Mapping and Archiving Complex Cultural Heritage Digital Objects: Literature Review

Laaiqah Hassim
Computer Science Department
University of Cape Town
Cape Town South Africa
HSSLAA002@myuct.ac.za

ABSTRACT

Advancements in technology have led to the development of an efficient mechanism for preserving cultural heritage in digital archives. A successful demonstration of a cultural heritage archive is Europeana. It is a large-scale search engine for cultural heritage from across Europe. Many successful cultural heritage platforms use available open-source software for their infrastructure, such as The Louvre using Drupal to manage digital content. Other content management systems include Omeka and Islandora. Despite the research going into the continuous development of cultural heritage archives, there is a gap that has yet to be filled. Cultural heritage archives, such as Europeana, allow for functionalities such as searching, browsing, and visualizing digital objects, however, no tool allows users to create maps with digital objects and have this composite structure exported. This paper draws on secondary literature to expand the functionality of digital archives by discussing related work on diagraming and exporting complex digital objects.

Keywords

Digital Archives; Cultural Heritage Preservation, Diagraming, Visualization, Digital Objects; Content Management Systems; .

1. INTRODUCTION

The advent of digitization of cultural heritage, such as cultural heritage archives, serves as an efficient mechanism to ensure the long-term preservation of heritage [16] while making it more accessible for students, researchers, and the community.

Cultural heritage archives provide storage, searching, browsing, and visualizing functions on digital objects. Storage allows the preservation of digital objects, searching and browsing allows the discovery of digital objects and visualizing allows digital objects to be viewed. These serve as the basic functionality of any digital archive system.

However, there is no efficient tool that enables end-users to create diagrams with digital objects, and allow these mappings to be exported. In this study, we will review material that can direct us to creating a tool to effectively map cultural heritage digital objects, and allow the complex object formed to be exported into a format that can be archived for future use.

For this research project, we will work with the Five Hundred Year Archive (FHYA). FHYA is a project-based in the History Department at the University of Cape Town, which attempts to document aspects of South Africa's pre-colonial heritage by collecting heritage material from archives and museums from across the globe.

This paper begins by outlining the building blocks of digital archives, focusing on digital objects and their functionalities, and provides an example of a successful digital cultural heritage archive called Europeana. It then introduces descriptions of tools providing digital archives functionality and follows with an analysis of the tools. Finally, this paper wraps up with conclusions drawn from this literature review.

2. Building Blocks

Digital objects are data stored in repositories, controlled by rules governing the interaction of the digital objects and its repository [2], called a Repository Access Protocol. Digital object types include but are not limited to, Text, Image, Video, and Audio. Below is an overview of storage and the interaction of digital objects within repositories and functions performed on digital objects.

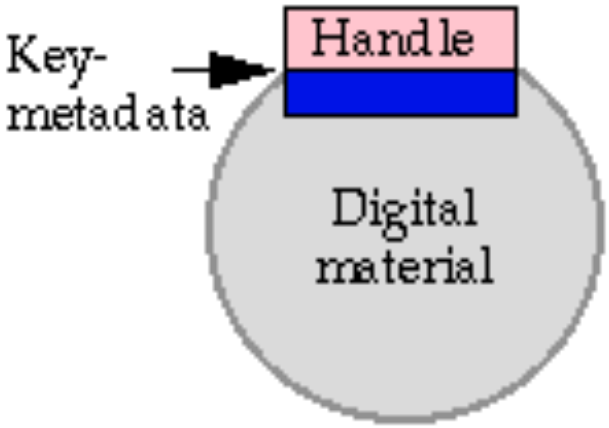

Figure 1. Digital Object Structure [2].

2.1 Open Archives Initiative Protocol for Metadata Harvesting

The Open Archives Initiative Protocol for Metadata Harvesting (OAI - PMH) serves as a protocol to transfer metadata from a source archive to a destination archive.

Metadata is the data that describes digital objects. It structures the data referring to digital objects in a way that allows it to discover and use the content of digital collections and repositories [12]. Since OAI-PMH uses an object's metadata, it needs a standard to structure its metadata. OAI-PMH requires that the digital objects must, at a minimum, offer unqualified Dublin Core metadata [15].

2.2 Metadata Standards

There are many standard metadata formats. The metadata format chosen to represent digital objects depends on many factors, such as the type of digital objects and its purpose. Below are two important metadata standards used in digital libraries.

2.2.1 Dublin Core

Dublin Core is an example of a descriptive metadata standard [12]. It is one of the simplest and most popular metadata formats, which is why OAI-PMH requires it as a minimum requirement [15]. It supports heterogeneity- an imperative feature to service and store diverse cultural heritage digital object types [10]. Its simplicity acts as an advantage but can also be a drawback. Cultural heritage digital objects store rich data and, by storing its metadata in this simple format, information of these objects can be lost [10].

2.2.2 METS

METS is a structural metadata type [12] that uses an XML document format to encode complex objects within digital libraries [12]. Where Dublin Core stores descriptive metadata, METS provides a framework for descriptive, administrative, and structural metadata [12]. It provides an efficient mechanism for storing complex digital objects such as those found in cultural heritage archives. It also supports interoperability, which is useful for aggregating digital content from different repositories [12].

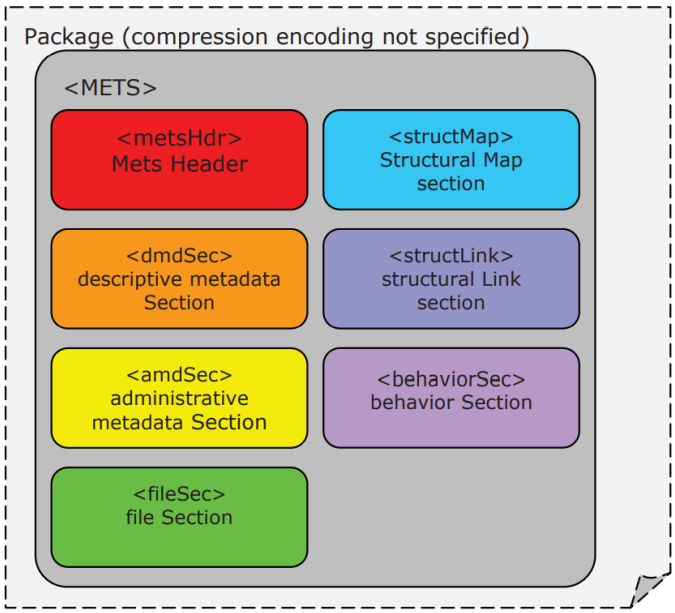

Figure 2. Internal structure of METS format [8].

2.3 Functionalities

Digital archives' main functionality is to preserve digital objects. Other important features include searching, browsing and visualizing digital objects.

Europeana aggregates metadata from libraries, archives, museums and galleries across Europe. It consists of heterogenous, multilingual digital objects, with different metadata standards. To allow storage of these vast standards a Europeana Data Model (EDM) was created to store the multiple' formats, with multiple languages [19].

Its technical architecture is built ultimately around the open-source search platform, Apache Solr6.

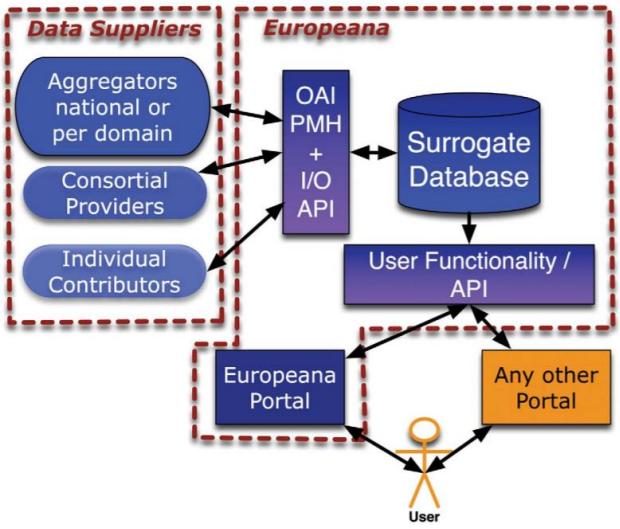

Figure 3. Overview of Europeana's Architecture [6].

3. Architecture Design

The high-level architecture of a digital archive consists of: a user-interface layer that enables the user to access cultural heritage digital objects, a service layer that allows the discovery and manipulation of digital objects, and the repository layer which stores and manages digital objects [1].

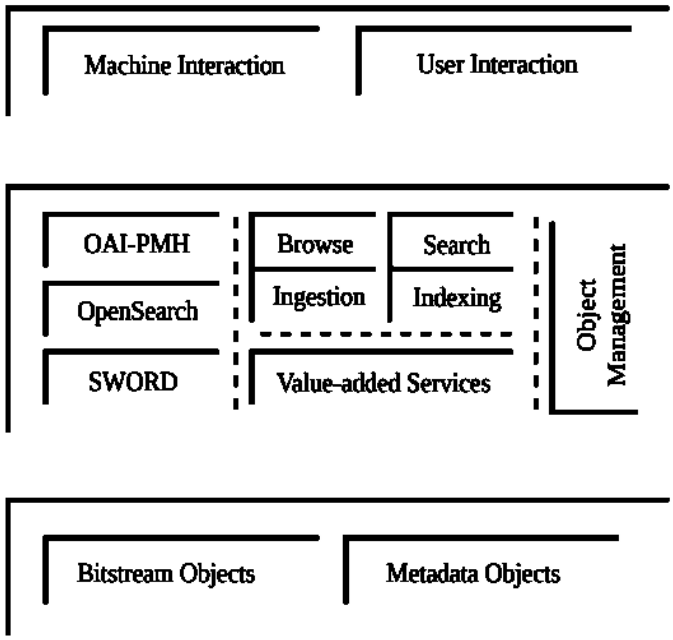

Figure 4. High Level Schema for Digital Archive System [21]

Many factors must be considered when creating an architecture for a cultural heritage system.

The system needs to be scalable. A successful cultural heritage digital archive tool architecture must support the heterogeneity of digital objects, considering the vast diversity of cultural heritage digital object types. It needs to consider the number of users accessing the system at any given time and the ways content of a system should be accessed should be decided upon [21].

4. Tools Analysis

4.1 Content Management Systems

Content management systems create, store, edit, secure, preserve, transform and publish original and acquired digital content. Below we discuss three important open-source content managing systems used in cultural heritage archive platforms. The ultimate difference between these content management systems is the way they manage their digital objects.

4.1.1 Drupal

Drupal is an open-source content management system that supports a range of tools, such as Apache web server and the MySQL database server, and various metadata standards and digital object types [3]. Although it is highly interoperability and customizable, making it good for integrating into other systems, it has a steep learning curve making it difficult for those who have never built a site on Drupal before [3]. Drupal does not offer complete hosting, which requires users to download the software and develop the site locally [3]. Drupal plays a significant role in digital cultural heritage projects. Currently, it is part of the infrastructure for The Louvre website. It is also used for managing and presenting content for the world's largest museum library system [13], The Smithsonian Libraries [3].

4.1.2 Omeka

Omeka is an exhibition creation platform that allows end-users to organize and share digital collections and archives [20].

End-users choose to either install Omeka for free on their servers or use a hosted version on Omeka.net [20]. The hosted version can be purchased in 5 ways. There is one free plan, and four paid plans [20]. Basic functionality includes uploading any form of digital object, adding metadata, and grouping objects so that it can be displayed. Omeka allows the personalization of visuals such as theme selection and creation of themes [20].

Omeka has features that are favorable to cultural heritage archives. Its underlying layer is structured heavily with a metadata scheme that allows interoperability with other content managing systems. It lets users build exhibitions that can be personalized with existing or user-created themes [20]. Omeka's user-friendly platforms allow people with a range of expertise to use it easily [20]. Omeka's adaptability and customizability serve as an advantage. It can be integrated with other applications and has a vast range of plugins that can be added to extend its functionality. In terms of cultural heritage exhibitions, plugins include maps, timelines, etc. [20]. Although Omeka can be used to store digital archives, create collections and visualize exhibits, it cannot export entire exhibits that users create.

4.1.3 Islandora

Islandora is a general-purpose repository platform that integrates three open-source tools [9]: Drupal, responsible for the user interface, Fedora, responsible for managing digital objects and Solr, responsible for indexing [15]. Just like Omeka, it can be used to organize, view, edit, and find digital objects [15]. As with Omeka, Islandora can store any digital object type, with any metadata standard [9].

4.2 Complex Object Systems

4.2.1 WARC

WARC (Web ARChive) is a container file standard used to store web content in its original context. The WARC format specifies a method to combines various digital content into a single archival file- together with its related information. The WARC format is a typical standard to structure, manage and store billions of collected digital resources. It can be used to build applications for harvesting, managing, accessing, and exchanging content.

The method used to create and WARC files will be created differs depending on the software and application implementations. Once a function calls to harvest a web page, metadata, called warcinfo records are added to the beginning of a file to generate a WARC file, as seen in figure 5.

```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2012-01-18T22:12:49.445Z
WARC-Filename: MY_WARC.warc
WARC-Record-ID: <urn:uuid:98187a24-8d74-a2b8-ec19-fbb6a958db9e>
Content-Type: application/warc-fields
Content-Length: 541

Software: WARCreate/0.4.1 http://matkelly.com/warcreate
ip: 128.82.5.133
hostname: cs.odu.edu
format: WARC File Format 1.0
conformsTo: http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf
```

Figure 5. Example of a WARC file generated with warcinfo records [14].

4.2.2 BagIT

BagIT is a way to transfer content for digital preservation [4]. Content is packaged in a bag that is designed to be flexible, allowing the transfer of different types of digital objects. A bag requires three things: a bag declaration tag, a list of the content files, and the content itself. Bags can be sent via computer networks or physically moved with the use of portable storage devices such as USBs. On receiving a bag, the computer analyzes it to see if the required files are present – which results in a successful transfer of digital content. Bags are containers with built-in inventory checking, so content transferred can be confirmed. BagIt helps support digital preservation by supporting the successful transfer of files from one computer to another.

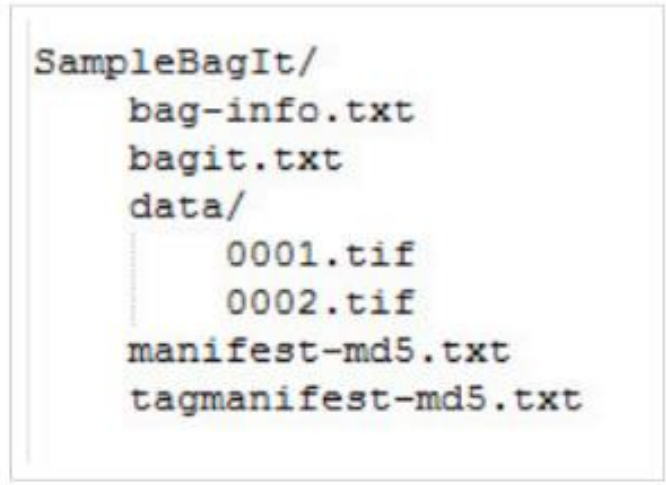

Figure 6. Example of a BagIt Bag and its contents [7].

4.2.3 Walden's Path

Walden's Paths provides annotated guided paths over World-Wide Web pages. Walden's Paths consists of three components, a Path Authoring Tool for creating and editing paths, a Path Database for storing, retrieving, and sharing paths, and a Path Server that provides access to published paths. These components work together as follows:

The Path Authoring Tool allows keyword searches for Web materials and displays it in a browser.

These authored paths are stored in the Path Database. This database provides each authorized path author with a working area for storing paths. When a path is ready for access by readers it is "published" to the Path Server. The Path Server is a Common Gateway Interface (CGI) program that creates the list of available paths and their presentation for readers. When a reader requests a path page, the Path Server constructs control-flow and annotation frames to appear at the top of the browser and requests the material from the source site on the Internet to appear in a lower frame. The reader can use the control-flow controls to step along the path. In addition, the links on the source page remain active.

5. CONCLUSIONS

Cultural Heritage archives serve as an efficient mechanism for end-users to access information from around the world, anywhere. There is a lot of research going into these archives and finding ways to make it more effective.

This literature review serves as a starting point in developing a tool that will allow end-users to create diagrams with cultural heritage digital objects, resulting in a complex object and allow this to be stored in a format that can be stored and retrieved from an archive.

6. REFERENCES

[1] Arms, W. 2001. *Digital libraries*. MIT Press, Cambridge, Mass.

[2] Arms, W., Blanchi, C. and Overly, E. 1997. An Architecture for Information in Digital Libraries. *Mirror.dlib.org*. http://mirror.dlib.org/dlib/february97/cnri/02arms2.html#digital-object.

[3] Avgousti, A., Papaioannou, G. and Gouveia, F. 2019. Content Dissemination from Small-scale Museum and Archival Collections: Community Reusable Semantic Metadata Content Models for Digital Humanities. *code{4}lib journal*. https://journal.code4lib.org/articles/14054.

[4] Bagit: Transferring Content for Digital Preservation. 2009. *The Library of Congress*. https://www.loc.gov/item/webcast-4682/.

[5] Belhi, A., Foufou, S., Bouras, A. and Sadka, A. 2017. Digitization and Preservation of Cultural Heritage Products. *IFIP International Conference on Product Lifecycle Management*, Springer, Cham.

[6] Concordia, C., Gradmann, S. and Siebinga, S. 2010. Not just another portal, not just another digital library: A portrait of

Europeana as an application program interface. *IFLA Journal 36*, 1, 61-69.

[7] Ferriter, M. 2021. BagIt at the Library of Congress | The Signal. *Blogs.loc.gov*. https://blogs.loc.gov/thesignal/2019/04/bagit-at-the-library-of-congress/.

[8] Godtsenhoven, K. and Vernooy-Gerritsen, M. 2009. *Emerging standards for enhanced publications and repository technology*. Amsterdam University Press.

[9] Jordan, M. and McLellan, E. 2016. PREMIS in Open-Source Software: Islandora and Archivematica. *Digital Preservation Metadata for Practitioners*. DOI: https://doi-org.ezproxy.uct.ac.za/10.1007/978-3-319-43763-7_16

[10] Kakali, C., Lourdi, I. and Stasinopoulou, T. et al. 2001. Integrating Dublin Core metadata for cultural heritage collections using ontologies. *International conference on Dublin core and metadata applications*, 128-139.

[11] Kent, A. 2014. Islandora: an open source digital repository solution. Computers in Libraries.

[12] Iris Xie, Krystyna K. Matusiak. 2016. Discover Digital Libraries.

[13] Kalfatovic, M., Kapsalis, E., Spiess, K., Van Camp, A. and Edson, M. 2008. Smithsonian Team Flickr: a library, archives, and museums collaboration in web 2.0 space. *Archival Science 8*, 4, 267-277.

[14] Kelly, M. and Weigle, M. 2012. WARCreate: create wayback-consumable WARC files from any webpage. *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 437–438. DOI: https://doi.org/10.1145/2232817.2232930

[15] Lagoze, C. and Van de Sompel, H. 2003. The making of the Open Archives Initiative Protocol for Metadata Harvesting. *Library Hi Tech 21*, 2, 118-128. DOI: https://doi.org/10.1108/07378830310479776

[16] Masenya, T. and Ngulube, P. 2019. Digital preservation practices in academic libraries in South Africa in the wake of the digital revolution. *SA Journal of Information Management 21*, 1. DOI: https://doi.org/10.4102/sajim.v21i1.1011

[17] McDonough, J. 2006. METS: standardized encoding for digital library objects. *International Journal on Digital Libraries 6*, 2, 148-158. DOI: **https://doi.org/10.1007/s00799-005-0132-1**

[18] Mitchell, E. 2015. *Metadata standards and Web services in libraries, archives, and museums*.

[19] Petras, V., Hill, T., Stiller, J. and Gäde, M. 2017. Europeana – a Search Engine for Digitised Cultural Heritage Material. *Datenbank-Spektrum 17*, 1, 41-46. DOI: https://doi.org/10.1007/s13222-016-0238-1

[20] Puckett, J. and Leslie, S. 2016. Omeka. *Journal of the Medical Library Association*. http://dx.doi.org/10.3163/1536-5050.104.4.030.

[21] Ruthven, I. and Chowdhury, G. 2015. *Cultural heritage information*. Facet Publishing, [London].

[22] WARC, Web ARChive file format. 2021. *Loc.gov*. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml#:~:text=The%20WARC%20(Web%20ARChive)%20format,file%20together%20with%20related%20information.&text=The%20WARC%20format%20generalizes%20the,exchange%20needs%20of%20archiving%20organizations.