LEGAL AND TECHNICAL DIFFICULTIES OF WEB ARCHIVAL IN SINGAPORE*

JHONNY ANTONIO PABÓN CADAVID**
JOHNKHAN SATHIK BASHA***
GANDHIMANI KALEESWARAN****

Introduction

Web archiving is a recent area of study that emerges at the end of the 20th century the purpose of which is to avoid a digital black hole in world history. The increasing publication of born digital information through the World Wide Web, and the international recognition of such publications as part of cultural heritage –digital heritage- have alerted governments to the significant national value of preserving web publications.

In Asia only South Korea, Japan, China and Singapore are doing national web archiving in their respective national libraries. Web archival is an issue of

* This is a short version of our final critical inquiry research paper at NTU 2013 and a conference presented at IFLA World Library and Information Congress Singapore 2013. Thanks to Ang Peng Hwa (Director of the Singapore Internet Research Centre, NTU) for his guidance and to Peter Lor (South Africa) and Brian Opie (New Zealand) for their comments and suggestions.

** Ph.D. Candidate, Victoria University of Wellington, New Zealand. Magister Science (MSc) in Knowledge Management, Nanyang Technological University, Singapore. Master of Arts in History (2008), PUJ, Colombia, LLB, Universidad Externado de Colombia, and has worked as a researcher on Intellectual Property in Colombia and New Zealand. E-mail: pabonjhon@myvuw.ac.nz

*** MSc in Knowledge Management, Nanyang Technological University, Singapore. Bachelors in Production Engineering from PSG college of Technology, Coimbatore, India (2008) and has worked with IBM for IT Consulting in Chennai, India. E-mail: SATHIKBAOOI@e.ntu.edu.sg

**** MSc in Knowledge Management, Nanyang Technological University, Singapore. Bachelor's degree in Engineering in Stream of Information Technology at Anna University, Chennai, India (2008) and has worked in the software industry as application engineer. E-mail: KALEESWAOOI@e.ntu.edu.sg. Fecha de recepción: 18 de marzo de 2014. Fecha de aceptación: 17 de mayo de 2014. Para citar el artículo: Pabón Cadavid, J.A.; J.S. Basha y G. Kaleeswaran, "Legal and Technical Difficulties of Web Archival in Singapore", *Revista La Propiedad Inmaterial* n.º 18, Universidad Externado de Colombia, noviembre de 2014, pp. 35-54.

international concern. Its development depends on cultural, legal, economical and political values in each country, which define the scope, objectives, functionalities and importance of web archival.

The complexity of web archival brings new legal challenges for the management of digital collections at the international and national level, especially within legal deposit legislation and copyright law. These difficulties cover the spectrum from the harvesting of contents to the use of the Web Archive by users.

Political-ethical aspects of web archiving

Web Archiving faces several ethical issues in any of the different areas of its information management cycle. What is to be preserved? Who is responsible for preservation? Who decides who gets access, under what circumstances, to what materials? These questions underline archival practices in the face of Web Archiving, but there are some characteristics of digital information and Web Archives that raise specific ethical issues.

One big difference between print and digital publications is the sense of permanence, which "is a vital principle of cultural heritage: the raison d'etre of collecting is to retain a cultural identity and to build up the resources—the cultural and research collections—that permit cultural enrichment, facilitate research, and bring wider social and economic benefits to the society that supports and finances that collecting activity" (Mason, 2007). Ephemeral print publications have been archived on a very small scale, but in Web Archiving a large number of digital publications that by intention tend to be ephemeral could be collected and preserved in the long term (Lor & Britz, 2012): for instance, twitter messages; comments in forums, opinions in articles, Facebook posts and other user-generated content which creators do not expect would be preserved by memory institutions for intergenerational purposes.

Another issue that can arise from web archiving is censorship, by which control over intellectual production in the Web can be exercised. For instance, a Singaporean caricaturist and blogger received a notification from the National Library regarding the archiving of his website and blog. He expressed his fear of vigilance for archiving in a comic way, commenting: (archiving is) "Bad in the sense that I got to be careful of what I post. Hahaha…" (Portrait Workshop, 2007). In addition, censorship through web archiving can be effected by limiting access to materials for researchers; the intentional deletion of websites or negligence to collect and preserve some websites for political or other reasons is a form of censorship affecting present and future researchers. For instance, copyright law could be used with political motivations to order the deletion and removal of material archived.

In some situations Web archives are research projects involving extremely sensitive issues as the creation of web archives of the dark web for the analysis of

terrorism (Chen et al., 2008). National and thematic web archives must follow ethical guidelines that guarantee to minimize possible harms.

National identity and National Library of Singapore

The development of National Library of Singapore (NLS) services in recent years has been in line with the aim of "nurturing a society of life-long learners who can accelerate the creation of intellectual capital and a new cycle of national innovation" (Seet, 2005, p. 151). Web archiving as a new tool for the National Library is oriented towards achieving these goals.

The objective of Web Archiving by the National Library Board (NLB) is to create a collection of Singaporean-originated digital information which has been published on the Internet, which registers different facets of the culture and heritage of Singapore. In addition it is expected that the Web Archive will help:

> to achieve a sense of community, national identity and rootedness among Singaporeans. This can be achieved by archiving information that shapes the national identity. The web is increasingly used as a tool for social communication and interaction. Over time, it would form a record of events that captures the milieu of a nation, which tracks how Singapore's national identity develops and evolves. Archiving this record provides an invaluable source of documented heritage for Singapore's present and future generations. This understanding will create a sense of community and belonging, communal feelings commonly fostered by a good and strong archive (National Library Board, 2012).

Web Archiving highlights new challenges for digital information, especially regarding new fields of study such as digital heritage and digital humanities. Digital Humanities is a new area applying computational techniques to research and communication in the humanities (Hayles, 2012).

Archives themselves do not create a sense of community and nationhood. It is access to and use of the archives that can achieve that outcome. One example of researching about national identity in Singapore using web information is the recent article titled "Singapore: The Politics of Inventing National Identity" which clearly shows how the Internet has been used by Singaporeans for rethinking and expressing their ideas of nationhood (Ortmann, 2009).

Questioning how NLB defines digital heritage and how digital humanities are nurtured in Singapore will assist in identifying the principal characteristics of NLS services required to produce the benefits offered by the Web Archive. Without clear policies for web archiving and for access to the contents of the Web Archive the purpose of WAS regarding national identity cannot be fulfilled.

Web archiving in Singapore (WAS)

Current scope of legal deposit in Singapore

With the creation of the National Library Board (NLB) in 1995, legal deposit functions in Singapore were made the responsibility of the NLB within the premises of the National Library of Singapore (NLS). One of the functions of the NLB is "To acquire and maintain a comprehensive collection of library materials relating to Singapore and its people" (Section 6, NLB Act Chapter 197, 1995). Library material is defined in a broad sense in Section 2, including any material or data that could be reproduced (Section 2). The interpretation of this section is that legal deposit in Singapore covers digital information that is distributed in a physical medium (such as DVDs, CDs, etc.) but online information is outside of the scope of current legislation.

In 2004, the NLB created a Task Force on Legal Deposit with the aims of reviewing the legislation and developing a strategy to build a broader collection of documentary heritage, including online information (Foo, 2005). The Task Force did a comparative study of legal deposit legislation, including issues related electronic deposit (e-deposit) materials and web archival. It recommended the immediate introduction of electronic voluntary deposit and in the long term a new legal deposit framework that would amend the NLB Act to include on-line material as part of heritage material under legal deposit (Singapore Report to CDNL, 2005-2006). Some eight years later, this amendment is still pending.

Development of web archival in the National Library of Singapore

In 2005 the Singapore Internet Research Center (SIRC) initiated the Asian Tsunami Web Archive in conjunction with the Internet Archive, and WebArchivist.org. Some members of the Task Force, especially academics from Nanyang Technological University, participated in this project. The project lasted about two months and archived approximately 1,600 websites, from around 40 different countries in 13 different languages. In the second stage of the preparation for the project it was decided as copyright policy to implement a takedown notice procedure (Wu & Heok, 2005). Nowadays the website of the Asian Tsunami Web Archive http://tsunami.archive.org/ is not available; displaying an HTTP error 404 - The page cannot be found-. In addition, that web archive has not been used for researching or any other purpose.

In 2006, with the experience gained from the Asian Tsunami Project and following the recommendations of the Task Force the NLS started Web Archive Singapore (WAS). WAS is aligned with the objective of the NLS to collect materials related to Singaporean national identity. WAS's appraisal criterion is to collect websites of national significance that could serve for future researchers (Chellapandi & Lian

San, 2009). It adopted a selective and thematic archiving model, selecting around 1,000 websites, especially from government agencies and official organizations. In addition, for events such as the Singapore General Elections in 2006, selected websites were harvested more frequently (Chellapandi & Lian San, 2009).

In 2007 NLB started to harvest the domain .SG, also on a selective basis (by 2009 NLS had the capacity of archiving 500 websites per month). The Singapore Network Information Centre (SGNIC) is the organization in charge of the administration of the national domain .SG. There are more than 100,000 domains registered under the .SG domain. In 2009, NLB signed a Memorandum of Understanding with SGNIC to facilitate access to the list of domains using .SG (Chellapandi & Lian San, 2009). With the use of the National Grid Pilot Platform more than 20,000 websites have been archived.

Without legal deposit legislation that allows NLS to do Web Harvesting, NLS is using a takedown policy and in some cases sending notifications (through letters or e-mail) to website owners regarding archiving and access to the materials (Chellapandi & Lian San, 2009).

The Task Force took the Web Archival congress held in Australia in 2004 as a point of reference (Foo et al., 2005). The development of WAS replicated some features of the PANDORA project of Australia, for example, the selective approach, the taxonomy driven interface for surfing the archive and is operating without adequate legal framework to facilitate crawling the entire national domain and to copy any material of interest to the National Library.

With the publication of the Library 2010 Report (National Library Board, 2005) the NLS started a group of initiatives to provide broader access to electronic resources, including the creation of the Digital National Library (including digitalization of the documentary heritage part of its existing legal deposit collections), the encyclopedic Singapore Infopedia, NewspaperSG (in 2007 the NLS established a copyright agreement with Singapore Press Holdings to digitize the Straits Times and made the newspaper internet accessible) and the ongoing project Singapore Memory (Chellapandi, Han, & Boon, 2010). In recent years, other memory institutions such as the National Archive of Singapore have had a proactive role in providing open access to documentary heritage (Beasley & Kail, 2009).

The mission of NLB stated in the Library 2010 Report "is to bring the world's knowledge to Singapore to create a positive social and economic impact" (National Library Board, 2005). However, Web Archival and born digital material as critical parts of Singapore's digital heritage were not mentioned in that report and were not included in the strategic plans framed by NLB in 2005.

Copyright challenges of web archiving in Singapore

The digital records management of web archiving entails several copyright issues with respect to collection, preservation, access and utilization of those collections.

International intellectual property treaties and agreements to which Singapore is a signatory, such as the Berne Convention, the TRIPS agreements and the bilateral Free Trade Agreement with the US (USSFTA), frame Singapore's copyright law. Nonetheless, recently the NLB stated that Singapore is working towards and supports flexibility in copyright law:

> NLB is working closely with the International Federation of Library Associations and Institutions (IFLA), the World Intellectual Property Office (WIPO) and the Intellectual Property Office of Singapore (IPOS) to lobby for changes in the copyright legislation in order to educate and sensitise the legislators to the need for library exceptions and exemptions in the Intellectual Property Rights legislation, both at an international level and in Singapore (Copyright Act). The NLB is revising its current Act to incorporate digital legal deposit (NLB Singapore, 2010).

It is not clear how NLB and IPOS are lobbying internationally, if at all, but at the national level the amendments for legal deposit have been pending for almost 10 years.

Singapore Copyright Law establishes that any reproduction of a protected work is an infringement unless there is authorization by the copyright holder or a legal limitation to copyright. Singapore's Copyright Law does not have any specific exception or limitation for either Web Archiving or for digital preservation. There is no legal certainty about the legality of web archiving in Singapore under the umbrella of the fair dealing clause.

Section 35 of the Singapore Copyright Act establishes a fair dealing general clause, which allows flexible criteria to evaluate a non-copyright infringement act depending on the following factors (Section 35(2)): the purpose and character of the dealing, including whether such dealing is of a commercial nature or is for non-profit educational purposes; (b) the nature of the work or adaptation; (c) the amount and substantiality of the part copied taken in relation to the whole work or adaptation; (d) the effect of the dealing upon the potential market for, or value of, the work or adaptation; and (e) the possibility of obtaining the work or adaptation within a reasonable time at an ordinary commercial price.

1. Collection

Copyright issues

As was mentioned WAS has passed through several stages of development, from a very selective approach of websites to a more comprehensive approach of harvesting the national domain in whole or part. Harvesting websites entails a reproduction of the contents. The NLS during the selection and crawling of websites adopted an opt-out policy, doing snapshots of the websites and sending to website administra-

tors a "Notification of Website Archiving" which set out some advantages of the Web Archive: "There are some benefits to you as a website owner in having your website and online publications archived by the NLB. You will be able to access earlier versions of your websites from the web archive. The NLB will also catalogue your publication, thereby increasing awareness and publicity of your website and publication among researchers using our Library network". In addition the notification states that: "The NLB reserves the right to take down any material from the web archive which in its opinion infringes Copyright". Accordingly, when the archiving of web material is done there is no assessment of any copyright issue; it is later through a Takedown policy that material could be deleted from the archive if someone requires it or the NLB decides to delete the information. A similar policy of take down is followed in the Netherlands (GLANVILLE, 2010).

The takedown or opt-out policy is the one implemented by the Internet Archive, and it has been proposed as an advantageous policy for web archives in the US (PATEL, 2007). However there is a risk of litigation because of this legal uncertainty, a risk that could be absolutely avoided by adequate legal deposit legislation like that of New Zealand, England or France. It is important to emphasize that legal deposit cannot be an absolute solution because it has territorial and technological limitations, for example, in New Zealand the National Library required to get permission for harvesting overseas content and local content protected by passwords (PAYNTER & MASON, 2006).

According to the WAS website the current policy of the website is to exclude from archiving material which they do not have copyright authorization for harvesting: "Note also that images and content that belongs to a Third-Party Copyright owner may not be captured as part of the archive due to copyrights concerns" and "NLB does not capture and archive information that hosted in other domains or servers" (WAS, NLB, 2013a).

2. Retention and preservation

Copyright issues

WAS has adopted a takedown policy where copyright holders can require NLB to retire the material archived. The takedown policy is very similar to the one adopted by the Internet Archive. NLB established a procedure of out-put notification by copyright holders and after investigation, when "the grounds for complaint are considered plausible, the material will be permanently withdrawn from the repository" (WAS, NLB, 2013b). Without legal deposit legislation permanent retention of the material collected is not granted, and depends on the agreement of copyright holders.

Web Archives in National Libraries are considered part of the cultural heritage of the country; as heritage, digital collections must be protected through a long-term preservation strategy. Digital preservation entails several uses of digital

contents affecting different exclusive copyrights. Limitations to rights to reproduce transform and, in some situations, "make available", are necessary if digital preservation is to achieve the goals set for it (National Digital Information Infrastructure and Preservation Program (U.S.), Joint Information Systems Committee, & OAK Law Project, 2008).

Provisions of the Copyright Law of Singapore regarding preservation are outdated and are insufficient in the digital realm. Sections 48 and 113 of Singapore's copyright law allow the National Library to make a reproduction of a work for preservation purposes. But this exception seems insufficient, considering that long-term digital preservation requires format shifting and emulation; both activities involve adaptation-transformation of the work that without a legal exception could result in copyright infringement. As a Singaporean scholar has pointed out, "an Archive has no definite assurance that its ingestion and storage processes will be fair dealing" (SENG, 2008).

3. Access to and use of web archives

If collecting and preserving information in national web archives raises complex management and technical issues, the provision of access is not different. One particularly sensitive area is copyright.

WAS provides access to a very small part of its collection. According to a recent interview with NLB stuff, they are waiting for an amendment to the NLB Act which will allow them to make a wider range of archived material accessible on the Internet (Yong Fu, 2013). Preservation provisions in Singapore's copyright law are insufficient for long-term access, allowing a narrow construction to authorize only reproductions of the work, impeding making available the web archive collections (SENG, 2008).

Sites that are accessible in the WAS website are the ones that have granted permission to the National Library or belong to the government. In the Archival notification the NLB, addressing website administrators, states that: "NLB has embarked on a web-archiving project, with the long-term goal of building a comprehensive collection of Singapore-related websites and publications to ensure that Singaporeans have access to their documentary heritage now and in the future. NLB has deemed your website xxx to be an important part of Singapore's documentary heritage and would like it to remain available to researchers and generations of Singaporeans in the future" (Appendix 1). However, this access could be stopped by the use of the takedown policy.

The only national library that allows open access to its Web Archive is the Icelandic Web Archive of the National and University Library of Iceland (www.Vefsafn.is), which restricts access only to material that requires payment and implements an out-put access policy.

The grounds for the restriction of accessibility to Web Archiving are that copyright does not allow making available works without prior authorization of

the copyright holder and that use can affect the economical interest of copyright holders. However, there is no empirical evidence that accessibility to Web Archives has an adverse economic impact (National Digital Information Infrastructure and Preservation Program (U.S.) et al., 2008).

Countries with Web Archives that provide on-site access authorize it with the objective of research or study, as educational and research exceptions to copyright law. In those circumstances the use of the Web Archive has been limited because the lack of copyright flexibility prevents the use of tools that digital humanities has adopted in recent years, for example, researchers in the National Library of France are facing restrictions in the use of the French Web Archive for data mining analysis (Stirling, Illien, Sanz, & Sepetjan, 2012).

Web archives have been used in many research projects such as the analysis of social actions, web historiography, and the ethical impacts of web archiving (Dougherty et al., 2010). The use and engagement of web archives will depend of the usability, completeness and also publicity of the archives.

Since its inception was and the nlb has failed to give information to the citizens about the project. A Singaporean Wikipedia community user discussing web archival notification by the nlb shows the failure of the nlb when describing his experience: "We made several calls to nlb to enquire about this heritage thing, no one knows... If this is meant for our heritage, it should be known to as many people as possible, and I was passed around all morning from one department to the next and they had no clue what this is about" (SG Wiki Editor, 2007).

Orphan Works

Orphan works are works for which the copyright holder cannot be identified or located (Copyright Office USA, 2005). This situation is a problem, because it generates a legal barrier for the use of these works. Copyright law established that any use of protected works required the authorization of the copyright owner and only by exception is the use of a copyright work without permission allowed.

Currently there is neither a solution to this problem or discussion about orphan works in Singapore. The relation between orphan works and web archives is that contents harvested could remain unused in the future if they lack clear identification of authorship and copyright.

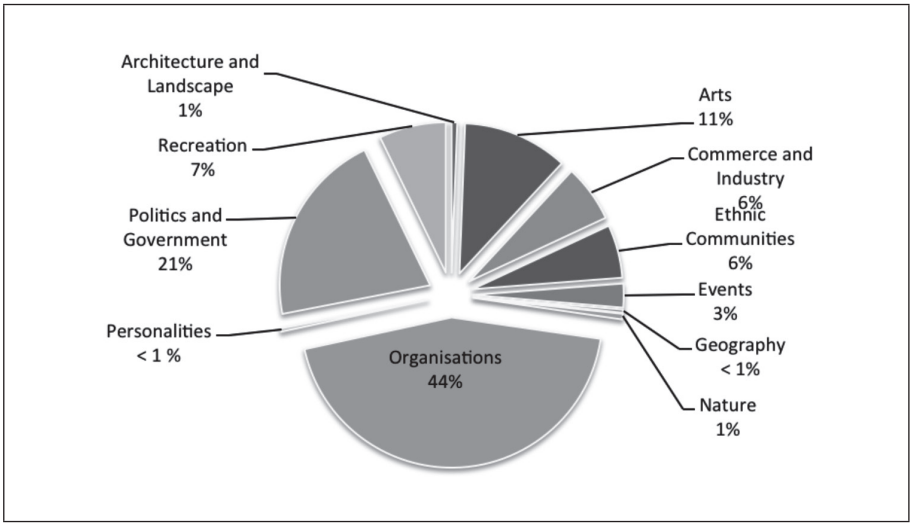
Technical issues

web archive taxonomy

Web Archive Singapore classified 1174 websites under 11 broad categories. Figure 2 shows their distribution. Organizations form the largest group with 44% (514) followed by Politics & Government (21%) and Arts (11%). The remaining categories

have less than 10% in each group. Out of these, 4 (Architecture and Landscape, Geography, Nature, Personalities) have 1% or <1% of the total archived websites. The category 'Personalities' has only 4 websites and most of well known Singaporeans (Lee Kuan Yew, Lee Hsien Long and so on) web links and their websites are ignored, a broad spectrum of websites of Singaporean people should be harvested so that Singaporean culture is represented as fully as possible. There is no criterion that explains why the specific archived contents of sites possess characteristics relevant to cultural heritage.

FIGURE 1. WEB ARCHIVE CATEGORIES IN WAS



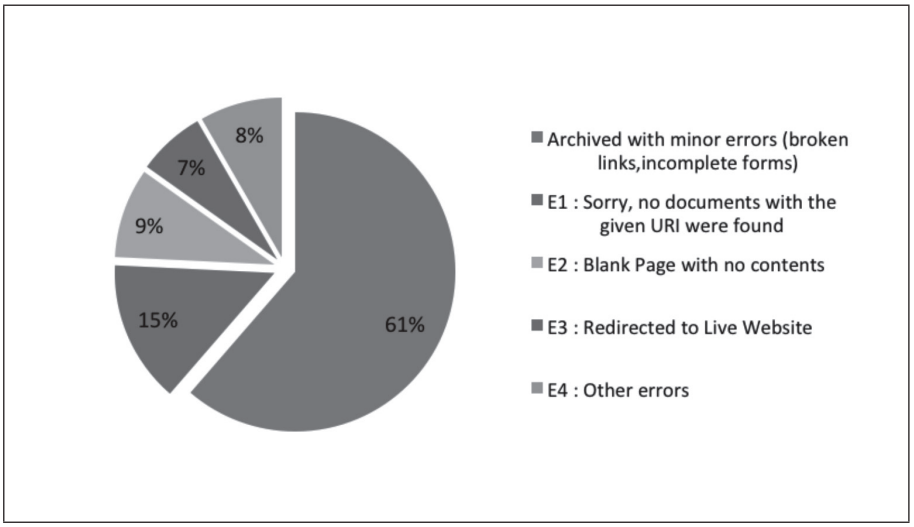
The 'Arts' category was chosen for examination with the aim of identifying the common errors encountered in the archived websites. Since the site included a note, "Some links, forms and search boxes may not function properly", only the archived contents, which did not show up any information, were deemed as error contents. Errors like websites with browser compatibility issues and websites with only a home page link were grouped under 'Others'.

The distribution of errors in the Arts category is set out in Figure 2. 39 % of the archived websites in this category have some errors in their last two versions. This evaluation indicates the low content quality of the WAS archives and stresses the need for technical improvements in the archiving process.

TABLE 1. LIST OF ERRORS ENCOUNTERED WHILE EXAMINING THE 'ARTS' CATEGORY

S.No	Defects /Errors
1.	Blank Page
2.	Browser Compatibility issues (Opens in one browser and not in other browsers)
3.	Main page appears and all other sub links lead to live websites
4.	NoSuchKey The specified key does not exist.top1.html
5.	Page remains forbidden
6.	Redirecting to Live Website
7.	Sorry, no documents with the given URI were found
8.	Status code 404 Page Not Found error

FIGURE 2. ERRORS ENCOUNTERED IN THE ARTS CATEGORY



Search and API's

The expectations and interest of experienced users of archive contents differ from each other. Scholars and researchers might look for web documents for their research and would cite or refer to those documents in their work. For generic users, who use the Internet as a communication medium, the archive might not be valuable. Currently most of the archived contents are preserved for research purposes.

In any case, a user might expect search features like those available in the live web. There is an urgent need to develop a better ways of accessing archived contents. Most web archive interfaces today are restricted to a URL-based lookup

with a basic search function using URL typing and categories. Wayback Machine, WERA (Web Archive Access) and Hanzo-WARC tools have only basic functions for searching archive content (COSTA SILVA, 2009).

With present developments in web archival it is only possible to index the text alone and not the actual audio/video file or the image files. This text indexing is done from the crawler-generated metadata present in file headers. Some web archive interfaces use Google for indexing and searching. Use of search engines allows automatic indexing and is quite influential as they are capable of managing and searching thousands of millions of documents. However, there is a need for web archive search tools like NutchWax[1], an extension of the Nutch web search engine for WARC/ARC file search, to meet the specific needs of web archive content searches. Though large-scale search engines (Google, Yahoo) claim to have image search and video search, they are still unable to provide accurate results from huge collections (LI WANG, 2008).

Cataloguing is an important factor that impacts the search function in web archives. Since libraries carry out large scale web archiving, they tend to apply traditional cataloguing procedures to cataloging web documents, which limits the effectiveness of the search functions. Traditional cataloguing is a not suitable option for all archiving practices except for selective archiving, which is usually small-scale (HALLGRIMSSON, 2006). This is because of the fact that only a low percent of web documents contains reliable descriptive metadata. In fact some studies show that cataloguing is not a viable option for providing access to archived contents; the final reports of the Minerva Project conclude that "The Library should rely on automatic indexing as the primary means for information discovery of websites and content within websites. The full text of all textual materials should be indexed on a periodic basis and searchable by users," and "The Library should not invest in extending MARC cataloguing or other manual methods of cataloguing to websites, except for sites of particular importance to the Library and its users" (MASANS, 2006).

Terminology evolution study

Full name (LiWA) is a project funded by the European Commission which focuses on extending current state of the art web archiving practices by focusing on innovative methods for content capturing, improving coherence and dealing with semantic evolution. One of the objectives of the LiWA project is to find terminology evolution in web archiving which is something that occurs in long time preservation. As society changes, the use of language also changes and this needs to be addressed in web archival (RISSE, 2011). For example, in a search for music players a user currently will type the word "ipod", but previously would

1. [http://nutch.apache.org/].

have used Discman and Walkman. A researcher in the future who wants to do research on a particular topic must be aware of all the earlier terms in order to find relevant information. Similar evolution might occur to location names and to names representing roles (chairman, president) as both the person holding a position and its title changes over time. Terminology evolution can impact web archival retrieval and usage in future research. To keep a web archive semantically accessible, LiWA has proposed clustering such common words within an archive as a solution. This would need to consider two new layers, an architectural layer and a linguistic and semantic layer which would map the terms used with their intended meaning. This kind of study is significant in archiving social media as it evolves very quickly and users of social media modify noun phrases and create new words very frequently.

The NLS could encourage the development of SOAP-based or other commercial web service interfaces that would deal with changes in the live web. In addition, the NLS must support development of visual tools for searching metadata.

Web analytics

Beyond the technical issues in archiving, institutions and organizations involved in web archiving practices must look to answer the following questions to enhance the user experience with archived contents.

– How do users look for relevant archived content (keywords used, user navigation path etc.)?

– How can low precise and un-indexed archived content be filtered?

– What type of archived content do users want to access?

– How can archived content be used more effectively in terms of index searching?

– How can usage patterns across archived web content be discovered?

Other similar questions could be answered with web mining techniques done in the live web. Web mining is the process of using data mining techniques to extract knowledge from the archived data such as log files which contain versions, network connection speeds, web servers, operating systems, and hyperlinks (COOLEY, SRIVASTAVA, and MOBASHER, 1997). This data centric view of web mining is more acceptable among researchers for its approach and it suits the needs of web archive mining.

Most of the institutions undertaking web archiving initiatives do not show any interest in preserving log files of web usage and focus only on web harvesting and external archiving processes. But it is essential to preserve log files that can be used for internal record management and web mining.

Web mining processes would help users to get relevant content during a search. It encompasses different techniques, such as web content mining, web usage mining, and web structured mining. In addition, web mining techniques would offer a better user experience by providing quality based ranking and display related web archive content.

Web mining could derive information based on the data collected throughout the lifecycle of web archiving. The data for web mining can be categorized into three types based on their source:

1. Metadata: metadata can further be classified into two types, object and technical metadata. Object metadata are data extracted from the web object such as the object size and creation date, whereas technical metadata has information about the type of web server and operating system used. Mining this data would improve authenticity and provenance of the archived content.

2. Usage data: a major source of usage data is web server logs. If a user sends a request to a web server regarding access to archived content, it may store the information like the date of request, number of bytes transmitted to that user and where it came from, the IP address of the user etc. This information helps the web mining process to analyze ratings of the archived content, traffic rankings etc.

3. Infrastructure data: this contains the backbone information of archiving infrastructure systems such as IP addresses, autonomous system numbers, number of CPU cycles connected with routers, bandwidth used by routers, etc. Through a data mining process, it is possible to analyze and measure the traffic updates of routers, Internet traffic, etc, and thus calculate the time taken for archived data packets to reach the destination host.

In a holistic view, all this knowledge would improve the performance of the web archiving process, and provide better categorization of contents and future prediction of usage patterns of archived data.

Challenges in infrastructure management

Cloud Computing and Grid Computing

A web archiving infrastructure should be reliably strong in storing large amounts of data and must be scalable to deliver these data to users while they are extracting it. Web archiving systems are intended to maintain their data for long-term preservation and so must be capable of retaining their archived content and other related data in case of system failure or shutdown.

Brewster Kahle (2002), the founder of the Internet Archive, has defined some basic requirements and operations for archival systems, which are still relevant:

– The system should use only commodity equipment.
– The system should not rely on commercial software.
– The system should not require a PhD degree to implement or to maintain.
– The system should be as simple as possible.

A web archiving infrastructure process flow involves storage, import, search, index, and access stages. The infrastructure must be compatible with these process requirements and it must also provide data security and resource management solutions. In addition, it must manage the IT resources and offer low cost solutions in terms of software,

hardware and other operational requirements. The major areas of concern for Internet archiving are interoperability, scalability and security (JAFFE & KIRKPATRICK, 2009).

To meet all these requirements web archiving institutions employ different types of computing systems based on their budgets. At present, cluster or distributed computing systems and grid computing systems are widely used among archiving institutions.

Grid computing provides a framework for identifying idle servers and other IT assets, and computes processor requirements for the archiving process. This in turn increases the efficiency of resource usage and creates balanced resource utilization. During an unpredictable peak, a particular task can be routed to an idle processor in the grid system and if the grid resource is fully utilized, it will suspend temporarily or cancel the lower priority task and set the way for archiving tasks with higher importance like capturing the index page of a website. Lower priority task subject will be carried out later according to the grid system resource specifications. It also calculates the unused storage disk spaces that are configured with the grid architecture (specifically called the data grid) and it creates a larger virtual data store able to archive contents more efficiently through a common remote machine as a shared storage resource.

Grid computing has been essential for the Web Archiving Singapore (WAS) Project, it facilitates sharing of storage space and other resources, like some specialized devices, software, services, licenses and so on. For instance, in an archival process, if a user needs to increase the speed of the internet to capture a snapshot of a multi-page website, the process can be split virtually in the grid architecture. The nature of the grid computing system allows parallel processing, an advantage for web archiving initiatives in terms of enhancing the IT infrastructure with managing resources, reliability and manageability. It also reduces cost and time, it needs less maintenance and manpower.

A grid computing system is compatible with internet archiving, but to some extent it faces issues in performing multi-level scheduling (FOSTER, ZHAO & SHIYONG LU, 2008). For instance, if one job is ready for archiving and it needs the capability of 100 processors each with available time of 60 minutes, it would have to wait until the location resource manager (LRM) can allocate 100 processors with the stipulated time availability. In the case of a cloud computing model, it looks entirely different. The particular task can be shared by all users at the same time processed by the dedicated scheduled queuing system and it allows applications to operate natively in the centralized storage cloud systems. The major drawback of grid computing is system failure; if one processor in the grid fails, the whole distributed network will be affected and data staging and scheduling will be stopped at a single point.

The prominent position of Cloud Computing

To overcome troubles that grid computing might encounter in future, the cloud computing system is a workable solution. Although grid computing is the backbone of cloud computing, it is designed to act as one complete infrastructure (infrastructure as a service-IaaS) and is automatically interchangeable so that the system would not halt when one network has failed to perform. Cloud systems provide the highest data security and compatibility in terms of storing data. In grid computing, one large task is divided into several sub tasks and executed on multiple processors based on the resource and time available, while in cloud computing architecture, users can harvest websites without interrogating the resources and perform multi-scheduled crawling. It can access shared servers on demand of resources, software and data.

A cloud computing service does not need to own software, application platforms or any individual infrastructure. It significantly reduces the upfront costs, operating and capital expenses.A cloud computing operation avoids server and network maintenance and as it allows access to multiple servers from anyplace through a common emulator, it is an eco-friendly method of computing. Most important of all, in cloud computing data can be restored. It has disaster recovery and data backup technology that would help to retain archived data at a time of system failure, like a server or data crash.

An example of cloud computing in Web Archival is the Latin American Government Documents Archives (LAGDA) initiative, which has implemented a cloud-computing system to archive ministerial and presidential web documents from eighteen Latin American and Caribbean countries (LAWAP[2], 2005). This is a joint project with the University of Texas Libraries, and the Texas Advanced Cloud Computing Center facilitates the cloud computing services. The current size of the LAGDA web archive is around 6 TB (OWENS, 2012).

Recommendations

Our recommendations are divided into three areas:

Ethical-Political:

a) The National Library must adopt a transparent and open framework regarding web archiving, making available its policies and sharing knowledge about WAS experience.

b) NLB must establish an ethical framework or ethics board regarding WAS.

2. The Latin American Web Archiving Project (LAWAP) aims to collect and preserve web resources from Latin America. The Internet Archive provides services to LAWAP.

c) NLB in line with the goals of supporting a knowledge economy in Singapore must rethink itself as a research library for providing support to digital humanities research, where WAS can have an important role as an expert source of research.

Legal-Copyright:

Singapore's copyright law in regard to web
archiving requires amendments that create

a) legal certainty for Web Harvesting, Digital Long-term Preservation, Long-term Access, and Use and reuse of the information archived for research and study; and

b) mechanisms that facilitate copyright clearance for reuse and transformation of copyrighted works.

Technical:

a) For better collection development, WAS must develop tools to improve utilizing metadata access in single websites (review of sitemap access tools).

b) Tools implementation (eg.Taverna) or development of new tools for an imporved work flow process.

c) WAS should start web mining on its preserved logs to improve the quality of web harvesting.

d) WAS should focus on creating better tools that allow better visualization of existing metadata.

e) WAS should study of terminology evolution and look for inclusion of a linguistic layer in its web archive architecture.

f) In future, for selective archiving of certain events, WAS could test using the cloud infrastructure.

Bibliography

Beasley, S. & Kail, C. (2009). "a2o: Access to Archives from the National Archives of Singapore". *Journal of Web Librarianship*, 3(2), 149-155. doi:10.1080/19322900902896531.

Chellapandi, S.; Han, C. W. & Boon, T. C. (2010). "The National Library of Singapore experience: Harnessing technology to deliver content and broaden access". *Interlending & Document Supply*, 38(1), 40-48.

Chellapandi, S. & Lian San, S. (2009, november). "Web Archiving Programme at National Library Singapore". *CDNLAO Newsletter*. Tokyo, Japan. Retrieved from [www.ndl.go.jp/en/cdnlao/newsletter/066/664.html].

Chen, H.; Chung, W., Qin, J., Reid, E., Sageman, M. & Weimann, G. (2008). "Uncovering the dark Web: A case study of Jihad on the Web". J*ournal of the American Society for Information Science and Technology*, 59(8), 1347-1359. doi:10.1002/asi.20838.

Cloud Computing vs. Grid Computing "CCSK Guide". (n.d.). CCSK Guide. Retrieved April 5, 2013, from [http://ccskguide.org/cloud-computing-vs-grid-computing/].

Copyright Office USA. (2005). Report on Orphan Works.

Costa M. and M. J. Silva. "Towards information retrieval evaluation over web archives". In S. Geva, J. Kamps, C. Peters, T. S. A. Trotman, and E. Voorhees, editors. Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, Boston, Massachusetts, July 2009. IR Publications, Amsterdam.

Datta, R.; Joshi, D., Li, J. & Wang, J. Z. (2008). "Image Retrieval: Ideas, influences, and trends of the new age". *ACM Computing Surveys*, 40(2), 1-60.

Difference between cloud computing and grid computing (n.d.). "Cloudways Managed Cloud Hosting & App Deployment Solutions". Retrieved April 3, 2013, from [/www.cloudways.com/blog/cloud-computing-vs-grid-computing-differentiated/].

Dougherty, M.; Meyer, E., Madsen, C., Van den Heuvel, C., Thomas, A. & Wyatt, S. (2010). "Researcher engagement with web archives: State of the art. Joint Information Systems Committee Report". Retrieved from [http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997].

Foo, S.; Wu, H. J., Lim, S. K., Haliza, J., Noryati, A. S., Mohamed, S. & Tim, Y. F. (2005). "Legal Deposit Development in Singapore: Future Challenges and Issues". In International Conference on National Libraries in the Knowledge Based Society. Bangkok.

Glanville, L. (2010). "Web archiving: ethical and legal issues affecting programmes in Australia and the Netherlands". *The Australian Library Journal*, 59(3), 128.

Hayles, K. (2012). *How we think: digital media and contemporary technogenesis*. Chicago & London: The University of Chicago Press.

Jaffe. E. & Kirkpatrick. S. (2009). *Architecture of the Internet Archive*. The Hebrew University of Jerusalem.

LiWA. (2013). "Livign Web archive web page". Web page retrieved 3 April, 2013, from [http://liwa-project.eu/].

Lor, P. J. & Britz, J. J. (2012). "An ethical perspective on political-economic issues in the long-term preservation of digital heritage". *Journal of the American Society for Information Science and Technology*, 63(11), 2153-2164, doi:10.1002/asi.22725.

Masans, J. (2006). "Access and Finding Aids". Web Archiving (pp. 131-151). New York: Springer-Verlag Berlin Heidelberg.

Mason, I. (2007). "Virtual preservation: How has digital culture influenced our ideas about permanence? Changing practice in a national legal deposit library". *Library Trends*, 56(1), 198-215.

National Digital Information Infrastructure and Preservation Program (U.S.), Joint Information Systems Committee, & OAK Law Project. (2008). *International*

study on the impact of copyright law on digital preservation. Washington, D.C.: Library of Congess. Retrieved from http://www.digitalpreservation.gov/partners/resources/pubs/wipo_digital_preservation_final_report2008.pdf

National Library Board. (2005). *Library 2010: libraries for life, knowledge for success*. National Library Board.

National Library Board. (2012). Singapore Country Report for the CDNL-AO Meeting 2012.

NLB Singapore (2010). "Singapore Country Report". In 15th CONSAL Executive Board. Bandung, West Java, Indonesia.

ORTMANN, S. (2009). "Singapore: The Politics of Inventing National Identity". *Journal of Current Southeast Asian Affairs*, 28(4), 23-46.

OWENS, T. (2012, June 6). "Web Archiving and Mainstreaming Special Collections: The Case of the Latin American Government Documents Archive | The Signal: Digital Preservation". Library of Congress Blogs. Retrieved March 23, 2013, from [http://blogs.loc.gov/digitalpreservation/2012/06/web-archiving-and-mainstreaming-special-collections-the-case-of-the-latin-american-government-documents-archive/].

PATEL, K. (2007). "Authors v. Internet Archives: The Copyright Infringement Battle over WEB Pages". *J. Pat. & Trademark Off. Soc'y*, 89, 410.

PAYNTER, G. W. & MASON, I. B. (2006). "Building a Web Curator Tool for The National Library of New Zealand". In New Zealand Library Association Conference.

Portrait Workshop. (2007, September 26). "Notification of Website Archiving - Portrait Workshop. Caricature Singapore". Retrieved from [www.caricature.com.sg/2007/09/notification-of-website-archiving.html].

POTTER, A. (2012) "A Vision of the Role and Future of Web Archives: Research Use | The Signal: Digital Preservation" (n.d.). Library of Congress Blogs. Retrieved March 20, 2013, from http://blogs.loc.gov/digitalpreservation/2012/05/a-vision-of-the-role-and-future-of-web-archives-research-use/

SEET, K. K. (2005). *Singapore's transformative library: knowledge, imagination, possibility*. Singapore: SNP Editions.

SENG, D. (2008). "International Study on the Impact of Copyright Law on Digital Preservation. Singapore's Legal Position". Presented at the WIPO International Workshop on Digital Preservation and Copyright, Geneve, Switzerland.

SG Wiki Editor. (2007, July 4). Discussion about Pacific Centennial Group Wikipedia Page. Retrieved from http://www.zaped.info/Wikipedia:Articles_for_deletion/Log/2007_July_2

SPANIOL, M.; MAZEIKA, A., DENEV, D. & WEIKUM, G. (2009). "'Catch me if you can': Visual Analysis of Coherence Defects in Web Archiving". In Proceedings of the 9th International Web Archiving Workshop (IWAW 2009), Corfu, Greece. Retrieved from [www.iwaw.net/09/IWAW2009.pdf].

Stirling, P.; Illien, G., Sanz, P. & Sepetjan, S. (2012). "The state of e-legal deposit in France: Looking back at five years of putting new legislation into practice and envisioning the future". *IFLA Journa*l, 38(1), 5-24. doi:10.1177/0340035211435323.

Sundaram, K. (n. d.). "Cloud Computing vs. Grid Computing - What is the Difference?". Find Science & Technology Articles, Education Lesson Plans, Tech Tips, Computer Hardware & Software Reviews, News and More at Bright Hub. Retrieved March 29, 2013, from [www.brighthub.com/environment/green-computing/articles/68785.aspx].

was, nlb. (2013a). faq. Retrieved from [http://was.nl.sg/faq.html].

was, nlb. (2013b). Takedown Policy. Retrieved from [www.microsite.nl.sg/TakedownPolicy.html].

Wu, P. H. J. & Heok, A. K. H. (2005). "Documenting Online Collaboration Between Researchers and Information Professionals: The Case of the Asian Tsunami Web Archive". *Singapore Journal of Library & Information Management*, 34, 75–85.

Yong Fu, J. L. (2013). "One Versus The Mob: A Case for the Wisdom of Crowds in Archival Appraisal 2.0". Thesis. Nanyang Technological University, Singapore.