Information Visualization in Semantically Enhanced Digital Libraries for Ancient Manuscripts and Books

Srushti Goud[1,*], Arianna Magnani[1] and Salvatore Sorce[1]

[1]Kore University of Enna, Piazza dell'Università, 94100 Enna, Italy

Abstract

Cultural heritage knowledge sharing through manuscripts and books has been an important method of documentation as well as communication since the use of writing media such as papyrus, bamboo, wooden strips, parchment (prepared animal skin), and paper. The origin of handwritten texts may be traced back to the early 3rd millennium BCE, when cuneiform was developed in Mesopotamia and hieroglyphs were used in Egypt. These tangible records of important activities have survived and provided great insights into the life and cultures of the greatest civilizations that exist. The move from physical libraries hosting these records to digital repositories has been unimaginable and necessary given the lifespan of tangible heritage such as manuscripts and books.

The advancement of semantic digital libraries has greatly improved the preservation and accessibility of ancient manuscripts in the digital era. Nevertheless, successfully exploring, comprehending, and deciphering these abundant cultural and historical resources poses a significant obstacle due to the intricate semantic information they contain.

In this paper, we examine the capacity of information visualization methods to enhance the accessibility and comprehensibility of ancient manuscripts and books in semantic digital libraries. This is done through the analysis of five case studies of digital libraries for manuscripts and books to appraise the state of the art. In doing so, we hope to bring out the key features of the information visualization identified from the case studies which can be beneficial for researchers aiming to build new libraries for ancient textual heritage.

Keywords

Digital Library for Ancient Texts, Information Visualization, Ancient Manuscripts and Books, Semantic Digital Libraries

Introduction

The evolution of communication from the oral to the written form facilitated the establishment of enduring records, which served as the foundation for cultural heritage and enabled the accumulation of wisdom and knowledge. The advancement of writing media has not only affected the dissemination of information but has also shaped societal structures, educational systems, and artistic endeavors. As a result, private and public libraries, along with archives, have emerged with the purpose of safeguarding documents and democratizing their accessibility for consultation. Earliest evidence of writing systems dates back to the 3rd millennium BCE

Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage (AVICH 2024), June 4th, 2024, co-located with the 17th ACM Conference on Advanced Visual Interfaces, Arenzano (Genoa), Italy.

*Corresponding author.

srushti.goud@unikore.it (S. Goud); arianna.magnani@unikore.it (A. Magnani); salvatore.sorce@unikore.it (S. Sorce)

0000-0001-5066-8584 (S. Goud); 0009-0005-2565-6895 (A. Magnani); 0000-0003-1976-031X (S. Sorce)

with the development of the cuneiform in Mesopotamia, and hieroglyphs in Egypt, followed by the pictographic origins of Chinese characters in China. Ever since then written forms of communication have become tangible evidence that have existed and transferred knowledge of the previous generations to their future. Ancient texts have been accounted as cultural heritage since they provide a window to the past with records tracing their necessity into the exploration of bygone eras.

While physical conservation, maintenance and access of these tangible texts is quintessential and irreplaceable, the digital world demands a systematic method of knowledge sharing of this medium. Digital environments represent a novel approach to doing and disseminating research, effectively becoming the new "covers" of scientific objects and replacing paradigmatic representation of knowledge domains through tangible objects [1]. The evolution from physical to digital repositories has been a significant advancement in preserving these valuable artifacts, guaranteeing their durability and availability for future generations. Nevertheless, the transition has also presented intricate challenges in examining, understanding, and deciphering the abundant semantic substance that these historical texts encompass. In this work, we present a study which examines how information visualization approaches might improve the accessibility and comprehensibility of historical manuscripts and books in semantic digital libraries. It offers a detailed analysis of five case studies of digital libraries that can help provide insights into the conceptualization and final outcome of these cases.

We undertook this study as a starting point for the project "M.A.R.E.: Manuscripts and books from Asia Reaching Europe. A semantically enhanced digital library mapping the Asian books circulation along the Silk Maritime Routes". The project aims to investigate a diverse array of texts in classical and oriental languages such as Chinese, Arabic, and Syriac, focusing on selected manuscripts and books from 700 to 1700. Its aim is also to map the trajectories through which these materials traversed to Europe, unveiling insights into the cultural exchanges and commercial networks along the Maritime Silk Routes connecting Asia and Europe. A pivotal challenge facing the project pertains to dealing with texts composed in varied oriental languages, often accessible solely to experts within specific disciplinary realms. The overarching goal is to render these materials accessible to interdisciplinary scholarly inquiry, thereby fostering innovative research perspectives. Hence, the present investigation into the current landscape of Data Visualization systems and visual interfaces for digital repositories of manuscripts and books stands as a foundational pillar of the M.A.R.E. project.

In the following section, we discuss the literature review followed by section 2, which details the methodology to select and analyze the cases. In section 3, we describe the five cases, followed by a discussion in section 4 highlighting key features for effective information visualization and an initial proposal for the library to be created within the M.A.R.E. project. In section 5, we provide preliminary conclusions and the way forward.

1. Literature Review

1.1. Transition from Traditional Libraries to Digital Libraries and Databases

The advent of the digital revolution has led to a substantial change in the conservation and availability of cultural knowledge, since physical libraries have been replaced by digital repositories.

The change has been prompted by the imperative to preserve fragile records from deterioration and to guarantee their accessibility to a global audience. Digital libraries offer unprecedented opportunities for researchers and the general public to effortlessly access and engage with historical records and books, overcoming the constraints of location and time [2]. However, this transition also presents challenges, including the need for technological infrastructure, expertise in digital abilities, and strategies to replicate the complex experience of engaging with physical books [3] [4] .

1.2. Semantic Digital Libraries

Semantic digital libraries represent an evolution in digital repository systems, using Semantic Web technologies to enhance information discovery and usage. The term 'semantic' in this context refers to the use of structured, interconnected data that enables more intuitive and effective organization of content. By implementing technologies such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL), these libraries not only store information but also understand and express the relationships between various pieces of content [5].

RDF is a standard model for data interchange on the web, allowing the libraries to represent information in a structured and machine-readable format. It uses triples, a simple data structure consisting of subject, predicate, and object, to make assertions about resources in a way that enables the integration of various data sources with ease. On the other hand, OWL is a more complex web ontology language that provides richer integration and interoperability between different data sets. It extends the RDF framework by adding additional vocabulary along with a formal semantics. OWL is used to explicitly represent the meaning of terms in vocabularies and the relationships between them.

Together, RDF and OWL form the backbone of semantic digital libraries, facilitating the creation of a well-structured and interconnected data ecosystem. This structure not only enhances the searchability of digital libraries but also enables sophisticated services such as context-aware recommendations and semantic querying, which significantly improve user interaction with digital content. For example, a semantic digital library could use these technologies to link various works by the same author, related historical documents, or academic papers citing similar studies, thereby providing a richer and more interconnected user experience.

The integration of semantic web technologies signifies a shift towards more advanced systems that are designed to meet the complex information needs of users, offering a more intuitive and enriched browsing and discovery experience in digital libraries[6].

1.3. Information Visualization

Information visualization plays a vital role in enhancing the accessibility and comprehension of complex data sets, particularly those found in digital libraries. Visualization techniques provide the capacity to transform intangible, semantic data into concrete, interactive encounters, especially in the domain of cultural heritage and ancient manuscripts and books [7]. Information visualization improves user engagement and understanding by using interactive maps to track the origins and dissemination of manuscripts, timelines to depict historical contexts, and network graphs to highlight connections between texts, authors, and subjects. These tools not only assist users in navigating through large digital collections but also uncover hidden patterns and relationships, thereby fostering the development of new intellectual concepts [8].

This literature review offered an overview of the importance of digital libraries and the innovative techniques of information visualization that improve the accessibility and interactivity of ancient writings and books. The following sections will discuss the research methods, provide detailed case studies of digital libraries that employ these methodologies, and analyze the consequences and future possibilities in this field.

2. Methodology

While numerous digital library and database projects exist today, encompassing both ongoing initiatives and established platforms, we focus on five case studies to illustrate the current landscape of digital preservation of cultural heritage. These case studies were selected to showcase the diverse approaches and methodologies employed in digitally conserving cultural artifacts. In particular, the projects were chosen based on their efficacy for information visualization, evaluated against the following criteria:

1. Projects implemented in the past 10 years and have the database available online for exploration were selected.
2. The libraries that visualize the texts on a graphical map were selected.
3. Projects having publications, descriptions to help understand its conceptual framework and development were chosen.
4. Projects with different types of textual medium like manuscripts, books, papyrus scrolls were selected to understand the requirements for different media.

2.1. Analytical Framework

To thoroughly evaluate the effectiveness of different semantic digital libraries, our analysis of each case study was structured around three primary aspects.

The Database's Background and Collection Focus: This was an examination of the historical and contextual background of each database, focusing on its foundational objectives and the specific type of content it aims to manage and deliver. For example, whether the database specializes in scholarly articles, historical documents, or multimedia resources. Understanding the collection focus helps in assessing how well the digital library meets its intended goals and serves its target audience.

The Semantic Technologies and Information Visualization Methods Used: This delved into the specific technologies and methods implemented by the digital libraries to manage, structure, and display content. Semantic technologies might include the use of RDF or OWL to create a structured, searchable data framework. Information visualization methods could involve interactive graphs, heatmaps, or dynamic querying interfaces that aid users in navigating complex information sets. This analysis helped to highlight the innovative aspects of each library and their effectiveness in enhancing data discoverability and usability.

Table 1
Case Studies of Digital Libraries

S.No.	Digital Library	Year of Library Implementation	Contents of the Library	Website
1.	Mapping Manuscript Migrations	2022	History and provenance of Western European medieval and early modern manuscripts.	https://mappingmanuscriptmigrations.org/en
2.	China Historical Christian Database	2022	Christianity in China: 16th – 20th century CE.	https://data.chcdatabase.com/
3.	Serica	2019	Texts, images and musical documents from the 2nd century. B.C. until the 19th and 20th centuries. A.D. concerning the Central Asian routes between China and Europe.	https://serica.unipi.it/
4.	Beta maṣāḥəft	2016	Christian manuscript tradition of the Ethiopian and Eritrean Highlands.	https://betamasaheft.eu/
5.	Reconstructing the Phillipps Manuscript Collection	2016	Manuscript collection of Sir Thomas Phillipps containing medieval and early modern manuscripts and documents.	http://personal-research-domain-burrows.nodegoat.net/viewer.p/32/895/scenario/1/geo/

How These Methods Enhance User Interaction, Comprehension, and Accessibility: The final assessment focused on the practical impact of the employed technologies and methods on the end user. It assessed how these tools improve user interaction with the library system, enhance comprehension of the content, and make information more accessible to diverse audiences. For example, it examines whether semantic tagging and advanced visualization tools lead to a more intuitive search experience or if they facilitate better understanding of complex datasets.

Out of the projects initially reviewed based on these criteria, we selected five projects for a detailed analysis.

3. Case Studies

Here are the five case studies considered based on the criteria mentioned above. Table 1 lists the five cases and their the year of its implementation, the contents of the library and the website for each case. The background and collection focus of the five projects can be summarized as follows:

1. **Mapping Manuscript Migrations (MMM)**: The MMM project integrates and studies premodern manuscript databases using Semantic Web technologies. It aggregates data from over 220,000 manuscripts from prominent databases such as the Schoenberg Database

of Manuscripts, Medieval Manuscripts in Oxford Libraries, and Bibale. This project aims to enhance manuscript studies by providing a collaborative platform for enriching and publishing content in a unified knowledge graph. The focus is on manuscripts' history, provenance, and the intellectual and cultural heritage they represent [9].

2. **China Historical Christian Database (CHCD)**: CHCD is a tool to study Christianity in China, focusing on creating a comprehensive geographic and relational database. While not a library in the conventional sense, it compiles data from various sources to analyze the Christian presence in China between the 16th and 20th centuries, including people, institutions, corporate entities, and events [6].
3. **Serica**: Serica is a digital archive on the Silk Road that represents an initiative to bridge East and West through historical documents that narrate the stories, cultures, goods, and individuals that traversed these routes [10].
4. **Beta maṣāḥəft**: Beta maṣāḥəft focuses on creating a comprehensive portal for the manuscript tradition of the Ethiopian and Eritrean Highlands. This project aims to encode and semantically relate descriptions of manuscripts, works, places, and persons involved in these traditions. Initially, there has been no comprehensive prosopography, gazetteer, or clavis for the Ethiopian tradition, with descriptions of manuscripts being scattered across various catalogues. The project seeks to fill this gap by structuring and producing information about primary and secondary sources for the study of the Ethiopic literary tradition and manuscript culture [11].
5. **Reconstructing the Phillipps Manuscript Collection (PHILLIPPS)**: The PHILLIPPS project focused on the manuscript collection of Sir Thomas Phillipps (1792-1872), which consisted of at least 40,000 medieval and early modern documents and manuscripts. This collection, now dispersed, is recognized for its crucial role in preserving a significant portion of Europe's cultural heritage. The primary goal was to aggregate and make accessible the data relating to the history of these manuscripts, facilitating their analysis and visualization [12].

4. Discussion

4.1. Key Features of Effective Information Visualization

In analysing the case studies and using their project repositories to the extent that they were available online, we identified the following as key features:

- **Semantic Enabled Search:**
  We examined repositories to identify the integration of user-friendly search features, such as semantic enabled searches that analyze search intent by considering both the words and context of the query. For instance, the use of semantic search in MMM allows users to effortlessly filter and locate relevant data. Employing Linked Open Data (LOD) services and SPARQL (SPARQL Protocol and RDF Query Language) [9] endpoints to improve data accessibility, as demonstrated in MMM, thereby increases the discoverability and accessibility of data to a broader audience, including individuals without technical proficiency. Ensure the platform is compatible with several devices and screen sizes, with a focus on optimizing user experience to support a wide range of accessibility requirements.

- **Dissemination of Ontology Infrastructure to achieve better Interdisciplinary Collaboration:**
  We analyzed the extent to which these projects engaged specialists from various disciplines to enhance content and ensure a comprehensive perspective on data. Projects MMM and Beta maṣāḥəft promoted a cooperative atmosphere for the generation and management of data, as there was clear evidence, in their reporting about engagement of specialists from other disciplines to enhance the content and to guarantee a comprehensive outlook on data. This was also seen as promoting the dissemination of ontology infrastructure, such as in the case of MMM and PHILLIPPS. This simplified the integration and interpretation of data across multiple disciplines. This could also enable comprehensive and intricate studies that encompass several fields of research.
- **Pre-determining and Beta-testing with Target Audience:**
  We looked into how beta-testing with the intended audience helped refine the features and functionalities of these digital libraries. It emerged that clearly defining the target audience for the digital library, as exemplified by Serica's focus on Silk Road enthusiasts and scholars was a necessary step. This allows for the customization of content and features to cater to their individual requirements and interests. Creation of the platform and its features needs to be done with the intended audience as the primary focus. Providing customized visualization tools and data presentation methods that specifically address the anticipated degree of proficiency and engagement improves the relevance of the project for the target audience. Beta-testing the early iterations of the project with the target audience can be greatly beneficial to refine the final product.
- **Ease of Comprehension:**
  Our analysis included examining the inclusion of informative materials and interactive tools within platforms. Once again this was achieved by going through the reported literature and also using the platforms directly, to the extent that they were available for use online. Inclusion of informative materials, instructions, and interactive prompts within platforms to assist users in exploring and comprehending the data, as demonstrated in Beta maṣāḥəft, improves overall understanding without causing users to feel overwhelmed. Utilization of easy visualization tools, such as interactive maps and charts in MMM and CHCD, to effectively display intricate data, enable people to comprehend complex linkages and understand historical context in an engaging manner.
- **Adaptive User Interface:**
  We reviewed how projects like CHCD design their user interfaces to be adaptable, allowing users to modify settings and filters according to their research interests or personal preferences. It is advisable to provide functionalities that cater to many types of user interactions, ranging from casual browsing to in-depth study. These features should provide varying levels of involvement based on the user's objectives and expertise with the subject matter. Make sure that the interface is capable of accommodating a variety of use cases and research queries, allowing for the opportunities to broaden or narrow down the scope of data investigation. This should represent the different interests of the target audience.
- **UI Tools for Complex inquiries:**

An adaptive user interface can sometimes include the use of complex tools for the UI. Naturally, our analysis aimed to understand how these tools facilitate a customized and interactive study of the content when catering to a broad spectrum of user requirements. Serica and CHCD projects prioritized the development of user-friendly interfaces that facilitate both general and specific, or complex inquiries. The use of temporal filters, thematic searches, and comparative analytical tools enables a more interactive and customized study of the content, accommodating a wide range of user requirements. The interface of CHCD is especially beneficial for researchers.

- **Multilingual Support:**
  In order to cater to the worldwide interest in cultural heritage, CHCD provides its platform in both English and Chinese languages, thereby expanding its reach and involving a larger audience. This approach highlights the significance of having multilingual support in digital libraries to cater to multinational research communities and the general public.

4.2. Implications for the Digital Library Development Proposal in the M.A.R.E. project

We proceed to share a list of lessons learnt that can be applied to the M.A.R.E. project or other projects of a similar nature:

- The information visualization should include a graphical map as seen in all of the case studies, highlighting the travel route of the texts along with description of the person/actor who was responsible (if available). This should be the starting point since the project focuses on the journey of each of these texts with respect to their origin and destination. The graphical map should depict the original position of the texts and while clicked upon show specifics of each of the texts as seen in cases CHCD and PHILLIPPS.
- Each text should contain the digitized version of the content semantically organized by filters like origin, destination, language, actor, year, type etc. This should be supplemented by an OCR (Optical Character Recognition) of the text in case interested people want to know more and read the contents of the text.
- Another feature that can be included is a timeline slider that provides an overview of the number of books and manuscripts positions by the year as seen in PHILLIPPS. This will provide dynamic information of the texts and the travel path they undertook before arriving in Europe.

5. Conclusions

In this study, we investigated the use of semantic technologies and information visualization approaches in digital libraries, with a specific focus on ancient manuscripts and books. The study demonstrated the transformative capacity of these technologies in improving accessibility, comprehension, and user interaction with digital repositories through the examination of five case studies. The main discoveries emphasized the significance of search features that are easy for users to navigate, the incorporation of Linked Open Data (LOD) to enhance the accessibility of data, and the use of interactive visualizations to uncover intricate connections within the data. Moreover, the study emphasized the importance of interdisciplinary collaboration and the creation of customized platforms for targeted audiences to ensure the efficient distribution and investigation of cultural heritage content. We hope to implement the key features uncovered by our exploration in the development of the library envisioned for the M.A.R.E. project.

Acknowledgements

This research is part of the PRIN (Projects of Relevant National Interest) 2022 PXTYCL "M.A.R.E.: Manuscripts and Books from Asia Reaching Europe. A Semantically Enhanced Digital Library Mapping Asian Books Circulation Along the Silk Maritime Routes," developed by the Kore University of Enna, (in collaboration) with the University of Pisa and the University of Salerno.

References

[1] C. Claire, G. V. R. E. Allen, Introduction : Ancient Manuscripts and Virtual, Classics@ 18 (2021). URL: https://classics-at.chs.harvard.edu/classics18-introduction/.

[2] Z. Manžuch, Ethical Issues In Digitization Of Cultural Heritage, Journal of Contemporary Archival Studies 4 (2017) 4.

[3] A. Baruzzo, P. Casoto, P. Challapalli, A. Dattolo, N. Pudota, C. Tasso, Toward semantic digital libraries: Exploiting Web2.0 and semantic services in cultural heritage, Journal of Digital Information 10 (2009).

[4] L. Candela, D. Castelli, P. Pagano, History, evolution, and impact of digital libraries, January, 2010. doi:10.4018/978-1-60960-031-0.ch001.

[5] P. Robinson, Creating and Implementing an Ontology of Texts , Documents and Works (2020).

[6] A. Mayfield, D. Ireland, E. Menegon, THE CHINA HISTORICAL CHRISTIAN DATABASE, Technical Report, Boston University, 2022.

[7] R. Adams, J. Glomski, Seventeenth-Century Libraries: Problems and Perspectives, Brill, Leiden | Boston, 2023.

[8] F. Kboubi, Semantic Visualization and Navigation in Textual Corpus, International Journal of Information Sciences and Techniques 2 (2012) 53–63. doi:10.5121/ijist.2012.2105.

[9] E. Hyvönen, E. Ikkala, M. Koho, J. Tuominen, T. Burrows, L. Ransom, H. Wijsman, Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12922 LNCS (2021) 615–630. doi:10.1007/978-3-030-88361-4_36.

[10] U. of Pisa, Serica, 2019. URL: https://serica.unipi.it/.

[11] P. M. Liuzzo, Digital Approaches to Ethiopian and Eritrean Studies, 2019. doi:10.2307/j.ctvrnfr3q.

[12] T. Burrows, Reconstructing the Phillipps Collection: Toby Burrows | News from an EU project aimed at reconstructing the manuscript collection of Sir Thomas Phillipps, 2016. URL: https://tobyburrows.wordpress.com/.