such as “Education and Research,” “Government, Law and Politics,” and “Medicine and Health.” This type of document retrieval system, used by Web search engines such as Google, should certainly be adopted by Web archiving institutions. Search engines can be envisaged for page titles and for content. The former is already implemented in the UK Web Archive interface. The latter is probably more helpful to users, though it is technically demanding for archiving institutions due to the large numbers of sites and pages and the large volume of data gathered in the archived collection.

These are but some few illustrations showing how search facilities on Web archiving interfaces could become more efficient for users. One thing is clear: if the

UK Web archive user interface remains unchanged it will become progressively less usable as the number of titles and instances archived increases. It may also be that in the future other national memory institutions (i.e. the National Archives of Scotland, the Public Records of Northern Ireland, other deposit libraries, etc) and major producers of Web content (broadcasters, universities and others) will join the project contributing to the formation of a single national collection of Websites. This will potentially raise de number of pages archived, making search for single titles and documents more arduous. The feasible future of whole domain harvesting after depository legislation have been approved, adding millions of new titles to the system, is another factor to be taken into consideration for the empowerment of current search facilities for the UK Web archive interface. With the order of hundreds or thousands of millions of single Web pages in the UK web space, and given the adoption of Web 2.0 features in an increasing proportion of these pages, it is clear that current UK Web collection interface needs to be adapted to better retrieve the sheer volume of information held by Web archiving institutions.

4. Metadata

Broadly speaking, metadata offers information about digital objects in the same way that cataloguing rules provides various levels of information to users and librarians about items hold in a library collection. Put in another way, metadata can be described as “structured data about digital resources that can be used to help support a wide range of operations” (UKoln, n.d.). Searchable metadata in Web collections can be applied at various levels of granularity. At the highest level, it can describe groups of titles (as with the present classification schemes); or it can refer to each individual title collected through selective harvesting. Metadata requirements such as those prescribed by the Open Archival Information System - OAIS (Digital Preservation Coalition, n.d.), the Metadata Encoding and Transmission Standards - METS classification compiled by The US Library of Congress (2008) and the Archiving Metadata Set proposed by the International Internet Preservation Consortium (2007) offer different guidelines and priorities,

usually leading to a plethora of metadata models not always compatible to each other. According to Dougherty et al (2010b, p.14), metadata description adopted by different institution should contain at least five essential fields of information:

1. *Provenance,* describing the custodial history of the object;
2. *Authenticity*, validating that the object is what it purports to, and has not been modified;
3. *Preservation activity*, describing actions taken to preserve the object;
4. *Technical environment*, describing the IT environment necessary to render the object faithfully;
5. *Rights management*, recording any property rights which may govern retention or publication of the object.

At present, most UK Web archive partners share a single repository, but perform Web archiving operations individually. Each institution makes selection decisions, executes the harvest, applies metadata to the material collected and most of this work is done in isolation from the other partners. It is probably not practical to consider applying the same metadata to individual instances, save perhaps in special cases. Partners have different requirements for indexing metadata for inclusion of Websites in their existing catalogues and that should be maintained as long as there is no conflictive information about a same digital object that has been duplicated in a shared collection. Apart from adopting the basic metadata descriptions, UK Web archive partner institutions should offer more contextual information about the objects archived in the collections, including the IP address from which a page was harvested to allow tracing of Website ownership at the time of publication.

If the essential driver of Web archive partnership is to harvest, grant public access and preserve indefinitely a number of Web pages considered to form part of the nation's cultural wealth, archiving institutions needs to produce collections that are totally compatible, so as to support any future collaborative presentation platform and resource discovery through sharable archives and Web interface. This applies mainly to similar harvesting procedures, curatorial selection and metadata

standards to be implemented by institutions. Therefore the choices here are to select a unique toolset, to be used by all partners or to agree on collecting criteria and then allow partners to use any toolset that adheres to the requested standards. Choosing a unique toolset has the advantage of ensuring compatibility; it also brings economies of scale for support and maintenance activities (Szydlowski, 2010). However, if partners identify requirements that are not shared, this becomes less advantageous. Agreeing on metadata standards allows partners to create coherent description of each archived Web page and select titles or combinations of titles that match their individual needs and perceptions.

5. Redirection

Both the UK Web Archive and The National Archives collections feature banners to highlight the fact that an archived Website is being viewed. This is a valuable feature given that the archived Web pages are retrievable and viewed in exactly the same ways as when their websites were live. Neither the Internet Archive nor PANDORA offer this feature. Informational banners are most useful for users when they are redirected to a Web archive when trying to access a URI that is not live or that has changed. Automated redirection, however can confuse users if it is not clearly signposted and explained: archived Web pages are clearly signposted in UK collections but archived content such as PDF files is not, due to technical limitations. UK Web archive institutions should work together in the development of software programs that could indicate when a PDF file has been changed or removed from public access through the World Wide Web.

6. Final considerations

Web archiving is a developing discipline, so some of the points raised in this section may be recognised by information professionals and archivists; while this allows us to understand and explain some of the issues, it does not help users to access archived content unless changes are made to UK Web archives' interface as proposed in this dissertation. In principle, options exist for the harvesting model – some combination of whole domain, selective, thematic, and deposit harvesting is needed. In practice, it is assumed that whole .uk domain harvesting will take place sometime in the near future; and it is certain that these harvests will not be heavily curated (because to do so would be prohibitive). It is therefore generally assumed that some fully curated selective and thematic harvesting (of some kind) will continue, with deposit harvesting perhaps playing a small part.

Web archives should appear as a single collection, regardless of how or where they are stored. In other words, collecting institutions may continue to expose their own collections of Web pages on their individual institutional Websites; but for the benefit of users they should additionally expose different Web collections in a single integrated view that includes all content archived by each institution. From a user standpoint, it matters little whether a national collection of Websites is implemented as a single repository, or as several repositories sharing a single access portal. The essence is to provide a single resource to allow users to access archived Websites, so that users do not need to worry about who owns the copy of a site before seeking access to it. The key recommendation of this work is that UK Web Archive partners, together with The National Archives and The National Library of Scotland, should work towards the creation of a "National Collection of Websites", integrating their collections to maximise their value. No particular form of integration is recommended; but several possibilities are identified, one such being the development of a portal that unifies access to separate Website repositories.

It is clear the present UK Web Archive membership consists of institutions that have demonstrated a commitment to taking a leadership role in web archiving. There would be benefit to enlarging the consortium, in particular to include other major cultural memory institutions, so users could have access to a more diversified and at the same time more complete collections of Web pages that are relevant for the cultural history of the UK.

Conclusion

Although still in its early stages of development, the importance and pervasiveness of the Internet and all of its accompanying digital contents is an undeniable characteristic of contemporary society. Given the changeable nature of digital objects and the rapid evolution of technology and new file formats applied to Websites, it is clear that archiving institutions need to begin archiving Web pages before their ephemeral content disappear altogether from the Web leaving behind an enormous information gap in the digital culture of our times for future generations. As I have briefly discussed in this work, memory institutions are becoming steadily aware of the urgency in preserving Web pages but, since Web archiving is a new activity initiated by these institutions, there is to date no final agreement about archiving standards, user interfaces, formalised strategic workflow, how to implement policies, how much should be archived, or even what to do with the resultant archives.

There are innumerable problems faced by archiving institutions when dealing with digital content. The limitations of harvesting software for crawling certain formats or script applications in Web pages is one of the major hindrances that needs to be overcome for Web archive repositories to be fully operational in the preservation of Web content. Short-term solutions such as conversion of more complex file formats and scripting languages into archivable content might offer an alternative for preserving digital material that cannot yet be harvested by crawling software. This situation, however, will become insurmountable in the future given the growing number of new file formats and diverse digital contents emerging in Web pages. To solve this problem, archiving institutions should work more closely with Web editors and software developers in order to create new products able to deal with the archiving of different Web formats and scripting languages without the need of human intervention for converting content into archivable material.

The lack of response from some national governments on issues related to copyright and deposit laws for digital material is another crucial point that stands in the way of how Web archiving institutions operate effectively. This is particularly problematic in countries like the UK where such laws are still waiting to be implemented. In my original dissertation project I intended to approach government plans for preservation and access of digital material published in the UK in the light of the Digital Britain report published by the labour government in 2009. The report, which presented the government's vision on the importance of digital content for the country and its aim to facilitate the inclusion of different sectors of the population into the digital world, is an analysis of the advantages of the Web in supporting the nation's economic growth:

> [s]hort term economic pressures have exposed areas of policy and regulation that need to be addressed, however, Digital Britain primarily seeks to position the UK as a long-term leader in communications, creating an industrial framework that will fully harness Digital Technology. The UK's digital dividend will transform the way business operates, enhance the delivery of public services, stimulate communications infrastructure ready for next-generation distribution and preserve Britain's status as a global hub for media and entertainment. Most importantly of all this approach seeks to maximise the digital opportunities for all of us, as citizens, where access to 21st Century technologies will be a key competitive advantage for generations to come (Department of Media, Culture and Sports and Department for Business, Innovation and Skills, 2009, p. 10).

According to the above statement, the report highlights the need for government intervention in providing public access to the Internet not from a cultural perspective but from an economic standpoint. Archiving and access to web content is not addressed in the report. The few passages discussing issues of copyright regulations and preservation of digital material are elusive and only show the position of the government in postponing the debate to another future report as evidenced in this excerpt:

> [t]he Government is however considering the scope to amend the copyright exceptions regime where we believe exemptions exist, in areas such as distance learning and the preservation of archive material and intends to announce a consultation on these later this year. Clearly, on the broader question of modernisation of fair use rights, further work remains to be done (p.113).

In the present context, UK Web archiving institutions need to develop a strategic plan to persuade the government about the importance of preserving the country's Web heritage for future generations and the implications that this activity has on the economic growth of the nation, bearing in mind that the preservation of and access to cultural content either in analogue or digital format is:

> besides an indispensable element for social cohesion and the reconstruction of an identity, an economic sector equally or even more important than any other productive sector of society. The economic transactions that take place in the deepest heart of culture generate positive economic effects such as learning and knowledge (Inter-American Council for Integral Development, n.d., p. 2).

In the second part of this dissertation I highlighted the importance of Web archiving institutions in interacting more closely with users, carrying out research on their information requirements, browsing experience and the materials they expect to find when accessing Web archive collections. By having more input from users, archiving institutions would be able to develop stronger online interfaces and collection catalogues in response to user needs and expectations. Following this argument, I indicated some browsing limitations I have encountered while using the UK Web archive interface and suggesting, where appropriate, how searching facilities could be empowered as to facilitate the retrieval of material requested by users.

In order to succeed in developing and offering service for the preservation and access to archived UK Web pages, archiving institutions need to create a single management system that will ingest, store and share the same metadata standards for Web material, providing better access to their archived collection. They must also ensure that the material is authentic and easy to find, and that users can view archived contents with contemporary applications, experiencing, where possible, the material in its original look-and-feel.

As a final recommendation, I have proposed that UK Web archive institutions should work more closely in partnership to avoid duplication of harvesting efforts and collections. They should also make their Web archiving collections available via a single access path and, if possible, invite other institutions (press, Web publishers, corporate business, etc) that are archiving their own Web content to join the consortium in making their material available to public users.

Bibliography

Anderson, P. (2007) What is Web 2.0? Ideas,Technologies and Implications for Education. [online] London: JISC. Available at: <http://www.jisc.ac.uk/whatwedo/services/techwatch/reports/horizonscanniing/hs0701.aspx> [Accessed 9 August 2010].

Bailey, S. and Thomson, D. (2006) UKWAC: Building the UK's first Public Web Archive. *D-Lib Magazine*, 12(1). [online] Available at: <http://www.dlib.org/dlib/january06/thompson/01thompson.html> [Accessed 17 August 2010].

Blue Ribbon Task Force (2010) Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. [online] Available at: <http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf> [Accessed 15 August 2010].

British Library (2010a) Driving UK Research: Is Copyright a Help or a Hindrance? A Perspective from the Research Community. [online] Available at: <http://pressandpolicy.bl.uk/imagelibrary/downloadMedia.ashx?MediaDetailsID=628> [Accessed 18 November 2010].

_____ (2010b) 2020 Vision. [online] Available at: <http://www.bl.uk/aboutus/stratpolprog/2020vision/2020A3.pdf> [Accessed 5 December 2010].

British Settlers in Argentina (n.d.) British Settlers in Argentina and Uruguay – Studies in 19th and 20th Century Emigration. [online] Available at: <http://www.argbrit.ar> [Accessed 7 October 2010].

Brown, A. (2006) *Archiving Websites: a Practical Guide for Information Management Professionals*. London: Facet Publishing.

Brügger, N. (2005) *Archiving Websites: General Considerations and Strategies*. Aarhus (Denmark): The Centre for the Internet Research.

Bünz, M. (2009) BBC and British Library to Take Joint Approach to Building Digital Archive, *The Guardian*, 11th December. [online] Available at: <http://www.guardian.co.uk/media/2009/Dec> [Accessed 8 July 2010].

Central Office of Information (2009) Archiving Websites. [online] Available at: <http://coi.gov.uk/guidance.php?page=239> [Accessed 16 June 2010].

ComScore (2010) UK Online Video Viewing up to 37 Percent During Last Year. [online] Available at: http://www.comscore.com/Press_Events/Press_Releases/2010 [Accessed 12 October 2010].

Charlesworth, A. (2003) *Web-Archiving: a Feasibility Study for JISC and the Wellcome Trust*. London: JISC and Wellcome Trust. [online] Available at: <http://www.jisc.ac.uk/whatwedo/programmes/preservation/webarchiving.aspx> [Accessed 15 August 2010].

Crook, E. (2009) Web Archiving in a Web 2.0 World. *The Electronic Library*, 27(5), pp.831-836.

Day, M. (2006) The Long-Preservation of Web Content. *In:* Julien Masanès (ed). *Web Archiving*, Berlin: Springer, pp. 177-199.

Department of Media, Culture and Sports (1989) The Copyright (Librarians and Archivists) (Copying of Copyright Material) Regulations. London: Her Majesty's Stationery Office. [online] Available at: <http://www.legislation.gov.uk/uksi/1989/1212/made> [Accessed 26 September 2010].

_____ (2003) Legal Deposit Libraries Act 2003. London: Her Majesty's Stationery Office. [online] Available at: <http://www.hmso.gov.uk/acts/acts2003/20030028.htm> [Accessed 12 September 2010].

_____ and Department for Business, Innovation and Skills (2009) Digital Britain. London: Her Majesty's Stationery Office. [online] Available at: <http://www.official-documents.gov.uk/document/cm76/7650/7650.pdf> [Accessed 12 March 2010].

Digital Preservation Coalition (n.d.) Reference Model for an Open Archival Information System (OAIS). [online] Available at: <www.dpconline.org/component/docman/doc.../284-digital-curation-reich> [Accessed 20 September 2010].

_____ (2005) UKWAC Evaluation Report. [online] Available at: <http://www.dpconline.org/component/docman/doc_download/15-ukwac-evaluation-report-ukwac-evaluation-report - 61.75 kB> [Accessed 15 September 2010].

_____ (2010) Who's Who: Sixty Second Interview with Iain Fleming, Public Records Office of Northern Ireland. [online] Available at: <http://www.dpconline/org/newsroom/whats-new/532> [Access 9 October 2010].

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

_____ (2010b) Researcher Engagement with Web Archives: Challenges and Opportunities for Investment. London, JISC. [online] Available at http://ssrn.com/abstract=1715000 [Accessed 5 December 2010].

Eisenschitz, T. and Shiozaki, R. (2009) Role and Justification of Web Archiving by National Libraries: a Questionnaire Survey. *Journal of Librarianship and Information Science*, 41(2), pp. 90-107.

Field, C. (2010) Proposal on the Collection and Preservation of UK Offline and Microform Publications. [online] Available at: <http://rylibweb.man.ac.uk/Rag2/…/ConsultationresponsetoHMG-LegalDeposit.pdf > [Accessed 18 November 2010].

Film Preservation Board (1993) Film Preservation Study. Washington D.C.: Government Office Public Hearing. [online] Available at: <http://www.loc.gov/film/hrng93dc.html> [Accessed 3 November 2010].

Forrester (2008) Forrester Press Release. [online] Available at: <http://www.forrester.com/ER/Press/Release/0,1769,1207,00.html> [Accessed 22 July 2010].

Gomes, D. et al. (n.d.) Design and Selection Criteria for a National Web Archive. [online] Available at: <http://xldb.fc.ul.pt/daniel/docs/papers/gomes06tomba.pdf> [Accessed 8 July 2010].

Hallgrimsson, T. (2006) Access and Finding Aids. *In:* Julien Masanès (ed). *Web Archiving*, Berlin: Springer, pp. 131-151.

Garret, J. and Waters, D. (1996) Preserving Digital Information. [online] Available at: <http://www.clir.org/pubs/reports/pub63watersgarrett.pdf> [Accessed 3 September 2010].

Highways Agency (2008) Web Archive Policy. [online] Available at: <http://www.highways.gov.uk/…/Highways-Agency-Web-Archive-Policy.pdf> [Accessed 20 August 2010].

Hockx-Yu, H. (2008) Archiving the UK Web. [online] Available at: <http://www.bka.gv.at/DocView.axd? CobId=32143> [Accessed 15 November 2010].

Inter-American Council for Integral Development (n.d.) Culture as an Engine for Economic Growth, Employment and Development. [online] Available at: <http://www.oas.org/udse/english/documentos/tema1estudio.doc> [Accessed 20 December 2010].

Internet Archive (n.d.) The Way Back Machine. [online] Available at: <http://www.archive.org/web.php> [Accessed 10 July 2010]

International Internet Preservation Consortium (2006) Use Cases for Access to Internet Archives. [online] Available at: <http://netpreserve.org/publications/iipc-r-003.pdf> [Accessed 12 September 2010].

_____ (2007) IIPC Web Archiving Metadata Set. [online] Available at: <http://iwaw.europarchive.org/05/masanes2.pdf> [Accessed 12 December 2010].

Joint Information Systems Committee (2008) The Preservation of Web Resources Handbook (PoWR). [online] Available at: <http://jiscpowr.jiscinvolve.org/wp/files/2008/11/powrhandbookv1.pdf> [Accessed 16 September 2010].

Jrank (2009) BBC.co.uk – History, Content. [online] Available at: <http://encyclopedia.jrank.org/articles/pages/1522856/bbc-co-uk.html> [Accessed 8 July 2010].

Kelly, B. and Pennock, M. (2006) Archiving Web Site Resources: a Records Management View. [online] Available at: <http://www.dcc.ac.uk/docs/WWW2006> [Accessed 10 March 2010].

Kobel, P. (2007) *Silent Movies: the Birth of Film and the Triumph of Movie Culture*. Boston and London: Little Brown.

Koerbin, P. (2008) The Australian Web Domain Harvests: a Preliminary Quantitative Analysis of the Archive Data. [online] Available at: <http://pandora.nla.gov.au/documents/auscrawls.pdf> [Accessed 18 November 2010].

Kuny, T. (1997) A Digital Dark Ages? Challenges in the Preservation of Electronic Information. *In: 63rd IFLA Council and General Conference*, 4th September, Copenhagen. [online] Available at <http://archive.ifla.org/IV/ifla63/63kuny1.pdf> [Accessed 20 July 2010].

Lasfarge, F. et al. (2008) Legal Deposit of the French Web: Harvesting Strategies for a National Domain. [online] Available at: <http://iwaw.net/08/IWAW2008-Lasfarges.pdf > [Accessed 7 August 2010].

Legal Deposit Advisory Panel (2010) Impact Assessment: Digital legal Deposit – Non-Print Off Line Publications. [online] Available at: <http://www.culture.gov.uk/images/publications/Impact_assessments-digitallegaldeposit2011.pdf> [Accessed 15 November 2010].

Marketing Gum (2010) Twitter Statistics for 2010. [online] Available at <http://www.marketinggum.com/twitter-statistics-new-stats-for-2010/> [Accessed 20 December 2010].

Massachusetts Institute of Technology (n.d.) Web Growth Summary. [online] Available at: <http://www.mit.edu/people/mkgray/net/web-growth-summary.html> [Accessed 15 October 2010].

Miller, R. (2000) Electronic Resources and Academic Libraries, 1980-2000: a Historical Perspective. *Library Trends,* 48(4): pp. 645-670.

Netcraft (n.d.) Most Visited Sites. [online] Avaliable at: <http://toolbar.netcraft.com/stats/topsites?c=UK> [Accessed 17 October 2010].

_____ (2010) UK Web Statistics, December. [online] Available at: <http://news.netcraft.com/archives/2010/12/01/december-2010-web-server-survey.html> [Accessed 20 December 2010].

Netpreserve (n.d.). International Internet Preservation Consortium – IIPC. [online] Available at: <http://netpreserve.org/about/mission.php> [Accessed 15 September 2010].

Nominet (n.d.) Statistics of Web page Registration. [online] Available at: <http://www.nominet.org.uk/intelligence/statistics/registration/> [Accessed 22 December 2010].

Palmer, J. (2009) Archives 2.0: If We Build it, Will They Come? [online] Available at: <http://www.ariadne.ac.uk/issue60/palmer/> [Accessed 3 October 2010].

Pedley, P. (2006) *Essential Law for Information Professionals*. 2nd edition. London: Facet.

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].

Public Record Office of Northern Ireland (n.d.) Online Archives. [online] Available at: <http://www.proni.gov.uk/index/search_the_archives.htm> [Accessed 19 October 2010].

Pymm, B. and Wallis, J. (2008) Archiving the Web: Does Whole-of-Domain Archiving = Information Overload? [online] Available at: <http://www.information-online.com.au/sb_clients/…PresentationB8.pdf> [Accessed 7 October 2010].

_____ and Simes, L. (2009) Legal Issues Related to Whole-of-Domain Web Harvesting in Australia. *Journal of Web Librarianship*, 3: pp. 129-142.

Ras, M. and van Brussel, S. (2007) Web Archiving Survey. [online] Available at: <http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/KB_UserSurvey_Webarchive_EN.pdf> [Accessed 24 October 2010].

Riley, J. (2009) Glossary of Metadata Standards. [online] Available at: <http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards_glossary_pamphlet.pdf> [Accessed 7 December 2010].

Rilling, P. (2006) Harvesting Web Content into MHTML Archive. [online] Available at: <http://www.codeproject.com/KB/IP/mhtmllib.aspx> [Accessed 16 August 2010].

Ruddock, B. and Stevenson, J. (2010) Moving towards Interoperability: Experiences of the Archives Hub. [online] Available at:

<http://www.ariadne.ac.uk/issue63/stevenson-ruddock/> [Accessed 18 October 2010].

Samouelian, M. (2009) Embracing Web 2.0: Archives and the Newest Generation of Web Applications. *The American Archivist*, 72(1): pp. 42-71.

Scott, K. (2010) Archiving Britain's Web: the Legal Nightmare Explored. [online] Available at: <http://www.wired.co.uk/news/archive/2010-03/05/archiving-britains-web-the-legal-nightmare-explored> [Accessed 7 September 2010].

Seneca, T. (2009) The Web-at-Risk at Three: Overview of an NDIIPP Web Archiving Initiative. *Library Trends,* 57(3): pp. 427-441.

Steinke, T. (2008) Harvester results in a digital preservation system. [online] Available at: <http://www.nla.gov.au/padi/metafiles/resources/34803.html> [Accessed 22 August 2010].

Sourceforge (n.d.) The WARC (Web ARChive) format, Version 9.0. [online] Available at: <http://archive-access.sourceforge.net/warc/warc_file_format-0.9.html> [Accessed 15 November 2010].

_____ (2010) Web Curator Tool. [online] Available at: <http://webcurator.sourceforge.net> [Accessed 12 October 2010].

Szydlowski, N. (2010) Archiving the Web: It's Going to Have to be a Group Effort. *The Serials Librarian*, 59: pp. 35-39.

Technorati (2010) State of the Blogosphere in 2010. [online] Available at: <http://technorati.com/blogging/article/state-of-the-blogosphere-2010-introduction/> [Accessed 20 December 2010].

The National Archives (n.d. a) DCMS Legal Deposit Consultation – Response from The National Archives. [online] Available at: <www.culture.gov.uk/.../National_Archive_-_Consultation_response_Digital_Legal_Deposit.rtf > [Accessed 16 October 2010].

_____ (n.d. b) Digital Repository Transfer System. [online] Available at: <http://www.nationalarchives.gov.uk/information-management/our-services/digital-transfer-system.htm> [Accessed 15 August 2010].

_____ (2009) Traffic statistics for UK Web archiving and The National Archive. [online] Available at: <http://www.nationalarchives.gov.uk/TNA%20UKWAC/Report/Deliverables/TNA%20traffic %20statistics> [Accessed 5 September 2010].

The National Library of Australia (n.d.) Preserving the Present into the Future. [online] Available at: <http://pandora.nla.gov.au/pan/49714/20080925-1031/www.unimelb.edu.au/records/web-archiving/index.html> [Accessed 11 April 2010].

_____ (2009) Pandora Monthly Statistics, 2009. [online] Available at: <http://stats.nla.gov.au/_reports/pandora/monthly/6-2009/awstats.pandora.keyphrases.html> [Accessed 23 September 2010].

The National Library of Scotland (n.d.) Digital Projects. [online] Available at: <http://www.nls.uk/about-us/working-with-others/digital-projects> [Accessed 23 July 2010].

The US Library of Congress (2008) METS Metadata Enconding & Transmission Standards. [online] Available at: <http://www.loc.gov/standards/mets/ mets-schemadocs.html> [Accessed 15 October 2010].

The United Kingdom Parliament (n.d.) Parliamentary Archives. [online] Available at: <http://www.parliament.uk/business/publications/parliamentary-archives/> [Accessed 23 October 2010].

Thomson, R. (2010) British Library Archives UK Websites with New Software. [online] Available at: <https://www.computerweekly.com/.../British-Library-archives-UK-websites-with- new-software.htm> [Accessed 13 October 2010].

Tuck, J. (2006) Collecting, Selecting and Legal Deposit. *In: DPC Forum on Web Archiving British Library,* 12th June, London. [online] Available at: <www.dpconline.org/component/docman/.../193-web-archiving-forum-tuck> [Accessed 8 October 2010].

_____ and Milne, R. (2008) Implementing E-Legal Deposit: a British Library Perspective. [online] Available at: <http://www.ariadne.ac.uk/issue57/milne-tuck/> [Accessed 20 October 2010].

UKlon (n.d.). Metadata. [online] Available at <http://www.ukoln.ac.uk/metadata/> [Accessed 20 August 2010].

Webcite (n.d.) Webcitation. [online] Available at: <http://www.webcitation.org/> [Accessed 11 October 2010].