Web Archive Metadata Elements for Selangor Royal Web Archive Content Collections

Noraizan Amran*, Farrah Diana Saiful Bahry*, Shamila Mohamed Shuhidan, and Haslinda Husaini

School of Information Science, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA (UiTM), 40150 Puncak Perdana, Selangor, Malaysia

noraizan@uitm.edu.my; farrahdiana.uitm.edu.my

Received Date: 20 August 2023
Acceptance Date: 15 September 2023
Published Date: 1 November 2023

Abstract. Describing web archive metadata in implementing web preservation is essential in the overwhelming digital born documents' generation. The primary purpose of metadata generation in this study is to encourage preservation initiative specifically for Selangor royal web collections. The Design Science Research (DSR) stages implemented in this study are significant indication that display the tasks involved made it possible for the web archiving implementation and metadata creation. A total of sixteen Selangor royal web archive metadata elements were created and identified after validating and mapping with existing metadata standards. Those metadata elements are functional in the implementation of object structure of web preservation repository. This study is indispensable to contribute quality of content retrieval and cultivating a good practice on web archiving initiatives both in social and practical context.

Keywords: Royal web content, metadata, web archive, cultural heritage, information management.

1 Introduction

The essential role of web archiving the increasing digital divide. Web archive can be referred as the copies of preserved web content which collected for permanent retention and access; or an organization which dedicated its commitment primarily in collecting and preserving web content (Dooley, 2018). Web archive repository which in general can be considered as the outcome from web archiving process. It is a central file storage location that has versioning control and facets on web archive contents that are sustainable, reliable and organized to function properly (Bahry et al, 2022). The effective web archive repository is crucial due to the short lifespan of digital objects in order to organize multiple versions of the same web content and maintain its semantic values.

Meanwhile, metadata is data about data or documentation that describes data. It is also defined as a structured data about an object in which it supports the functions that related with the designated object (Giannoulakis et al., 2018). In the context of memory institutions, the precise interpretation of metadata is the structured information regarding resources in which it describes, locates (discover or place), manages, provides access and uses digital information resources (Khan & Rahman, 2019).

While preserving the digital collections, metadata plays an important role to describe, locate, retrieve, manage and use the information resources. Without the implementation of metadata on the digital collections, it would be impossible for people to retrieve the collections from storage based on their information needs as well as unable to discover the content of any particular digital collection. The same concept applied to the archived web collections that being stored in web archive repository. The collections definitely in need of proper metadata description for storage, usage and preservation purpose. The web pages are easily updated, changed, moved and removed after its initial creation. It portrays that web page is dynamically changed which make it challenging to ensure the contents are consistent (Ozawa and Uehara, 2012; Agata et el., 2014). The web archiving initiative in Malaysia is negligence due to issues like massive storage used, privacy act violation and awareness of Malaysian citizen towards the value of digital contents (Bahry et el., 2019).

Web archive metadata in the context of this study is referring to the structured information that is essential to describe Selangor royal cultural heritage collection in web format. It has the same concept as cataloguing that is used to locate or retrieve the digital collections from the repository. The primary purpose of essential metadata identification in this study is to encourage preservation initiative specifically for Selangor royal web collections. This is aligned with research by Formenton & Gracioso (2022) that stated among of digital preservation strategies include an effective implementation of metadata standards with the aim to support management and to interpret as well as preserve digital objects of informational media. In addition, digital preservation of website content also included as the digital preservation strategy. Therefore, taking consideration of limited web archiving initiative in Malaysia especially on the royal institution related information and collections, the objectives of the study were written as below:

1. To preserve the web content related to Royal Malaysia which have cultural and historical values.
2. To map selected Royal Malaysia web content with relevant metadata element

The metadata identification contributed to the existing metadata standard and beenmapped specifically to Selangor royal web collections, however it might also applicable to other state royal collections.

Cultural heritage has been defined by UNESCO as the inheritance of physical artifacts and intangible attributes belongs to a group or society inherited from previous generations, preserved in the present and granted to provide benefit for future generations (Wijesundara & Sugimoto, 2018). Referring to the National Heritage Act 2005, Laws of Malaysia, heritage art is included in the category of "intangible heritage". It consists of performing arts, musical arts, customs and culture, traditional games, heritage food, fine arts and traditional clothing. In the 10th Malaysia Plan (2011-2015), Perbadanan Adat Melayu dan Warisan Negeri Selangor (PADAT) has carried out research and held a "Selangor State Identity Convention" to identify heritage arts that can be used as the identity of the State of Selangor. PADAT is an organization that focuses on exhibition, research, publication and collection activities for museums in the state of Selangor since 1975 under Enactment No. 6 of 1975. Now, PADAT under the patronage of D.Y.M.M Sultan of Selangor functions specifically to develop, preserve and disseminate Malay customs and heritage of the State of Selangor (Dewan Negeri Selangor, 2023). Selangor Royal Highness Sultan Sharafuddin Idris Shah Alhaj Ibni Sultan Salahuddin Abdul Aziz Shah Alhaj was proclaimed the 9th Sultan of Selangor on 22 November 2001, or 6 Ramadan 1422 Hijrah, following the demise of his father, Sultan Salahuddin Abdul Aziz Shah Alhaj on 21 November 2001. He was later enthroned and crowned in a ceremony full of Royal Customs in Balairung Seri, Alam Shah Palace, Klang on 8 March, 2003 or 5 Muharram 1424 Hijrah (Selangor State Government, 2023).

The rest of this paper is stratified into four parts. Section 2 showcases the proposed methodology. Section 3 reveled the results obtained while section 4 uttered the discussion of the study. Finally, the study is concluded in section 5.

2 Research Methodology

2.1 Materials and Components

The study adopting the application of Design Science Research (DSR) methodology in producing a data model for Selangor Royal Cultural Heritage Web Collections. The DSR stages implemented in this study was constructed inline with we archiving process, metadata creaation and database repository implementation. In order to assure the research aim implemented successfully, this study proposed a set of activities which based on the comparative studies on few related procedures on DSR applications, web archiving process and database design activities as follow:

Table 1 Initial stage of DSR procedure

DSR phases	Details process steps	Outputs
1. Awareness of Problem	Literature Review Problem identification	Research Proposal
2. Suggestion	Archiving Website Metadata Mapping	Tentative Design

The awareness of problem stage refers to identification of an interesting problem that needs a specific solution (Mdletshe et al., 2021) where the awareness can appear from various sources such as developments in industry or in a reference discipline. The output of this phase is a proposal of new research endeavor (Ardakan & Mohajeri, 2009). Literature review of previous studies was conducted to understand the concept of web archiving from different perspectives as well as to identify the issues and concerns revolve around the initiatives of preserving cultural heritage. This has led to the discovery of problems and concerns regarding the current state of web archiving initiatives especially in Malaysia which require specific solutions.

Next, the suggestion stage involved with integration of web archiving process and metadata mapping which adopted from Bahry et al. (2022). Prior to the database design process for web archiving repository which will take place in the development process, the study found the needs to identify the essential metadata elements to describe the web collections of Selangor royal cultural heritage. The identified websites or webpages on the Selangor royal cultural heritage will be crawl by using web archiving tools HTTrack and Conifer. Other tools as suggested in previous studies (Khan & Rahman, 2019; Dooley & Bowers, 2018) might be considered if it can precisely capture the content of the selected web collections. Based on the outcome from the web content analysis, the metadata elements of Selangor cultural heritage will be identified and mapped to be included in the data model of web archiving repository.

The section elaborates partly of whole study that end with metadata mapping. The final metadata elements and data model for visualizing Selangor royal cultural heritage will be included in the next stage of DSR as the main outcome of implementing the Royal Web archive database repository which directly indicate the accomplishment of the main research objectives.

3 Results

Metadata mapping was conducted to identify the essential elements for knowledge discovery of Selangor royal cultural heritage which aligned with the objective of this study. Mapping the metadata indicates the process of gathering all existing metadata standards, identify the characteristics and purpose of each standard before make comparison to select the most suitable elements to describe the web collections of Selangor royal. Findings from the website archiving process was beneficial in contributing to this metadata mapping step. Some of the categorization made during the website analysis were incorporate into the metadata elements as it was found suitable in describing the web collections of Selangor royal. Further understanding on the metadata concept was crucial in this study through well comprehension on the separate categories of metadata that reflects the key functionalities in a particular system. Hence, metadata categories found from previous studies based on its functionalities have been defined as follow Formenton & Gracioso (2022):

1. Descriptive Metadata - Comprise of detail for location, identification, or understanding which may include elements such as title, author, and subject. The main usages are for discovery, presentation and interoperability.

2. Structural Metadata – Details that explain the internal structure of the digital archive and the hierarchical relationships of the resources that are part of each other. The main usages are for navigation and presentation.

3. Administrative Metadata – Provide information that reinforce the management lifecycle such as creation, selection, and description of information resources which may include elements like file type and size, creation date/time and others. The main usages are for interoperability, digital object management and preservation.

4. Markup Languages – Comprise of metadata and flags for other structural or semantic features in the digital content such as paragraph, name, list and date where the main purposes are navigation and interoperability.

On the other hand, there are few metadata standards across different functionalities being identified in previous studies as listed in Table 2. From all these metadata standards, this study has selected five (5) standards that being used as the main reference in mapping the metadata for Selangor royal web collections. Table 3 presenting the five (5) standards that have been selected and its individual elements while Table 4 presents the description of each element that has been selected as the essential metadata elements for knowledge discovery of Selangor royal cultural heritage from web collections.

Table 2: Metadata Standards

Purposes/Uses	**Metadata Standard**
Main standard in Web domain	Dublin Core
Archival and museological domain	Encoded Archival Description (EAD) Visual Resources Association (VRA) Core Categories for the Description of Works of Art (CDWA)
Bibliographic domain	Machine Readable Cataloging (MARC) Metadata Object Description Scheme (MODS)
For describing websites/archived site collections	WAM Data Dictionary
For textual and image-based works	Metadata Encoding and Transmission Standard (METS)
Others	PREservation Metadata: Implementation Strategies (PREMIS) UNESCO's World Heritage metadata

There are few metadata standards across different functionalities being identified in previous studies such as Dublin Core, EAD, VRA, CDWA, MARC, MODS, WAM Data Dictionary, METS, PREMIS and UNESCO world Heritage metadata. The metadata standards can be referred at Table 1 with the purpose and use of the metadata standard.

From the metadata standards, this study has selected five (5) standards that being used as the main reference in mapping the metadata for Selangor royal web collections. Table 3 presenting the five (5) metadata standards that have been selected and its individual metadata mapping elements with Selangor Royal Metadata, which are Qualified

DC, WAM data dictionary, MODS, UNESCO world Heritage metadata and metadata standard (neutral standard).

Table 3: Selangor Royal Metadata elements comparative and mapping

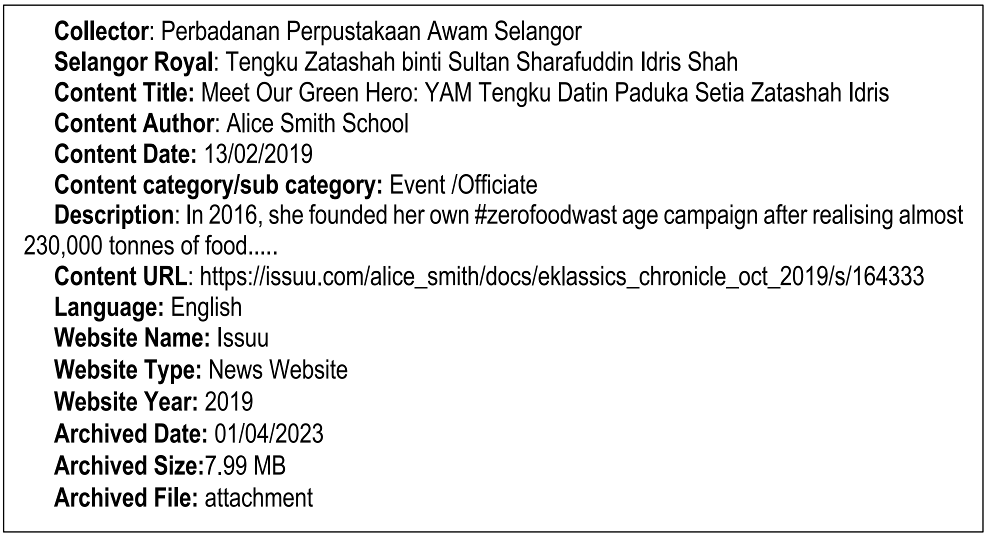
Selangor Royal Metadata	**Qualified DC (version 1.1) (Formenton & Gracioso, 2022)**	**WAM data dictionary (Dooley & Bowers, 2018)**	**MODS (version 3.7) (Formenton & Gracioso, 2022)**	**UNESCO's World Heritage metadata (Tan et al. (2019)**	**Mfigures descriptive metadata (Bahry et al. 2022)**
Collector	Title	Collector	Title information (<titleInfo>)	Heritage ID	Collector
Royal family	Creator	Contributor	Name (<name>)	Title	Extent
Content title	Subject	Creator	Resource type (<typeOfResource>)	Heritage Type	Source of description
Content author	Description	Date	Genre (<genre>)	Heritage Period	Contributor
Content date	Contributor	Description	Origin Information (<originInfo>)	Heritage Time Span	Genre/Form
Content category, sub-category	Date	Extent	Language (<language>)	Purpose (reason recorded)	Subject
Description	Type	Genre/Form	Physical Description (<physicalDescription>)	Recording Device Parameters	Creator
Content URL	Format	Language	Abstract (<abstract>)	Secondary Device	Language
Language	Identifier	Relation	Table of Contents (<tableOfContents>)	Environmental Conditions	Title
Website name	Source	Rights	Note (<note>)	Submitter and Date of Submission	Date
Website type	Language	Source of description	Subject (<subject>)	Rights Given	Relation
Website year	Relation	Subject	Related item (<relatedItem>)	Author (Copyright Holder)	URL
Archived date	Coverage	Title	Identifier (<identifier>)	Sponsor (client)	Description
Archived size	Rights	URL	Location (<location>)	Date (of recording and manipulation)	Rights
Archived file	RightsHolder		Access Condition (<accessCondition>)	Location (Latitude/Longitude and direction)	
	Provenance		Part (<part>)		
Royal Genecology			Extension (<extension>)		

	Record Information (<recordInfo>)

The metadata elements and description of each element that has been selected as the essential metadata elements for knowledge discovery of Selangor royal cultural heritage from web collections which are collector, Selangor Royal, Royal Genecology, content title, content author, content date, content category, sub-category, description, content URL, language, website name, website type, website year, archived date, archived size and arcived file as in table 4.

Table 4: Selangor royal Medata elements

Selangor Royal Metadata	
Metadata elements	**Description of element (s)**
Collector	Institution that collects the web content for preservation
Selangor Royal	Selangor Royal figure or the institution itself
Royal Genecology	Royal Geancology
Content title	Title of a web content or article
Content author	Author of a web content or article
Content date	Published date of a web content or article
Content category, sub-category	Category of a web content/article such as event, campaign, background, Selangor history
Description	Brief description of a web content/article
Content URL	URL of a web content/article
Language	Language used to deliver a webs content/article
Website name	Name of website that reported the content/article
Website type	Type of website such as news website, corporate website, informative website etc
Website year	Copyright year at a point of web content being archived
Archived date	Date of archiving process being conducted
Archived size	Size of web content that has been archived
Archived file	Attachment of an archived web content



Collector: Perbadanan Perpustakaan Awam Selangor
Selangor Royal: Tengku Zatashah binti Sultan Sharafuddin Idris Shah
Content Title: Meet Our Green Hero: YAM Tengku Datin Paduka Setia Zatashah Idris
Content Author: Alice Smith School
Content Date: 13/02/2019
Content category/sub category: Event /Officiate
Description: In 2016, she founded her own #zerofoodwast age campaign after realising almost 230,000 tonnes of food.....
Content URL: https://issuu.com/alice_smith/docs/eklassics_chronicle_oct_2019/s/164333
Language: English
Website Name: Issuu
Website Type: News Website
Website Year: 2019
Archived Date: 01/04/2023
Archived Size:7.99 MB
Archived File: attachment

Figure 3: Sample data for Metadata Elements of Selangor Royal Cultural Heritage

4 Discussion

There are different categories of metadata during DSR phases, with different functionalities which include Descriptive Metadata, Structural Metadata, Administrative Metadata and Markup Languages. These categories are also representing the purpose of metadata itself that must no be disregarded during metadata elements identification. In order to adopt an effective metadata on the digital collections, the purpose must be achieved. Hence, Formenton and Gracioso (2022) in their research has highlighted that preservation metadata schemes may include elements from multiple categories as mentioned above. It means that a metadata implementation is not necessarily adopt a single category but all the stated categories.

In this study context, the metadata categories that have been adopted were made up of two categories which are descriptive and administrative metadata. The descriptive metadata was adopted as its purpose is to locate and identify the digital collections, in which for the Selangor royal web collections metadata it comprises of collector, title, author, theme, and URL. The second category which is administrative metadata was adopted as the primary uses are for digital object management and preservation. The essential elements selected for this category include the website type, archived content date, archived content file, size and date of the web being archived. Both metadata categories chosen for describing the web collections of Selangor royal are mainly targeted for locating the collections, describing the important details of the web content and preserving the web contents.

Information content of Selangor Royal can be easily retrieved and compiled base on the Royal Selangor searching criteria, example using keyword “Zatashah”. However, the search result is depends on how the Royal Selangor name being entered in the web archive repository. Therefore, the genealogy information of the royal Selangor family needs to be identified first. Selangor Royal Highness Sultan Sharafuddin Idris Shah Alhaj Ibni Sultan Salahuddin Abdul Aziz Shah Alhaj has been blessed with two Princesses, Tengku Zerafina and Tengku Zatashah, as well as a Prince, Tengku Amir Shah, Raja Muda of Selangor. Sultan Sharafuddin Idris Shah Alhaj has married Tengku Permaisuri Hajah Norashikin at the Royal Palace Mosque, Istana Alam Shah, Klang (Selangor State Government, 2023). The Selangor Royal data can be expanded to Selangor Royal full name, title, birth of date, genealogy and others, for more specific searching. Website which contained information of Duli Yang Teramat Mulia Tengku Amir Shah Ibni Sultan Sharafuddin Idris Shah Alhaj can be found at Selangor State Government website (https://Www.Selangor.Gov.My/Index.Php/Pages/View/1993?Mid=896). This valuable information can be stored in the repository and be used for future references.

The content title can be stored as description and can be searched using wild searching for any information content. However, for easy retrieval, it can be categorized by similar activities, name and others, example by event, location, person, international etc. Other than that, there are various website which contained information related to Royal Selangor Palace such as Istana Alam Shah (https://www.visitselangor.com/istana-alam-shah/), Alam Shah Palace (https://selangor.travel/listing/alam-shah-palace/), Royal Klang Town Heritage Walk(https://www.mpklang.gov.my/en/node/3691), Istana Alam Shah (https://www.mpklang.gov.my/ms/istana-alam-shah) and Istana Bandar Jugra (https://www.heritage.gov.my/index.php?option=com_content&view=article&id=78&Itemid=521&lang=ms).

The web archive metadata plays an important role to manage the web archive content and retrieve the web collections from digital storage based on the information needs by royal institutions, countries, agencies, researchers and others. This intangible heritage information or valuable digital asset of Royal Selangor can be easily retrieved and can be referred as the identity of the State of Selangor for exhibition, research, publication and collection activities for museums in the state of Selangor.

5 Conclusion and Recommendations

In conclusion, web archiving implementation requires comprehensive life cycle that accompanied with detailed steps at each phase of flow process. Various different aspects need to be addressed, defined and close monitoring including filtering and selection of websites to be archived, platform or tools to be used, storage as repository to preserve the archived content, metadata to describe the web collections, organization of the digital repository and many more. In order to successfully adopting web archiving initiative to preserve any website collections, all the essential aspects must be properly defined during the planning phase.

In the context of this study, it caters only to the part of web archiving life cycle that associated with metadata and storage aspects. However, the process to identify essential metadata elements and designing the data model which part of web archiving life cycle has already involved an extensive analysis and exploration on the website contents available on the internet. This indicates that a complete and comprehensive web archiving implementation that any organization may interested to pursue, must be conducted with detailed and comprehensive planning to avoid any issues that may hinder the implementation all together. The outcome from website analysis and metadata mapping of this study were used as the main input to create the data model. Entity Relationship (ER) model has been chosen as the data model to present the essential metadata elements for Selangor royal cultural heritage web collections. Hence, this study is able to contribute in existing research on web archiving initiative both in social and practical context.

Acknowledgments

We would like to express our sincere gratitude to Dana UiTM Cawangan Selangor (DUCS) for the provision of financial support to conduct this research, UiTM Research Management Center (RMC) for coordinating and monitoring this research to ensure it is progressing as scheduled, our Research Assistant (RA) Miss Masturah Mohd Nor and others who have contributed to this study. This is part of work under the DUCS-P grant (600-UITMSEL (P.1 5/4)(049/2022).

References

Agata, T., Miyata, Y., Ishita, E., Ikeuchi, A., & Ueda, S. (2014, September). Life span of web pages: A survey of 10 million pages collected in 2001. In *IEEE/ACM Joint Conference on Digital Libraries* (pp. 463-464). IEEE

Bahry, F. D. S., Amran, N., Putri, T. E., & Ramli, M. I. (2022). Database design of the Malaysia public figures web archive repository: a social and cultural heritage web collections. *Collection and Curation*, *41*(4), 133-140.

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.

Dewan Negeri Selangor. Retrieved on 6th September 2023 at https://dewan.selangor.gov.my/question/adat-dan-seni-warisan-selangor/

Dooley, J. (2018). Descriptive metadata for web archiving: recommendations of the oclc research library partnership web archiving metadata working group. OCLC Research.

Formenton, D., & de Souza Gracioso, L. (2022). Metadata standards in web archiving technological resources for ensuring the digital preservation of archived websites/Padroes de metadados no arquivamento da web: recursos tecnologicos para a garantia da preservacao digital de websites arquivados. Revista Digital de Biblioteconomica e Ciencia da Informacao, 20(1), NA-NA.

Giannoulakis, S., Tsapatsoulis, N., & Grammalidis, N. (2018). Metadata for intangible cultural heritage. In *Proceedings of the 13th international joint conference on computer vision, imaging and computer graphics theory and applications (VISAPP 2018)* (pp. 634-645).

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

Musa, S. H., Sauti, N. S., Nasir, F. M., Abdullah, N. A., & Ahmad, Y. (2020). The Development of GISBased Digital Archive for Heritage Building in Melaka World World Heritage Site (UNESCO), Malaysia. International Journal of Engineering Advanced Research, 1(3), 1-16.

Ozawa, R., & Uehara, M. (2012, September). Long term management of web cache for web archive. In 2012 15th International Conference on Network-Based Information Systems (pp. 639-644). IEEE.

Selangor State Government. Retrieved 11th September 2023 at https://www.selangor.gov.my/index.php/pages/view/77?mid=523

Tan, Y. Y., Rafi, A., Musa, S. N., & Anuar, A. A. M. (2019, May). Design and Development of e-Warisan Senibina Portal: A Web-based Knowledge System for Architectural Virtual Heritage Data and Heritage Education. In Proceedings of the 2019 5th International Conference on Education and Training Technologies (pp. 130-134).

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: Library and Information Science Research Electronic Journal, 28(2), 58-80.