E-B
Environment - Behaviour
Proceedings Journal

ITMAHuB2023

https://kppim.uitm.edu.my/research/conferences/itmahub-2023
2ⁿᵈ International Conference on Information Science,
Technology, Management, Humanities, and Business
Hotel UiTM, Shah Alam, Malaysia, 26-27 Sep. 2023
Organised by College of Computing, Informatics and Mathematics, UiTM, Malaysia

e-IPH
e-International
Publishing House Ltd.,
United Kingdom

Exploration of Web Contents of Selangor Royal Family using Web Archiving Tools

Masturah Mohd Nor, Farrah Diana Saiful Bahry, Shamila Mohamed Shuhidan, Haslinda Husaini

College of Computing, Informatics and Mathematics,
Universiti Teknologi Mara, Malaysia, Kampus Puncak Perdana, Shah Alam, 40150, Malaysia

farrahdiana@uitm.edu.my, noraizan@uitm.edu.my, shamila@uitm.edu.my, haslinda3145@uitm.edu.my
Tel: 0192084038

Abstract
This study aims to explore the web contents of Selangor Royal family as an initiative to preserve the contents of social and cultural values. Initially, an exploration of websites' profiles and the contents of Selangor Royal family by using tools BuildWith and XML-Sitemaps. Then, the selected web pages were then archived using web archiving tools, HTTrack and Conifer. The findings indicate that there are different types of websites and website owners covering the Selangor Royal family-related content. The recorded information and lifestyles of royal family represent value in building Malaysian identity and are significant to be preserved for future discovery.

Keywords: web contents exploration; web archiving; web preservation; Selangor royal family

1.0 Introduction

The emergence of web archiving is closely related to the growing creation of online web content, which exists in two prominent forms of: text and multimedia. This is due to the need to preserve the published data or information that possesses value as the web contents are prone to changes, updates, modifications, and even vanish from the internet. Krol (2021) stated that all content published on the websites may be lost irretrievably through each upgrade or replacement of a website with a new one. The rapid growth of technology has led to the born-digital heritage over the internet, hence the need for web archiving. This situation has resulted in the existence of pioneers in web preservation initiatives where, some of which are the Internet Archive (based in the United States but with international coverage), the Australian National Library project 'PANDORA' (Preserving and Accessing Networked Documentary Resources of Australia) and many more (Rockembach, 2018). Information on social and cultural values is among the web content topics highlighted in previous research that need to be preserved.

In Malaysia, royal families are recognized as having a strong role in reflecting good motivation and vibes to citizens and society based on their high reputation. Each state in Malaysia has a Sultan who governs the state as a constitutional advisor and the Islamic religion in each state (Shah, 2022; Muslim et al., 2020). Nowadays, royal families' activities in their work and personal lives have been revealed and exposed throughout social media and online channels. Some of the web content have been publicized either by local or international web channels. Various online content types relate to the Malaysia Royal families, especially the Selangor Royal family,

such as their royal genealogy or history, professional growth, achievements, social activities, and commitment to the nation. All of the contents are scattered on the World Wide Web and become references to the society building Malaysian identity and nation-state. Some web content may be useful in the future and become royal memories and historical evidence. Archiving such royal memories and historical evidence's web contents is crucial and must be neglected and addressed. However, the web archiving initiatives in Malaysia are still limited, leaving the royal historical heritage collections at risk of losing and inaccessible in the future. Besides, given Malaysian practices, web archiving initiatives are still considered at the immature point. The initiative is negligence due to issues like massive storage use, privacy act violations, and Malaysian citizens' awareness of the value of digital content (Bahry et al., 2019).

Hence, our research aim is to explore the web contents of the Selangor Royal family as an initiative to preserve the contents of social and cultural values available over the internet using a web archiving approach. This study highlights the nature of information in the web format that are prone to updates, changes, modification, and removal which put its originality, values, and lifespan at risk. Consequently, it provides significant practical contributions to the web archiving initiative and preservation in Malaysia, which is very limited and lacks awareness. The data analysis from web exploration and web archiving tasks presents the essential aspects and characteristics of web contents that need to be recorded when preserving information in this form which is different from other formats of information. The technical aspects of the web content greatly influence the web archiving performance. Future research or web archiving implementation may improve the practices and approaches for more efficient processes. Next, this study also aims to provide social contributions by explicitly focusing the preservation initiative on the Selangor Royal family. As the royal family is recognized in Malaysia as setting a good example among citizens and encouraging excellent nation-building while maintaining our identity, thus a need to preserve their reported achievements, participation, contributions, and others scattered all over the internet.

2.0 Literature Review

2.1 Web Archiving
Web archiving is a process that involves gathering, storing, and providing information on past events from the World Wide Web, commonly known as the Web, for future researchers (Rockembach, 2018). The International Internet Preservation Consortium (IIPC) interprets web archiving as the process of (1) collecting web parts or pieces, (2) preserving the collected parts in an archival format, and eventually (3) making the collections available for access and use. These definitions give insight into the idea that web archiving is associated with a series of processes primarily involving web collections to preserve, access, and/or utilise the said collections. It is essential to understand that rapid technological changes are one of the factors that urge people or institutional bodies and researchers to archive web content, whereas at the same time, technology advancement also provides the methods and tools to implement and improve web archiving activities. Web archiving implementation is more familiar and widely known as the activities carried out by memory institutions, including "Libraries, Archives, Museums, and others whose collective mission is to manage the documentary and cultural heritage of their nations" (IGI Global, 2022).

2.2 Royal Cultural Heritage
UNESCO defined cultural heritage as the inheritance of physical artefacts and intangible attributes belonging to a group or society inherited from previous generations, preserved in the present, and granted to benefit future generations (Wijesundara & Sugimoto, 2018). According to Ibrahim et al. (2022), Malays are significant with the culture of being loyal to their king, or the formal term is Sultan. They perceive the Sultan as the ruler they must obey and be loyal to. Muhammad & Rosdi (2022) stated that the Sultan, also referred to as the ruler, has complete jurisdiction over a state with assistance from the royal family members. This portrays how notable the existence of the Sultan and his descendants are to Malaysia as a whole and to individual states. However, the study related to royal institutions is minimal, especially in the context of web content preservation and a study on the Selangor royal family. The previous research conducted on Malaysian royal institutions involved other different aspects, such as architecture and environment by Muhammad & Rosdi (2022), related explicitly to Kedah Royal. Another architecture-related study of the royal institution is conducted by Sodangi et al., (2020) for the Royal Museum of Seri Menanti located in Negeri Sembilan. Besides, a study by M. K et al., (2020) was conducted about heritage plants in the royal town of Perak. In contrast, Ibrahim et al., 2022 conducted a study related to royal traditional clothing and the culture of Malay ceremonies.

2.3 Web Archiving Flow Process/ Approach
In general, studies on web archiving involve a wide range of aspects. It is challenging to find research on web archives specific on preserving web collections of royal cultural heritage. The most found studies on web archiving centred around technical aspects in which the research involved a few numbers of technology tools or software. These can be referred to the study by Yu et al. (2009), Chu et al. (2010), Kawano (2011), Anjum (2012), Holzmann et al. (2016), and Bahry et al. (2019) that involved technology application as the research approach in which many of them used experimentation method. Previous studies also experimented with the best practices and technology tools to be used in any step involved in archiving the web collections. In addition, the institutional web archiving standard, focusing on data harvesting (e.g. Heritrix), data storage (WARC), and data playback to users (e.g. Wayback), provides a vital set of tools for web histories, but also not the only available method to preserve the web histories (Hegarty K., 2022).

Among the flow processes or approaches identified from previous research is the Web Archiving Life Cycle Model introduced by the Archive-It team in 2013 (Bragg & Hanna, 2013). This life cycle model is the product of a project from Archive-It team members' collaboration as their first white paper. The project was conducted to design web archiving best practices and processes to be shared

with organisations that are interested in implementing web archiving initiatives. It has a circular shape, which is used to indicate the repetitive nature of the steps in the life cycle. The elements that made up this model include policy, vision and objectives, resources and workflow, access/use/reuse, preservation, risk management, metadata/description, appraisal and selection, scoping, data capture, storage and organisation, and quality assurance and analysis. Khan and Rahman (2019) constructed a model called the Systematic Approach for the Web Preservation Process. Detailed steps are included in this approach, such as defining the scope of the web archive, understanding the web structure, identifying the web resources, identifying the designated community, and others.

3.0 Methodology
This research has employed part of the model developed by the Archive-It team in 2013 called the Web Archiving Life Cycle Model. The selected tasks were extracted from the inner orange circle of the model, which the original life cycle model comprises of (1) appraisal and selection, (2) scoping, (3) data capture, (4) storage and organisation, and (5) quality assurance and analysis (Archive-It, 2013). However, in conducting the website analysis activity, only the first three tasks were employed in this study. These tasks were categorized as analysing the website as the main activity. This research also considered previous studies' recommendations and suggestions to undertake this task. Besides, a few online tools have been utilized to conduct this website analysis, including BuiltWith, XML-Sitemaps, HTTrack, and Conifer. The crawler tools used are easy to use, especially by newbies in web archiving, and affordable with a free subscription-based. Figure 1 shows the study approach, which illustrates iterative individual tasks of website analysis that have been constructed based on the aforementioned life cycle model.
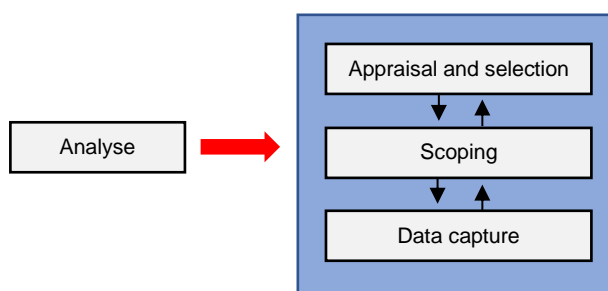

Fig. 1: Iterative Individual Tasks of Website Analysis

The selected tasks being implemented in this stage, as shown in Figure 1 are further elaborated as follows:
1) Appraisal and selection – This task involved selecting specific websites we want to archive and preserve. In the context of this study, we have chosen websites with information related to the Selangor royal family, which aligned with the aim of this study. The inclusion criteria used here were broad in terms of content related to "Royal Selangor", "Kesultanan Melayu Selangor", "Raja Selangor" and any name of Selangor royal genealogy. Next, the website types vary from news to corporate websites, blogs, or even social media. Another requirement is that all websites pass the pre-crawler process to estimate the crawler time and website size.
2) Scoping – After choosing the sites to archive, scoping is the step where website range or scope on the Selangor royal-related content to be archived has opted from portions of a website, whole sites or entire web domain as well.
3) Data capture – The stage where we have to deal with the suitable crawling software. HTTrack is one of the web archiving tools that enables the website to be archived. Hence, once the crawling process is complete, it can be verified by checking the HTTrack web archival repository with the actual website content.
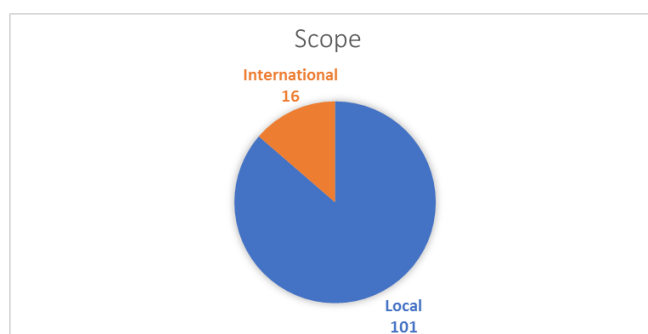
4.0 Results

4.1 Appraisal and Selection

Fig. 2: Scope of the Websites

Analysing websites is an essential initial step before archiving the website content. One hundred seventeen websites' contents related to Selangor royal family information or news have been selected to preserve the information regardless of its form. This step includes identifying and extracting the vital information of the selected websites, which comprise the Website Name, Website Type, Scope, Website Owner, and Topic of the web content. Besides, the websites' profiles were also identified during this phase to retrieve information such as Content Management System, Operating System & Server Operating System & Server, and SSL Certificates using an online tool called BuiltWith. Meanwhile, using an online tool called XML-Sitemaps, the processing time and web content size were identified as part of this step.

Figure 2 shows the graph of the website's scope being archived. A total of 117 selected web content related to the Selangor Royal Family's content, 101 of them are from local websites. In contrast, only 16 international websites were found to have covered news related to Selangor Royal. All the selected web contents belong to diverse websites' owners, representing either an organisation or individual, of which the majority are from all over Malaysia and few from international bodies.
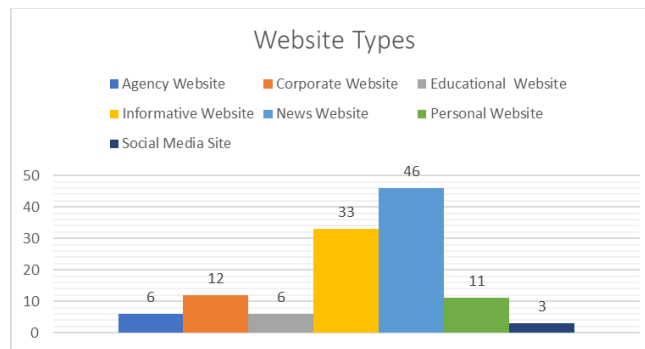

Fig. 3: Type of the Websites

The website types that have been identified related to Selangor Royal families comprise an agency website, corporate website, educational website, informative website, news website, personal website, and social media site. Figure 3 presents the number of websites based on their types. The type of website with the most content related to Selangor Royal families is a news website with a total of 46 web contents. This is followed by the informative website with 33 web contents, the corporate website with 12 web contents, and the personal website with 11 web contents. Meanwhile, educational websites and agency websites have the same number of web content, which is 6. The least number of website types goes to social media sites, with only three web contents.

4.2 Website Archive Implementation
The website archive implementation in this study was made up of the other two remaining tasks, which are scoping and data capture. These tasks were successfully accomplished by utilizing two web archiving tools, HTTrack and Conifer, as mentioned earlier. The decision to utilize two different archiving tools is to see the performance of the tools in archiving the web content. It is essential to understand that when it comes to technological tools, there must be advantages and disadvantages among them that we can identify and compare. HTTrack is a browser utility that is freely available online and crawls different types of files from a server to a local directory, such as HTML, images, and other files. It also allows offline viewing of the website. (Khan & Rahman, 2019)
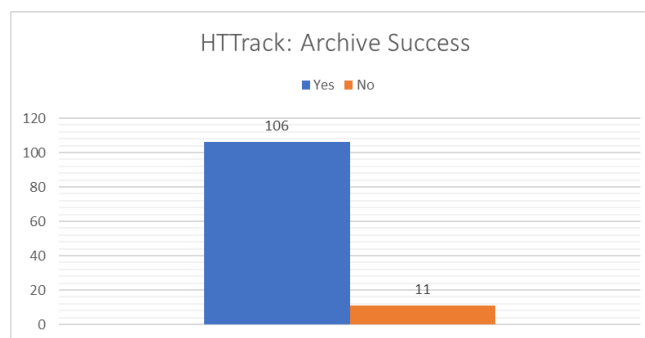

Fig. 4: HTTrack Archive Success

The status of the web archiving process on the 117 selected web pages related to the Selangor royal families is demonstrated in Figure 4. It indicates the number of successful and unsuccessful web pages archived using HTTrack. One hundred six web pages have successfully been archived, while another 11 were found to be unsuccessful in the archiving process.
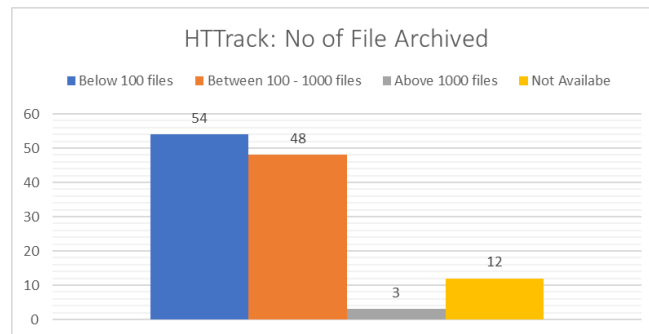
Fig. 5: HTTrack No of File Archived

On the other hand, Figure 5 presents the number of files archived for each selected web page. This is an important indicator because the number of files available on each web page or website depends on the website's structure and content. The outcome of the study analysis based on the archived files altogether has shown that 54 web pages have below 100 files, while 48 web pages have between 100 and 1000 files. These two categories are not huge gaps. Only three web pages have above 10000 files and the remaining falls under the unsuccessful attempt of archiving.
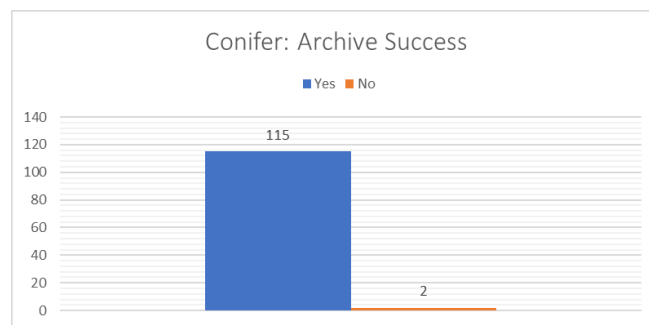

Fig. 6: Conifer: Archive Success

As mentioned earlier, Conifer is another web archiving tool that this study has used for archiving the web content. According to Jayanetti et al. (2022), Conifer also referred to as the web archive platform, which was formerly known as Webrecorder, which provides a service for web users to record and replay websites. Figure 6 shows the archive success status using Confier. It can be seen that 115 of the selected web pages have successfully been archived, and only two web pages have not been able to be archived. However, this does not affect the validity and reliability of the study since, technically, only some websites are expected to be archived.
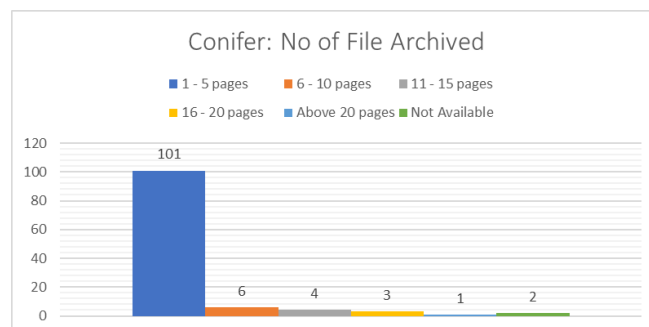

Fig. 7: Conifer: No of File Archived

In contrast with HTTrack, the web page archived by Conifer is in the form of a web page itself, which is stored on its cloud storage. Thus, this study presented the archived web page outcome based on the number of pages captured by Conifer, as shown in Figure 7. The range with the greatest number of web pages where their pages were successfully archived is between 1 – 5 pages of 101 websites. This is followed by a total of 6 websites with a range between 6 – 10 pages being archived, 4 websites between 11 -15 pages, three websites between 16 – 20 pages and the least number of websites with above 20 pages being archived is only one website. The remaining two websites are the ones with unsuccessful archive status.

5.0 Discussion

Website exploration and analysis are essential steps before archiving the web content itself to obtain as many details as possible regarding the websites that we have selected to archive. The scope of selected websites is dominated by locally owned websites, which is expected. We believe that the media coverage is still limited locally and even more internationally. Various topics cover the Selangor royal family on those selected websites, including background, participation, experiences, opinions, campaigns, sports and others. As a royal family that people look up to and are significantly known for representing the nation, especially Malay culture, there should be more information and media coverage on the Selangor royal family so that many important occasions and information are captured for future knowledge discovery as well as to preserve its value. Besides, it indicates the reputation of the Selangor royal family itself. Furthermore, a few websites covering the Selangor royal family were identified during website analysis. Among the website types, news websites appeared to be the media that covered the Selangor royal family story the most, with a percentage of 40 per cent.

The news websites include SelangorKini, MalaysiaKini, Astro Awani, The Star, Sinar Harian, Utusan Malaysia and a few others. More news websites should cover the news, especially events or participations involving Selangor royal families, to increase awareness of the people and set an example among them. This is because many of the Selangor royal families' participation is beneficial to the people themselves, such as environmental campaigns, motivational talks, sports, donations and others. Besides, other types of websites that record the small coverage of the Selangor royal family story is also suggested to take part more in publishing their stories. This is considering that each website type has its own group of audiences, which might be different from one another. Thus, the more websites covering the Selangor royal family story, the more people will eventually read about it. Garcia-Santiago and Olvera-Lobo (2021) mentioned that the development of worldwide access to information using WWW has resulted in vast numbers of users with particular needs.

The web archiving process, which was conducted through the crawling process, shows different outcomes between two different crawling tools. In terms of the web archiving success rate, the Conifer tool has more successful websites that have been archived as compared to the HTTrack, which is different by nine websites. The websites successfully and unsuccessfully archived by these two tools are also different. The two websites that were unable to be archived by Conifer were, on the contrary, successfully archived by the HTTrack. The reasons for these differences need further investigation and research. As stated by Ruest et al. (2021), who conducted research specifically on the web archiving tool named Heritrix, advanced knowledge of how its user needs the processing chain works. In addition, comprehension of the method to debug odd web behaviour during crawling process encounters by Heritrix is also important. Hence, this situation requires a developer and is apparently not fit for researchers of information professionals in conducting web archiving. Another difference between HTTrack and Conifer is the outcome of the archived websites. HTTrack will archive the web page (based on the link provided) together with its embedded objects and resources. All these archived files will be directly saved in the user's preferred location on the laptop. On the contrary, Conifer will archive in the form of web pages and store them on the user's online storage with the condition that users need to create an account to have continuous access.

6.0 Conclusion

This study is the first, to the best of the authors' knowledge, that conducted a web archiving initiative to preserve the web contents of the Selangor Royal family with social and cultural values. The web archiving process indicates that preserving web content for future knowledge discovery is essential as well as requiring a detailed and meticulous set of activities. This is due to the rapid advancement of technology, hence increasing challenges in archiving web content despite the development of a few web archiving tools. It can be concluded that more web archiving initiatives specifically for web content related to the Royal Institution should be conducted to preserve its value for long-term accessibility. The limited expertise related to technology is the limitation in conducting research on web archiving initiatives as we may perceive that technology can improve the web archiving process efficiently. Thus, it is recommended that future research involves technology expertise in the web archiving initiative. However, this is only possible by spreading extensive awareness on web content vulnerability and how significant it is to protect the content from factors that can threaten its value and accessibility. Therefore, only with the awareness of a particular group of people, such as researchers, archivists, information professionals, and others, will they feel the urge to plan and execute web archiving initiatives. It is crucial to highlight the high possibility of web content changes and permanent loss from the internet. Once it disappears from the Internet, it will not be able to retrieve anymore, and this is something that must be prevented. In addition, a crucial part of web archiving is the contribution to knowledge discovery through web exploration. Aside from exploring the websites' profiles, the content coverage of the website is also being assessed to ensure it is related to Selangor royal families to avoid the waste of resources in archiving irrelevant web content. Moreover, future research should explore other web archiving tools and implement different approaches to identify the most efficient web archiving process and improvise present approaches suitable to the advancement of technology.

References

Archive-It Blog. (n.d.). Announcing the Web Archiving Life Cycle Model. [online] Available at: https://archive-it.org/blog/post/announcing-the-web-archiving-life-cycle-model/ [Accessed 26 Jun. 2023].

Anjum, A. & Anjum, A. (2012). Aiding web crawlers; projecting web page last modification. In 2012 15th International Multitopic Conference (INMIC) (pp. 245-252). doi: 10.1109/INMIC.2012.6511443.

Bragg, M., & Hanna, K. (March, 2013). The web archiving life cycle model. Archive-It. https://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

Bahry, F.D.S., Amran, N., Peter, P.A., Anyi, V.U., Othman, M.S., Mat Sedi, N.H.N., Sahid, S.S. & Munawar, A.A.A.A. (2019). Web Archiving Implementation for Retrieving JAWI Writing Collection in Web Sphere. Research Hub, 5(11), 17-25.

Chu, C., Wang, S., & Lai, Y. (2010). A study of Web 2.0 based digital archives system using Kano Model. International Computer Symposium (ICS2010), 171-176. doi: 10.1109/COMPSYM.2010.5685524.

García-Santiago, L. and Olvera-Lobo, M.-D. (2021). How accessibility guidelines are used in Spanish World Heritage websites: an exploratory study. Library Hi Tech, ahead-of-print(ahead-of-print). doi:https://doi.org/10.1108/lht-05-2019-0113.

Hegarty, K. (2022). The invention of the archived web: tracing the influence of library frameworks on web archiving infrastructure. Internet Histories, pp.1–20. doi:https://doi.org/10.1080/24701475.2022.2103988.

Holzmann, H., Goel, V., & Anand, A. (2016). ArchiveSpark: Efficient Web archive access, extraction and derivation. IEEE/ACM Joint Conference on Digital Libraries (JCDL), 83-92. http://dx.doi.org/10.1145/2910896.2910902.

Ibrahim, N.R., Zakaria M. S.,Zainuddin, N., and Zakaria, M. (2022). A Study on Children Customary Clothes in Malay Head Shaving - Cukur Jambul Ceremony for The Malay Royal Tradition. 7(2), pp.50–58. doi:https://doi.org/10.24191/idealogy.v7i2.335.

Jayanetti, H. R., Jones, S. M., Klein, M., Osbourne, A., Koerbin, P., Nelson, M. L., & Weigle, M. C. (2022). Creating structure in web archives with collections: different concepts from web archivists. *Research Gate*. https://doi.org/10.48550/arXiv.2209.08649.

Kawano, H. (2011). Hierarchical storage systems and file formats for web archiving. 21st International Conference on Systems Engineering, 217-220. doi: 10.1109/ICSEng.2011.46.

Krol, K. (2021). Digital cultural heritage of rural tourism facilities in Poland. Journal of Cultural Heritage Management and Sustainable Development, 11(4), 488-498. https://doi-org.ezaccess.library.uitm.edu.my/10.1108/JCHMSD-10-2019-0130

Khan, M. & Rahman, A.U. (2019). A systematic approach towards web preservation. Information Technology and Libraries, 38(1), 71-90.

M.K, I. K., M. Tahir, O., Ismail, N. A., & Muslim, Z. (2020). The heritage plant: The nostalgia from the past enhance the place identity for the royal town of Kuala Kangsar, Perak, Malaysia. International Journal of Advanced Research, 8(3), 328-336. http://dx.doi.org/10.21474/IJAR01/10632

Muhammad, S.A. and Rosdi, N.M. (2022). Understanding the Socio-cultural Sustainability of Kedah Royalty and the Influences on the Palace's Architecture. 1102(1), pp.012069–012069. doi:https://doi.org/10.1088/1755-1315/1102/1/012069.

Muslim, N., Hassan, W.Z.W. and Jodi, K.H.M. (2020). Constitution and Building of Nation-State in Malaysia. Journal of Politics and Law, [online] 13(3), p.122. doi:https://doi.org/10.5539/jpl.v13n3p122.

Rockembach, M. (2018). Web archiving: International case studies and the brazilian case. Digital Journal of Library and Information Science, 6(1), 7-24. doi: 10.20396/rdbci.v16i1.8646067.

Ruest, N., Fritz, S., Deschamps, R., Lin, J., & Miligan, I. (2021) From archive to analysis: Accessing web archives at scale through a cloud-based interface. International Journal of Digital Humanities, 2, 5–24 (2021). https://doi.org/10.1007/s42803-020-00029-6.

Shah, D.A. (2022). The 'Three RS' in Malaysia's Struggle for Constitutional Democracy. Federal Law Review, p.0067205X2210874. doi:https://doi.org/10.1177/0067205x221087460.

Sodangi, M., Kazmi, Z., Bakri, M. H., & Anwar, F. (2020). Protection of buildings with historic, architectural & cultural values: The case of royal museum of Seri Menanti. 2020 Second International Sustainability and Resilience Conference: Technology and Innovation Building Designs, 1-6. doi: 10.1109/IEEECONF51154.2020.9319987.

Wijesundara,C & Sugimoto, S. (2018). Metadata model for organizing digital archives of tangible and intangible cultural heritage and linking cultural heritage information in digital space. Libres, 28(2), 58-80. doi: 10.32655/LIBRES.2018.2.2.

www.igi-global.com. (n.d.). What is Memory Institutions | IGI Global. [online] Available at: https://www.igi-global.com/dictionary/collaboration-on-public-programming-by-memory-institutions-in-botswana/69080 [Accessed 12 December. 2022].

Yu, F., Wenbin, C., Dan, F., & Liang, T. (2009). Developing web archive for digital preservation of university. International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, 87-92. doi: 10.1109/EEEE.2009.35.