IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 24.06.2018

Web archiving issues and challenges in State Government of Sarawak (Malaysia): Do they really need their website to be archived?

Jassalini Jamain
Sarawakiana Division, Sarawak State Library, Kuching, Malaysia
jassalij@sarawak.gov.my

Ayu Lestari Yahya
Archives Management Division, Sarawak State Library, Kuching, Malaysia
ayulestari@sarawak.gov.my

Natalia Muhammad
Sarawakiana Division, Sarawak State Library, Kuching, Malaysia
nataliam@sarawak.gov.my

Musa Ayob Abdul Rahman
Website and Web Application Development Division, Sarawak State Library, Kuching, Malaysia
musaaar@sarawak.gov.my

Abstract:

Sarawak State Web Archive (SSWA) is Sarawak State Library's (Pustaka) initiative. Website contents of the Sarawak State Civil Service (SSCS) entities obtained from World Wide Web (WWW), are archived for the purpose of preserving non-library resources, as part of the Legal Deposit requirements of Sarawak State Library Ordinance, 1999. Web preservation is considered as a common practice at international level, whereas in Malaysia this is still at a minimal level. Since 2009, Pustaka has been harvesting 132 websites of Sarawak State Government departments and agencies. However, Pustaka faced challenges in performing web archiving works. This paper focuses on the general issues and challenges in preserving corporate information heritage to make it available for future reference.

Keywords: Web archiving; legal deposit; corporate information heritage; Sarawak, Malaysia

Introduction

In general, web archiving is the processes of gathering data, which is uploaded on the World Wide Web, storing it, ensuring the data is preserved, archived, and making them available for research purposes.

Library of Congress defined Web Archiving as "The process of creating an archival copy of a website. An archive site is a snapshot of how the original site looked at a particular point in time". In recent years, various studies were undertaken pertaining to the advent of methods and procedures in web archiving [1]. For example, In 2006, Masanès comprehensively reviewed the state-of-the-art methods, tools, and standards to build a web archive [5].

Pustaka's Sarawak State Web Archive (SSWA) focuses on selecting, scoping and preserving website contents of Sarawak State Civil Service (SSCS) entities derived from the World Wide Web (WWW), to fulfill the legal deposit requirements of *Sarawak State Library Ordinance, 1999,* as well as for the preservation of non-print library resources.

Literature Review

Government agencies' websites play a vital role in disseminating accurate information to the public. The contents on the websites may change daily, and some may be permanently removed. This can lead to the loss of important information, which might be valuable for future research purposes. Niu (2012) defined "Web archiving is the process of gathering data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research"[2].

According to The National Archive, United Kingdom, 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 disappearance of important information for the future. Lala and Joe (2006) in [3] mentioned that "in the twenty first century information is being produced in vast rates, not only through traditional forms of formats, but also increasingly in electronic, or digital formats".

Web archiving is one of the initiatives in preserving the information and capturing the data on the websites. However, there are still few who do not recognize the value and the importance in preserving the websites. The National Archives, UK (2011) stated "web archiving is a vital process to ensure that people and organization can access and re-use knowledge in the long-term, and comply with the needs of retrieving their information" [5].

Web archiving implementation is not a simple task since it involves technology and big data. There are many aspects to be measured such as the storage format as well the storage capacity, selection of the website that need to be archived and a host of other considerations. Pennock (2013) highlighted "Capturing large and complex sites on a recurring basis, whilst maintaining and clearly identify the relationship between different versions of a site and simultaneously managing the artificial boundaries that inevitably occur in an 'extracted' collection, requires a more complicated solution" [4].

Objectives of Sarawak State Web Archive

The main objectives of Sarawak State Web Archive are as follows:

a) To preserve and maintain evidence of the web contents published by departments and agencies of the Sarawak State Civil Service (SSCS) .
b) To comply with the requirements of the *Sarawak State Library Ordinance, 1999.*
c) To contribute to information availability and accessibility for research

Background

Pustaka's web archiving processes include capturing of websites published by entities in the SSCS, comprising of ministries, local authorities, state statutory bodies and government-owned companies and resident offices. The capturing of websites are inclusive of still images, sound recordings, motion pictures and other multimedia formats available on the websites. The processes of capturing are undertaken once in every two months.

"Continuous Improvement Program" (CIP) on web archiving

A "Continuous Improvement Program" (CIP) workshop was held on February 2017, to identify areas of improvement to the web archiving processes and outcomes. The CIP committee identified the challenges in performing the web archiving services. One of the major challenges is the non-availabilty of maintenance of the current solution used for web archiving. As such, the functions of the web archiving solution is not updated to meet the current requirements of Pustaka.

Web Archiving Process

a) Capturing and selection of Sarawak Government Websites

As of May 2018, Pustaka has harvested and archived a total of 132 websites. However, after several state government departments underwent restructuring and there were several departmental mergers in 2016, the total websites are now reduced to only 92 websites. Because of that reason, archiving the websites has become very important especially to the website owners in order to keep all the information of web content that are feared to be lost if they are not properly kept and stored.

Table 1: Statistics of web archiving for Sarawak State Civil Service for 2017

No.	Category	Total Websites
1	Ministries	6
2	Chief Ministry Departments	6
3	State Departments	21
4	Local Authorities	25
5	Residents & Districts Office	11
6	Statutory Bodies	23
	TOTAL	**92**

Source: Sarawak State Library Annual Report 2017

Issues in web archiving

a) Authorization to capture Websites

Pustaka was given the mandate and authority to archive all SSCS websites through an executive directive issued by Sarawak State Secretary. All the website owners consented with the instructions by allowing their website contents to be captured. However, some government websites still require permission in order to completely harvest data from their websites.

b) Legal Requirement

Pustaka has full mandate, under *Section 14 (1), of Sarawak State Library Ordinance 1999*, in preserving digital content or information published in Sarawak. As stated in the *Ordinance,* Pustaka is "to provide for the preservation and the use of library resources or materials published in Sarawak."[11]. The processes of selection, scope and preservation of websites content for the purpose of preservation of non-print library resources fulfils the Legal Deposit requirements of the *Ordinance*.

c) Preservation

The preservation of websites becomes so important nowadays. By capturing their websites, Pustaka preserves the organizations' vital information, which might change and may not necessarily need to be available on their current websites.

d) Dedicated Webmaster

An observation on government websites was conducted in 2016, and the findings indicated that some websites were not updated due to lack of manpower or dedicated staff to maintain the websites. Some of those websites were not updated since they were first published. It was recommended to these agencies on the need to assign webmasters to maintain the websites to ensure currency of their agencies' information.

e) Lifetime value of IT equipment

By early 2018, most of the existing hardware for web archive in Pustaka are no longer able to support the principle. Faulty hardware needs to be replaced with refurbishment parts, and availability of parts will need a minimum of 8 weeks to be delivered. The pricing for back-to-back support also will cost more than buying a new server.

Challenges

Web archiving is regarded as a means to preserve big data as the population is increasingly active online. Based on the statistics in 2017 by Malaysian Communications and Multimedia Commission (MCMC), from 32 million people in Malaysia, 24.5 million users (76.9%) have access to Internet, and the other 7.5 million (23.1%) do not access to the internet. Almost all Malaysian Internet users (96.3%, 23.59 million) use Internet for text communication such as WhatsApp, Facebook Messenger, WeChat and so forth [14]. This statistics shows that people prefer to go online to make their lives became easier. Therefore, digital information is important and need to be stored as a proof of law and as evidence to evolving culture and current situation. Due to this situation most of countries in the world have realized on the important of web archive in preserving the nation's intellectual heritage for future use.

There are three main challenges in web archiving works in Pustaka:

a) Insufficient data harvested

Though web archiving work was done on a quarterly basis and each website were successfully harvested, the majority of these websites cannot be viewed because of insufficient data were harvested. Heritrix status code (9998-robots.txt rules precluded fetch) can be seen in the web's archive log viewer file indicating that these websites had been harvested but, not enough materials to be viewed e.g., only 6.52 kilobytes. This is because some agencies restrict the public from harvesting the content of their websites and even though the Pustaka has obtained the permission from agencies involved. The web archiving system is unable to capture the full details of the websites. These are illustrated in Table 2 to Table 4 below.

Table 2: Data downloaded for the web archiving process

Results

Id	Name	Harvest Date	State	Owner	Run Time	Data Downloaded
8880251	Majlis Perbandaran Samarahan	07/06/2017 12:16:14	Harvested	I. one	00:05:29:27	257.57 MB
8618390	Majlis Perbandaran Samarahan	22/04/2017 12:34:48	Harvested	I. one	00:08:27:46	243.75 MB
8618123	Majlis Perbandaran Samarahan	02/03/2017 12:06:08	Harvested	I. one	00:07:12:35	248.5 MB
8519812	Majlis Perbandaran Samarahan	25/01/2017 17:29:13	Harvested	I. one	00:04:31:55	323.13 MB
8421509	Majlis Perbandaran Samarahan	13/12/2016 16:01:00	Harvested	I. one	00:04:11:56	353.47 MB
8356006	Majlis Perbandaran Samarahan	19/11/2016 14:11:41	Harvested	I. one	00:01:33:58	46.87 MB
8159458	Pejabat Daerah Samarahan	10/08/2016 07:23:45	Harvested	I. one	00:00:00:01	6.52 KB
8159396	Pejabat Residen Samarahan	02/08/2016 11:31:32	Harvested	I. one	00:00:00:32	6.55 KB

Source: Sarawak State Library Web Archive

Table 3: Insufficient web archiving data harvested

Results

Id	Name	Harvest Date	State	Owner	Run Time	Data Downloaded
8880310	Pejabat Daerah Simunjan	07/06/2017 16:19:08	Harvested	I. one	00:00:00:04	6.49 KB
8880308	Pejabat Daerah Serian	07/06/2017 16:17:55	Harvested	I. one	00:00:00:14	6.49 KB
8880306	Pejabat Daerah Sarikei	07/06/2017 16:16:42	Harvested	I. one	00:00:00:04	6.48 KB
8880304	Pejabat Daerah Saratok	07/06/2017 16:15:34	Harvested	I. one	00:00:00:02	6.49 KB
8880300	Pejabat Daerah Pakan	07/06/2017 16:06:54	Harvested	I. one	00:00:00:01	6.48 KB
8880295	Pejabat Daerah Mukah	07/06/2017 15:02:30	Harvested	I. one	00:00:00:03	6.49 KB
8880292	Pejabat Daerah Meradong	07/06/2017 15:01:29	Harvested	I. one	00:00:00:06	6.5 KB
8880290	Pejabat Daerah Matu	07/06/2017 15:00:06	Harvested	I. one	00:00:00:03	6.49 KB

Source: Sarawak State Web Archive

Table 4: Heritrix status code (9998-robots.txt rules precluded fetch) can be seen in the web's archive log viewer file

b) Technology

The rapid development of the latest technology requires urgency in actions to be taken in archiving the websites. Electronic content is constantly changed, erased, or lost permanently and cannot be recovered when the website is being updated. In such instances, web content becomes inaccessible.

The ever changing digital storage technologies has affected how the Pustaka copes in terms of migrating from existing technology to newer ones. Pustaka is in the process of enhancing the existing storage format to make it in line with current technologies.

c) Data Storage & Backup Recovery

One of the crucial problems is on the storage capacity in keeping the archived websites. A server with a huge space capacity must be provided to ensure the web archiving activity runs smoothly.

Another issue is the inconsistency of backup recovery work done to the web archiving server. Previously, back up recovery for web archiving was done using CD, but now the harvested websites are saved directly to the server.

Due to the frequency of harvesting done in every two months, it is required that the largest storage capacity is to be prepared for data keeping from whole websites for a longer time period.

Pustaka Web Archives milestone

Table 5 shows the evolution of web archiving practices at Pustaka from 2009 until 2018 and future planning from 2019 and beyond. The 8 years of web archiving was a learning process for Pustaka in web archiving. There is no public access for the archived websites at this point of time until implementation of new web archiving solution.

Table 5: Evolution of web archiving at Sarawak State Library

Period	2009 – 2010	2011-2016	2017 -2018	2019 and beyond
Software	HTTrack	Web Curator Tool	NetArchive Suite5 (Trial process)	Enhancement process
Frequency of crawling	Bimonthly	Quarterly	Bimonthly	Bimonthly
Number of websites harvested	138 websites	110 websites	92 websites	92 websites
Storage system	In house server	In house server	Trial process at Central Data Centre	To be hosted at Central Data Centre
Backup recovery	Save into CD	Save directly to Server	Trial process at Central data centre	Central Data Centre
Accessibility	No public access	No public access	No public access	Open public access

Source: Legal Deposit Unit, Sarawak State Library

Enhancement of New Project Web Archiving Solution

Starting 2017, the development of web archiving system is currently developed by a government-owned IT company, Sarawak Information System (SAINS). This project is still in progress and awaiting the final approval from a higher authority.

This enhancement project will be using *NetArchive Suite 5* which is highly recommended by the web vendor for web archiving process with criteria as follows:

a) NetArchive Suite 5 is the latest software which is user-friendly in usage, easy to administer, configure and customize.
b) Although this software does include a web viewer to view the results, however, it has a basic verification viewing and separate installation for web viewer.
c) Operating system is available for Linux.
d) Up-to-datedness of software (latest updated on 10 March 2017)
e) Maintenance requirements is very minimal from user side as operations are highly automated.
f) High accuracy of harvesting with a success rate of 98%.
g) Time taken to crawl the websites range from the fastest of crawling in 3 minutes and 15 seconds (8MB), to the slowest time in 6 hours 44 minutes and 20 seconds (6.5GB).
h) This software allows different generations of download/overwrite existing (e.g v1.0, v2.0).
i) Fully customizable depth Level via the versatile Heritrix crawler's configuration to Heritrix instances (Access into Deeper Level (from 3 to 5 layer deep))

This project is still at the trial process and will be hosted at a Central Data Centre as soon as the project is approved.

Purpose of the Project

The main purposes of the new web archiving project are as follows:

a) To implement a web archiving solution that allows Pustaka to archive Sarawak State Civil Service (SSCS) websites with minimal error.
b) To archive websites to be retrieved and viewed.
c) To setup a support and maintenance process to ensure that the web archive solution implemented will be upgraded and functionalities enhanced when necessary.

As this project continues and matures, there are several potential improvements and new possibilities that should be considered of which is the enhancement of accessibility and visibility of those websites that are already archived.

This project can be strengthened by collaborating with other collecting institutions such as The National Archive of Malaysia who maintains the Federal Malaysia Government websites. Active collaboration with state government agencies who own these websites will benefit all parties.

Conclusion

As a conclusion, web archiving is important to preserve and maintain evidences of the Sarawak State Civil Service (SSCS) and made accessible for future research purposes. For future development of web archiving, the processes must be supported by concrete policies and framework. Websites are used daily by the citizens for access to information However, the frequent changes to this information requires web archiving, not only for reference and research in the future but also as rich text of records as evidence of the evolution of the SSCS agencies. It necessitates for a web archiving solution which works optimally, and that the system is to be centrally hosted.

Acknowledgement

We would hereby like to express our profound and gratitude to Sarawak State Library's Top Management, the Chairman, YBhg. Tan Sri (Dr) Hj Hamid Bugo and Board Member of the Sarawak State Library, Chief Executive Officer, Puan Hajah Rashidah Haji Bolhassan, Deputy Chief Executive Officer, Puan Arpah Adenan and all Sector Heads who have given full support an opportunity for staff participating in IFLA WLIC 2018.

Equally thankful to Puan Hayati Haji Sabil, Continous Improvement Program Team (CIP) and the Sarawak Information System (SAINS) by providing information and assistance pertaining to the State Web Archive System at Pustaka.

Thanks are due to all Sarawak State Government Department/Agencies who have been involved in giving us the permission to capture their websites.

We also would like to extend our gratitude to our mentor, Mr. Edison Ricket for his guidance and other colleagues at Pustaka especially to the staff at Depository Services Sector, Pustaka's IFLA WLIC 2018 Committees, Mr. Bronny Lawrence Nawe, Puan Dayangku Horiah Awang Gani, Miss Norasemah Drahman and ICT Team. Not forgetting, Mr. Chuah Kee Man the proofreader, and for those who are really involved directly and indirectly in completing this paper.

Finally, we are indebted to our family for their support, love and encouragement.

REFERENCES

[1] Library of Congress. Saving the World Wide Web. Available at: http://www.digitalpreservation.gov/series/challenge/web_harvest_challenge.html. Access on 15 May 2018.

[2] Niu, Jinfang. (2012). An Overview of Web Archiving. D-Lib Magazine, Volume 18, Number 3/4.

[3] Lala, Vanita & Joe, Susanna. (2006). Web Archiving At The National Library of New Zealand.

[4] Pennock, Maureen. (2013). Web Archiving. Great Britain: Digital Preservation Coalition.

[5] The National Archives. (2011). Web Archiving Guidance. United Kingdom: The National Archives.

[6] Masanes, Julien. (2005). *LIBRARY TRENDS,* Vol. 54, No. 1.

[7] Grotke, Abigail. (2017). Collaborating to Preserve Federal Government Websites. *INFORMATION OUTLOOK*, Volume 21, No. 3, 5-7.

[8] Kleiber, Eleanor. (2014). Gathering the 'Net: Efforts and Challenges in Archiving Pacific Websites. *The Contemporary Pacific*, Volume 26, Number 1, 158-166.

[9] Ahmed AlSum. (2014). Web Archives Services Framework For Tighter Integration Between The Past And Present Web. *Ann Arbor, Michigan: ProQuest LLC.*

[10] Hockx-u, Helen. (2011). Web Archiving at the British Library. Available at http://www.loc.gov/today/cyberlc/feature_wdesc.php?rec=5272. Accessed on 15 May 2018.

[11] Sarawak State Library (Amendment) Ordinance, 1999, Chapter 29 Section 14 (2010).

[12] FOR A.tv. The Wayback Machine: Preserving the History of Web Pages. YouTube Channel 23 December 2011. Available at: https://www.youtube.com/watch?v=JsL1TADosN0&feature=youtu.be. Access on 25 May 2018.

[13] Library of Congress. Web Archiving. YouTube Channel 30 November 2009. Available at: https://www.youtube.com/watch?v=T0943YkhLWU. Access on 16 May 2018.

[14] Malaysian Communications and Multimedia Commission. (2018) Internet Usage Statistics in Malaysia for 2017 (online). Available at: http://iamk.com.my/articles/2018/01/12/internet-usage-statistics-in-malaysia-for-2017/. Access on 28 May 2018.

[15] Sarawak Information Systems Sdn. Bhd. (2017). Project Proposal for Pustaka Negeri Sarawak: Web Archiving Solution. November 2017, version 2.0.