Techniques for Mapping Complex Digital Objects for Cultural Digital Libraries

Literature Review

Yashkir Ramsamy
Department of Computer Science
University of Cape Town
Cape Town, WC, South Africa
rmsyas003@myuct.ac.za

ABSTRACT

New tools and fields have emerged because of technology advancements, particularly in the subject of Digital Humanities, notably in its sub-field of Digital Archiving. In the past, heritage material would only be available for viewing through physical archives. Recently, much of this heritage material has been digitized and stored in digital databases as complex digital objects with the objective of preservation, as well as viewing these objects. With the addition of these digital archives to the field, arose the need for tools for digital library exploration, digital object creation, and virtual exhibition for viewing heritage material. Organizing digital objects according to the relationships among them and producing an archivable output is an area where the existing tools lack functionality. This paper reviews key concepts that would make up such a tool and covers traversing and mapping of (complex) digital objects and libraries through browsing and searching, concept maps and topic maps. Content Packing standards are also reviewed, where metadata and content containment types are discussed, and, finally, existing tools that offer similar functionality to what is required for best designing a tool that allows creation for the mapping of the relationships between digital objects from any given digital library.

KEYWORDS

Digital Libraries, Digital Archives, Complex Digital Objects, Metadata, Content Packaging

1. Introduction

The purpose of a mapping tool that allows for the organization of various digital heritage materials should allow for a better understanding of the material and its context that users can create [7]. An example of such material can be found in the Five Hundred Year Archive (FHYA)[1] [1], a digital library assembled from archives and museums from around the world that are specifically related to South African history 500 years before colonialism. An integral part of understanding this material stems from the visualization of relationships among the heritage material, thus this project aims to introduce a tool where this material can be structured in a Knowledge-Graph format using diagrams such as flowcharts, maps and organizational diagrams that should be self-contained, machine-readable and archivable.

This paper will discuss the related literature, such as searching through digital libraries, techniques that can be used for mapping the relationships among digital objects, content packaging standards, and the existing tools that provide similar functionality to what is required. The information drawn from these areas will aid in developing a tool for mapping the relations between digital objects or any other digital archive that conforms to the standards of a digital object.

2. Traversing Digital Libraries

This section describes the different methods for traversing through content in a digital library. Two exploration methods that will be reviewed are Searching and Browsing. In most digital libraries, both features are offered by different services as they utilize different techniques for displaying and searching for content [25].

2.1 Faceted Search

The conventional or traditional search technique known as lookup search assumes that a user knows the content they are searching for [22]. This is a limitation of a lookup search. When users are not sure of what to search for, it becomes difficult to construct a search query [22]. Faceted Search is one technique that is used to aid users in searching across dimensions, called facets, for broad search queries [25].

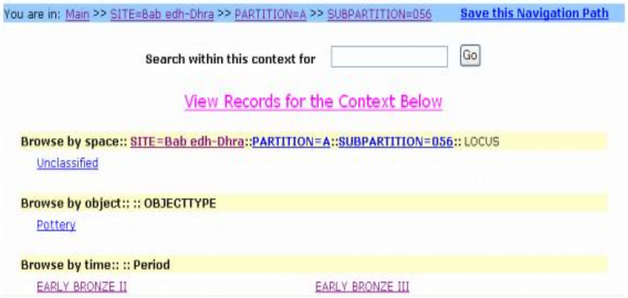
[1] The Five Hundred Year Archive. https://fyha.org

One practical example of a faceted search implemented in a digital archive is in the FHYA which incorporates a JavaScript-based search engine [39] that allows a searcher to drill through results based on the content features as a criterion for the resulting search [41].

2.2 Browsing Techniques

2.2.1 Multi-Dimension Browsing

In ETANA-DL[2], a digital library for archaeological data, it is possible for a user to browse through various dimensions simultaneously [25].



You are in: Main >> SITE=Bab edh-Dhra >> PARTITION=A >> SUBPARTITION=056 Save this Navigation Path

Search within this context for Go

View Records for the Context Below

Browse by space:: SITE=Bab edh-Dhra::PARTITION=A::SUBPARTITION=056:: LOCUS
Unclassified

Browse by object:: :: OBJECTTYPE
Pottery

Browse by time:: :: Period
EARLY BRONZE II EARLY BRONZE III

Figure 1. ETANA-DL's multi-dimensional browsing feature [25]

Figure 1 above depicts this browsing feature, as a user navigates through each path, a representation of the content that each dimension describes will appear [25]. Thus, this also serves as an exploratory search tool [25].

3. Mapping of (Complex) Digital Objects

Digital libraries are composed of (Complex) Digital Objects [17]. A Digital Object is defined as an object with more than one content file with corresponding metadata [5]. These objects thus contain multiple layers of complexity that should be accounted for when preserving them [9]. Since Digital Objects contain varying levels of information, this could imply the possible existence of related information or events between any two digital objects. The following are the techniques that can be used for mapping these relationships between digital objects.

3.1 Mapping Techniques

3.1.1 Concept Mapping

Diagrams that organize information in enclosed shapes with visually represented relationships are known as concept maps [27]. The use of tools that generate interlinking information in an organized manner is referred to as concept mapping [27].

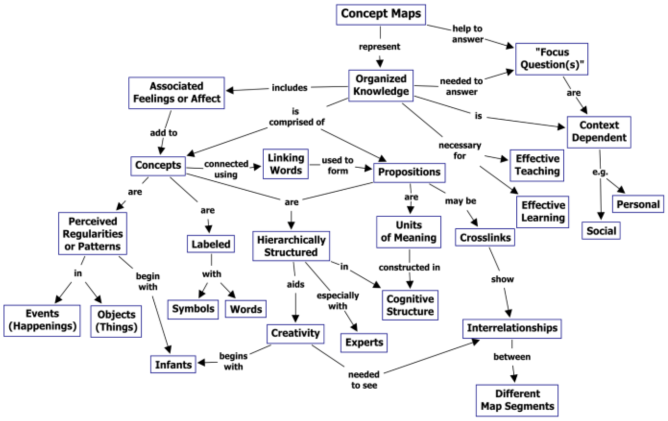


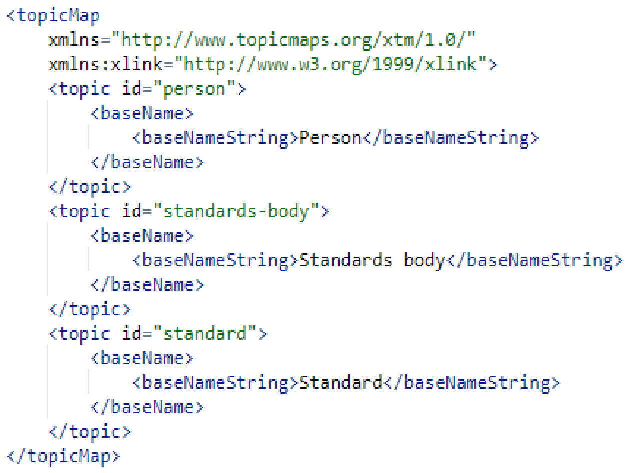
Figure 2. A Concept Map [27]

Figure 2 displays an example of such a concept map.
Concept maps allow their viewers to understand relationships between objects in a concise, reduced fashion [6]. Despite this, an increasing amount of information on a concept map is inversely proportional to its utility, attributable to the high number of related links between objects [35]. In an instance where a researcher requires the input of more information or a change of relational links, it may prove to be difficult [19,35].

3.1.2 Topic Maps

Topic maps are like concept maps in that they also visually represent relationships between informational objects. The difference between concept maps and topic maps is that topic maps represent this information with formally defined and structured graphs [20]. Topic maps are constructed with a specialized version of eXtensible Markup Language (XML), XML Topic Maps (XTM) [40]. Topic maps are defined in an ISO standard, known as ISO/IEC 13250 [20, 26]. This means that Topic Maps are an industry standard and are machine-readable maps. This is also implied by its use of XML which is designed to be machine-readable [16]. Topic maps are not restricted to a domain, or the types of data that it can model [26].

[2] ETANA-DL: A Digital Library for Integrated Handling of Heterogeneous Archaeological Data. http://www.etana.org/



```
<topicMap
    xmlns="http://www.topicmaps.org/xtm/1.0/"
    xmlns:xlink="http://www.w3.org/1999/xlink">
    <topic id="person">
        <baseName>
            <baseNameString>Person</baseNameString>
        </baseName>
    </topic>
    <topic id="standards-body">
        <baseName>
            <baseNameString>Standards body</baseNameString>
        </baseName>
    </topic>
    <topic id="standard">
        <baseName>
            <baseNameString>Standard</baseNameString>
        </baseName>
    </topic>
</topicMap>
```

Figure 3 XTM 1.0 implementation of a Topic Map [42]

Figure 3 above demonstrates the XML-based XTM specification is used to construct an arbitrary topic map.

4. Content Packaging Standards

To allow for appropriate exportation of diagrammatic mappings, outputs of such a system would need to be machine-readable and self-contained with the necessary data that describes itself. Metadata provides a facility for this, as well as its preservation [11]. For containment, 3 file formats will be discussed, each that deal with concatenating multiple layers of information into a single preserved object.

4.1 Metadata Standards

4.1.1 Dublin Core Metadata Initiative (DCMI)

A formal ISO[3] standard (ISO 15836) [15], the Dublin Core is a set of properties used for describing digital resources and physical material [15]. The Dublin Core is defined in Resource Description Framework (RDF) semantics, a framework that enables organised metadata to be encoded, exchanged, and reused [23], for Linked Data [8], or XML, JSON, UML or relational databases for non-RDF data [8].

The Dublin Core standard in some instances is ambiguous for applications that require a high level of granularity [43]. The names of some elements can cause uncertainty about what information it should retain and how it should be used, according to a study that looked at the interpretation of the Dublin Core's metadata semantics by information professionals [30]

4.1.2 Encoded Archival Description (EAD) & Encoded Archival Context (EAC)

EAD is used for encoding descriptive information concerning archival information [32] and contains over 140 properties [36, 10] for describing information. It is also possible for some properties of EAD to be converted to other metadata structural standards, one being Dublin Core [37]. Its latest version is EAD3 and is defined as an XML schema or Document Type Definition (DTD) [36].

EAC pertains to the information relating to the creators of digital repositories or archived materials in a digital library [31]. It is defined in XML and can be used in conjunction with or independently of EAD and other metadata standards [31].

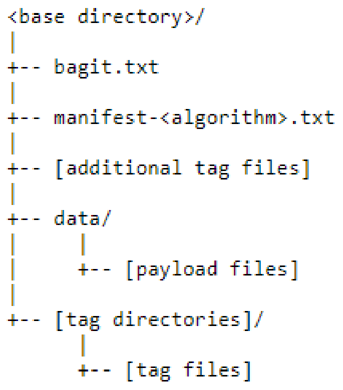
4.2 File Packaging Formats

4.2.1 Web ARChive (WARC)

WARC is a formal ISO standard (ISO 28500) [14], WARC is a file packaging format for storing any number of HTTP requests and responses with a catalogue of metadata for the corresponding messages and is optimized for long-term preservation [2].

4.2.2 BagIt

BagIt is a formal IETF[4] standard (RCF 8493) [18], BagIt is a set of hierarchical file layout specifications for the transfer and storage of digital objects [18].



```
<base directory>/
|
+-- bagit.txt
|
+-- manifest-<algorithm>.txt
|
+-- [additional tag files]
|
+-- data/
|     |
|     +-- [payload files]
|
+-- [tag directories]/
      |
      +-- [tag files]
```

Figure 4: A BagIt "Bag" Structure [18]

Figure 4 above depicts the structure of a Bag. The following is a description of the structure in the hierarchy in figure 4 [18]:

1. *<base directory>:* the root directory.

2. *bagit.txt:* the bag declaration, which identifies the bag version and encoding characteristics.

3. *manifest-<algorithm>.txt:* a list of checksums and the list of each corresponding file in the payload. *<algorithm>* denotes the checksum algorithm.

[3] International Organization for Standardization. https://www.iso.org/

[4] Internet Engineering Task Force. https://www.ietf.org/

4. *data/:* is the payload directory containing the digital objects of the bag.
5. [tag directories]/: a directory that contains tag files, where a tag holds metadata about its containing bag or files in the payload of the bag.

The BagIt system offers the benefit of assurance of completion of a transferred collection attributable to its inventory and file checksum features [24].

4.2.3 Self-contained Information Retention Format (SIRF)

SIRF is a storage container format designed for long-term retention [34]. It has recently become a formal ISO standard (ISO 23681) [15]. SIRF is self-contained, where both its data and metadata are stored as a single unit, self-described, and extensible, where additions are accounted for if necessary [34].

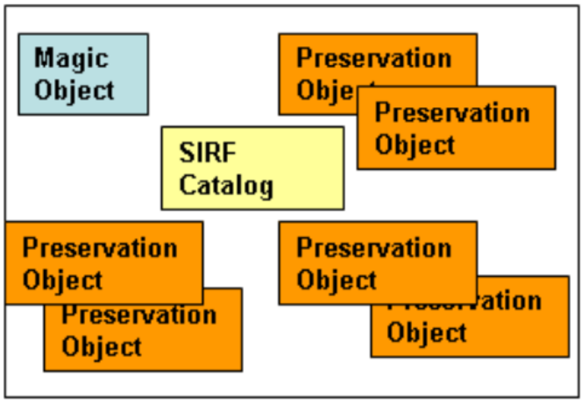


Figure 5. A SIRF container depicting its components [34]

Figure 5 above shows the definition of a SIRF container and its composition of components:

1. Magic Object: holds SIRF container identifier and version [34].
2. Preservation Object: immutable content to be preserved [34].
3. Catalog: contains metadata about the container and its preservation objects [34].

4.3 SIRF vs BagIt

SIRF is comparable to BagIt in the sense that both BagIt and SIRF attempt to preserve digital objects, however, BagIt aims to preserve a single object whereas SIRF aims to preserve a collection of objects [34].

5. Existing Tools for Complex Digital Object Visualization

5.1 Omeka

Omeka is an open-source content management system (CMS) for use by digital libraries. It offers exhibition creation, content organization and display [46]. Its open-source nature has facilitated the creation of plugins for its plugin architecture [46]. This provides a framework for the visualization of digital objects exhibited on Omeka. The metadata standard that Omeka utilizes is Dublin Core and Omeka can ingest any file format [33]. Omeka is a web-based solution and requires a web server and is often compared to WordPress[5] [33]. Omeka comes with a robust exhibit creation tool, which allows users to create exhibits with web pages that are composed of digital objects from the repository stored with the Omeka instance [29].

5.2 Collective Access

Collective Access is an open-source digital library management tool and digital exhibition creation software. [44] Collective Access is built on a web-based core called Providence [44]. Providence is an application that manages data modelling, a media framework that can manipulate and convert images, videos, audio, text and documents, an interface for cataloguing, and traversing collections [44]. Collective Access supports a wide array of metadata standards of which include EAD and Dublin Core [44]. Collective Access makes use of its system, like Providence, Pawtucket for the display of digital objects housed in Collective Access [45]. Pawtucket uses either default or user-defined display templates which format objects' metadata for viewing on-screen or as a PDF output [45].

5.3 Fedora Commons

Fedora Commons is an architecture that defines a framework for digital asset management and offers functionality for the management and discovery of digital assets at the expense of the actual implementation of certain features for the reason of high flexibility [4]. This means that it is possible to develop a feature with no constraints imposed by the framework that can allow users to view and create digital objects. Many repository management tools use Fedora as its backend; Islandora falls under this category.

5.4 Islandora

Islandora is a Fedora Commons backend, with Drupal[6] as the front-end solution [21]. Islandora extends the Fedora Commons file and metadata ingestion methods, meaning that Islandora accepts the same files as Fedora Commons [33]. The default metadata standard is the XML Schema definition of the

[5] WordPress. https://www.wordpress.org/
[6] Drupal. https://www.drupal.org/

Dublin Core [33]. Islandora offers Solution Packs, which are custom Drupal modules tailored to Islandora's functions for digital object management and display but none of the provided solution packs offers any additional diagrammatic ability to visualize digital objects [3].

6. Conclusions

The material presented above emphasizes key considerations when creating a tool that works with content from any digital library. Digital Archives are made up of complex digital objects that store multiple layers of data. Users may not always be aware of the content that resides on an archive other than what corresponds to their knowledge boundaries, hence this information must be easily discoverable through searching and browsing. When diagrammatically mapping digital objects and their relationships, the produced output should be of a format that can store metadata about its mappings and has utility when it results in something of a large scale. The resulting outputs should also be of a relevant mapping technique type, the lack of recent literature about topic maps and documentation of its language could indicate that such technology is no longer relevant, despite being an ISO standard. The analysis of content packaging standards posed some important considerations. The digital objects that make up digital libraries are defined by their metadata. This means that any application that uses a digital object should provide a facility for accessing this metadata. Secondly, resulting outputs should be archivable and accessible in the long term. WARC provides a good foundation for this as output formats are likely to be HTML web pages. Many of the tools reviewed offer viewing of digital objects in addition to the creation and management of them. However, none offer the precise features that are required such as manually mapping complex digital objects and exporting the resulting map as a self-containable object.

REFERENCES

[1] [n.d.]. The Five Hundred Year Archive. Retrieved May 25, 2021 from http: //www.apc.uct.ac.za/apc/research/projects/!ve-hundred-year-archive

[2] Alam, S. Web ARChive (WARC) File Format. City, 2018.

[3] Becker, S. Islandora (2019).

[4] Castagné, M. Institutional repository software comparison: Dspace, eprints, digital commons, islandora and hydra. University of British Columbia, 2013.

[5] CDL. University of California Digital Library: Glossary of Digital Library Terms. Retrieved from https://cdlib.org/resources/technologists/glossary-of-digital-library-terms/.

[6] Daley, B. J. Using concept maps in qualitative research (2004).

[7] Danks, M., Goodchild, M., Rodriguez-Echavarria, K., Arnold, D. B. and Griffiths, R. Interactive storytelling and gaming environments for museums: The interactive storytelling exhibition project. Springer, City, 2007.

[8] DCMI Usage Board. DCMI Metadata Terms. (2020). Retrieved from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[9] DELVE, J. and ANDERSON, D., 2014. *Preserving complex digital objects*. 1st ed. London: Facet Publishing, pp.10-11.

[10] Encoded Archival Description Tag Library - Version EAD3 (EAD Official Site, Library of Congress). (2021). Retrieved from https://www.loc.gov/ead/EAD3taglib/EAD3.html

[11] Good, J., 2010. A Gentle Introduction to Metadata. metaOER.

[12] Iso, B. S. and Standard, B. Information and documentation—The Dublin Core metadata element set. City, 2003.

[13] Iso, I. S. O. 28500: 2009 Information and documentation-WARC file format. International Organization for Standardization (2009).

[14] Iso. I. S. O 23681: Information technology — Self-contained Information Retention Format (SIRF) Specification (2019)

[15] Khare, R. and Rifkin, A. XML: A door to automated Web applications. IEEE Internet Computing, 1, 4 (1997), 78-87.

[16] Kozievitch, N. P. Complex objects in digital libraries. Bulletin of IEEE Technical Committee on Digital Libraries, 5, 3 (2009).

[17] Kunze, J., Littman, J., Madden, E., Scancella, J. and Adams, C. The bagIt file packaging format (v1. 0). Internet Engineering Task Force (2018)

[18] Larkin, J. H. and Simon, H. A. Why a diagram is (sometimes) worth ten thousand words. Cognitive science, 11, 1 (1987), 65-100.

[19] Le Grand, B. and Soto, M. Topic Maps, RDF Graphs, and Ontologies Visualization. Springer, City, 2006.

[20] Leggott, M. A. Islandora: a Drupal/Fedora Repository System. Georgia Institute of Technology, City, 2009.

[21] Lin, Y., Ahn, J. W., Brusilovsky, P., He, D. and Real, W. Imagesieve: Exploratory search of museum archives with named entity-based faceted browsing. Proceedings of the American Society for Information Science and Technology, 47, 1 (2010), 1-10.

[22] Miller, E. An introduction to the resource description framework. Bulletin of the American Society for Information Science and Technology, 25, 1 (1998), 15-19.

[23] Minor, D., Sutton, D., Kozbial, A., Westbrook, B., Burek, M. and Smorul, M. Chronopolis digital preservation network (2010)

[24] N. S. Vemuri, R. D. S. Torres, E. A. Fox, W. Fan and R. Shen, "Exploring digital libraries: integrating browsing, searching, and visualization," Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06), 2006, pp. 1-10, doi: 10.1145/1141753.1141755.

[25] Newcomb, S. R. and Biezunski, M. A Draft Reference Model for ISO 13250 Topic Maps. ISO/IEC JTC, 1 (2002).

[26] Novak, J. D. and Cañas, A. J. The theory underlying concept maps and how to construct them. Florida Institute for Human and Machine Cognition, 1 (2006).

[27] Novak, J. D., Gowin, D. B. and Bob, G. D. Learning how to learn. cambridge University press, 1984.

[28] Omeka. 2020. Omeka Classic: Exhibit Builder. Retrieved June 2, 2021 from https://omeka.org/classic/docs/Plugins/ExhibitBuilder/

[29] Park, J.-r. and Childress, E. Dublin Core metadata semantics: An analysis of the perspectives of information professionals. Journal of Information Science, 35, 6 (2009), 727-739.

[30] Pitti, D. V. Creator description: encoded archival context. Cataloging & classification quarterly, 38, 3-4 (2004), 201-226.

[31] Pitti, D. V. Encoded archival description: An introduction and overview (1999).

[32] Puckett, J. and Leslie, S. Omeka. Journal of the Medical Library Association: JMLA, 104, 4 (2016), 374.

[33] Rabinovici-Cohen, S., Baker, M. G., Cummings, R., Fineberg, S. and Marberg, J. Towards SIRF: Self-contained information retention format. City, 2011.

[34] Rueda, U., Arruarte, A. and Elorriaga, J. A. From scalable concept maps to scalable open student models. City, 2009.

[35] Stevenson, J. Encoded archival description tag library, version EAD3. Taylor & Francis, City, 2016.

[36] Stockting, B. Time to settle down? EAD encoding principles in the access to archives programme (A2A) and the research libraries group's best practice guidelines. Journal of Archival Organization, 2, 3 (2004), 7-24.

[37] Suleman, H. Investigating the effectiveness of client-side search/browse without a network connection. Springer, City, 2019.

[38] Suleman, H. Reflections on Design Principles for a Digital Repository in a Low Resource Environment (2019).

[39] TopicMaps.Org Authoring Group, XML Topic Maps (XTM) 1.0. (2001). Retrieved from https://topicmaps.org/standards/

[40] Tunkelang, D. Faceted search. Synthesis lectures on information concepts, retrieval, and services, 1, 1 (2009), 1-80.

[41] UC3M. 2010. Universidad Carlos III de Madrid OpenCourseWare. Topic Maps. Retrieved June 2, 2021 from http://ocw.uc3m.es/ingenieria-informatica/information-engineering/lecture-notes-1/

[42] Vogel, D. Qualified Dublin Core and the scholarly works application profile: a practical comparison. Library Philosophy and Practice (2014), 0_1

[43] Whirl-i-Gig. CollectivAccess Documentation Release 1.8. Retrieved from https://providence.readthedocs.io/en/latest/index.html

[44] Whirl-i-Gig. CollectiveAccess Pawtucket Documentation Release 1.8. Retrieved from https://pawtucket2.readthedocs.io/en/latest/.

[45] Zhu, Q., Gonçalves, M. A., Shen, R., Cassell, L. and Fox, E. A. Visual semantic modeling of digital libraries. Springer, City, 2003.