International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I12P110

Volume 72 Issue 12, 83-88, December 2024

Review Article

Metadata Harvesting in the Digital Humanities: A Case Study of The Ohio Digital Library

Flavia Ann Albert Saldanha

Independent Researcher, Cincinnati, Ohio.

Corresponding Author : flavia.a.saldanha@gmail.com

Received: 31 October 2024 Revised: 26 November 2024 Accepted: 11 December 2024 Published: 30 December 2024

Abstract - *This review and research investigates the practical application of metadata harvesting at the Ohio Digital Library (ODL). It explores how standardized metadata collection and consolidation methods contribute to advancing digital humanities research. The study delves into the technical aspects of metadata harvesting, including the protocols used to extract information from diverse digital repositories. It examines the strategies to ensure data consistency and compatibility across different formats and sources. The research addresses the challenges encountered when integrating harvested metadata into the ODL's infrastructure. By analyzing the ODL's experiences, this paper aims to highlight the significance of metadata harvesting in enhancing the discoverability and accessibility of digital resources. It provides insights into metadata harvesting practices' potential benefits and limitations, offering valuable lessons for other digital libraries and cultural heritage institutions.*

Keywords - *Application programming interface, Digital library, Digital humanities, DISTRIBUTED processing, metadata.*

1. Introduction

We are in the era of digital information, where efficient discovery and access to digital resources are vital for the larger and more efficient success of digital innovations. Digital libraries across the globe have emerged as essential tools in this transformation, providing platforms for organizing, storing, and sharing digital resources. However, digital libraries' effectiveness depends on robust metadata practices that ensure seamless discoverability and accessibility of resources across diverse repositories. Metadata harvesting is a key fundamental process for modern digital library systems, enabling efficient search and access capabilities to its vast digital resources. Efficient metadata management practices can improve discoverability, better data quality, and faster access to insights, thus helping reduce overall costs and better user experience. This review and research shall analyze the practical application of metadata harvesting performed at The Ohio Digital Library (ODL), a well-known part of the State Library of Ohio that performs harvesting to supplement digital artifacts to the national collection in the Digital Public Library of America (DPLA). The review study identifies technology and process gaps in the existing metadata harvesting techniques utilized by ODL, including inconsistent metadata standards, data redundancy, and integration complexities, if any, hamper resource discoverability and usability. By exploring standardized metadata collection and consolidation methods, the research aims to bridge these gaps and provide solutions or recommendations to improve metadata interoperability and scalability.

2. Background and Context

2.1. Metadata Harvesting

Metadata harvesting is an automated process of collecting information about digital resources and organizing them with labels to systemically improve the searchability, accessibility and management of digital information across different systems and platforms. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a widely used method for metadata harvesting that uses XML over HTTP to exchange data on the World Wide Web. Using OAI-PMH,[1] the data providers expose the structured metadata, and the service providers request to harvest that metadata via OAI-PMH extractors. This collected data is then stored and managed on the search interface of the digital libraries and accessed by the requested repositories.

2.1.1. Benefits

The key advantage of metadata harvesting using techniques like OAI-PMH includes enhanced discoverability, allowing users to locate resources across various platforms with minimal effort efficiently. By providing a centralized approach to metadata, the process streamlines the search experience, ensuring that relevant information is accessible regardless of the originating system. OAI-PMH enables interoperability and scalability, allowing data providers and service providers to communicate and exchange metadata regardless of their systems. Furthermore, OAI-PMH is open and inclusive, promoting the visibility and accessibility of digital resources. This allows data providers to expose their

metadata to a range of service providers and enables them to offer diverse services to users. Interoperability plays a crucial role in bridging the gaps between diverse systems. It ensures that heterogeneous platforms can work together seamlessly, creating a cohesive environment where data and functionality are shared effectively. The optimization of resources is another significant benefit, as it reduces the duplication of records. By centralizing metadata, systems avoid redundancy, saving time and effort while maintaining an organized and efficient data repository. Metadata management can provide a wide range of benefits to the institution or organization adopting standardization and governance around its metadata practices.

Management metadata [2] describes concepts, relationships and rules in management, such as staff roles, job responsibilities and management processes. The implementation of metadata management can provide an overall view of enterprise data and how to use them. It also guarantees the quality of data, expressly, completeness, consistency and accuracy. Since the development history of software products and tools is diverse, metadata in each data supply chain process stage may not be exchanged efficiently when missing a unified metadata model (metamodel). Metamodel is a conceptual model of metadata which offers a detailed description of metadata units and their relationship. Without the appropriate metadata management tools and practices, over 80% of a researcher's time is spent searching for and preparing data. A metadata management system can accomplish these tasks in seconds rather than hours.

2.1.2. Challenges

Despite its benefits, metadata harvesting faces challenges, such as inconsistent metadata standards, incomplete records, and scalability issues when handling large datasets. For instance, OAI-PMH only supports the data about the digital objects, i.e. the metadata and not the digital object itself.

It does not guarantee the consistency or completeness of the metadata capture across different data providers and could compromise the extraction quality as it depends on the stability and performance of the network and systems involved in the process. Additionally, OAI-PMH is static as it does not support any feedback or interaction between data providers and service providers, nor does it allow service providers to query or filter the metadata. As a result, service providers must monitor and refresh their metadata collections on their schedules.

2.2. The Ohio Digital Library Infrastructure

The Ohio Digital Library [3] serves as a central hub for digital collections across Ohio's academic and public libraries, managing millions of digital objects that span a diverse range of materials. Its robust digital infrastructure accommodates various content types, including historical manuscripts and photographs, oral histories, digital books and journals, cultural heritage artifacts, and archaeological data. These collections highlight the region's rich history, culture, and scholarly resources, making them accessible to a wide audience.

2.3. Digital Humanities

Digital humanities are an interdisciplinary field that integrates traditional humanities disciplines—such as literature, history, philosophy, and cultural studies—with digital technologies to analyze, interpret, and preserve human culture and historical artifacts. This approach utilizes computational tools and methods, enabling innovative ways to engage with large datasets, visualizations, and interactive models to understand complex patterns in human culture and society.

The Ohio Digital Library is pivotal in advancing digital humanities by providing access to a rich and diverse digital collection. This includes historical manuscripts, rare books, maps, photographs, oral histories, and multimedia resources. These materials span a variety of subjects, including regional history, genealogy, social movements, and artistic endeavors, offering scholars and students invaluable resources for research and education.

For example, textual collections are often digitized and made searchable, allowing for computational textual analysis techniques like [3] word frequency studies, sentiment analysis, and network mapping of historical relationships.

Visual and multimedia collections contribute to geospatial studies, virtual reconstructions, and cultural storytelling through digital exhibits. By integrating technology with humanities scholarship, the Ohio Digital Library exemplifies the transformative power of digital humanities, bridging the gap between traditional humanities scholarship and the dynamic possibilities of the digital age.

3. Methodology

In this case study, we will analyze and explore the metadata issues in the metadata collection of oral histories submitted to the service provider ODL [4], in this case, for making the metadata available in the DPLA for the digital humanities category. The aim is to provide a detailed overview of the process, identify challenges, propose improvements, and leverage advanced tools and techniques to enhance metadata quality and consistency.

For the investigative analysis, the metadata from DataCite.org was referenced using OAI-PMH extractors with the following base URL: https://oai.datacite.org/oai

These endpoints allow us to perform various OAI-PMH operations, such as retrieving metadata records, identifying available metadata formats, and listing identifiers. The Key OAI-PMH Extraction Strategies are explained below:

3.1. Technical Metadata Extraction Methodology

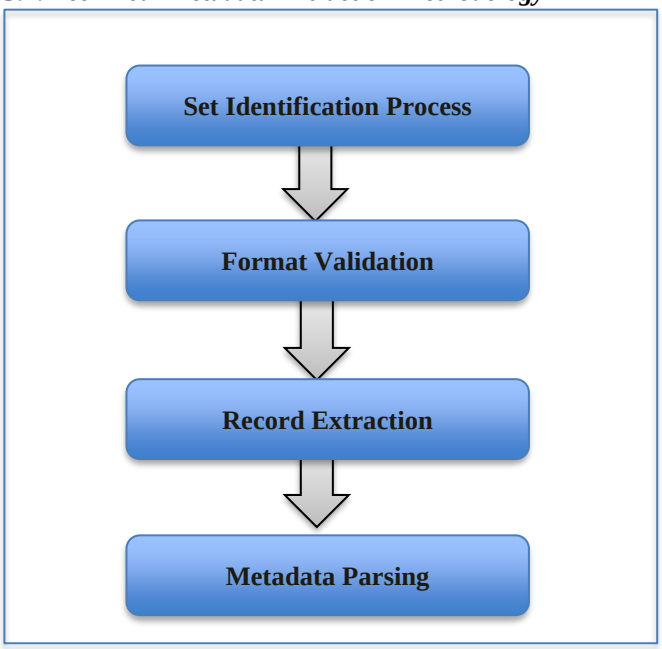

Fig. 1 Harvesting process workflow

3.1.1. Harvesting Process Workflow

- Set Identification: Determine relevant metadata sets
- Format Validation: Confirm metadata format compatibility
- Record Extraction: Retrieve XML-formatted metadata records
- Metadata Parsing: Extract and standardize metadata elements

Each step in the workflow was performed to review and analyze the metadata elements, as illustrated in the next section.

3.2. Set Identification

- Endpoint: https://oai.datacite.org/oai?verb=ListSets
- Purpose: Retrieve available metadata sets for targeted harvesting
- Functionality: Allows filtering and segmentation of metadata records based on predefined sets

To retrieve a list of available sets within the DataCite repository, which can be used to filter records during harvesting, one can use the ListSets verb: https://oai.datacite.org/oai?verb=ListSets

3.3. Metadata Format Discovery

- Endpoint: https://oai.datacite.org/oai?verb=ListMetadataFormats
- Purpose: Identify supported metadata format standards
- Significance: Ensures compatibility and standardization of metadata records

For example, to list metadata formats from the DataCite service, you can use: https://oai.datacite.org/oai?verb=ListMetadataFormats

This request will return a list of all metadata formats from the DataCite repository, as shown in Figure 3.

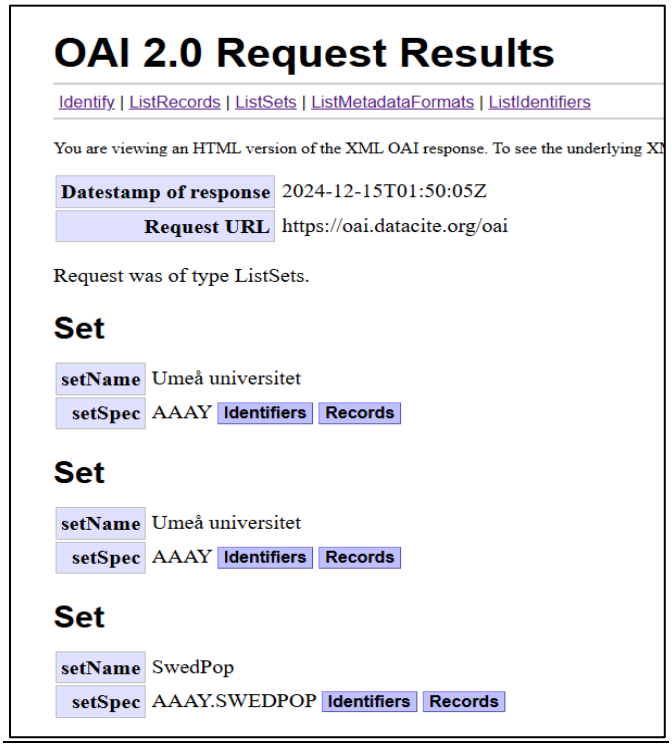

Fig. 2 ListSet extracted for review

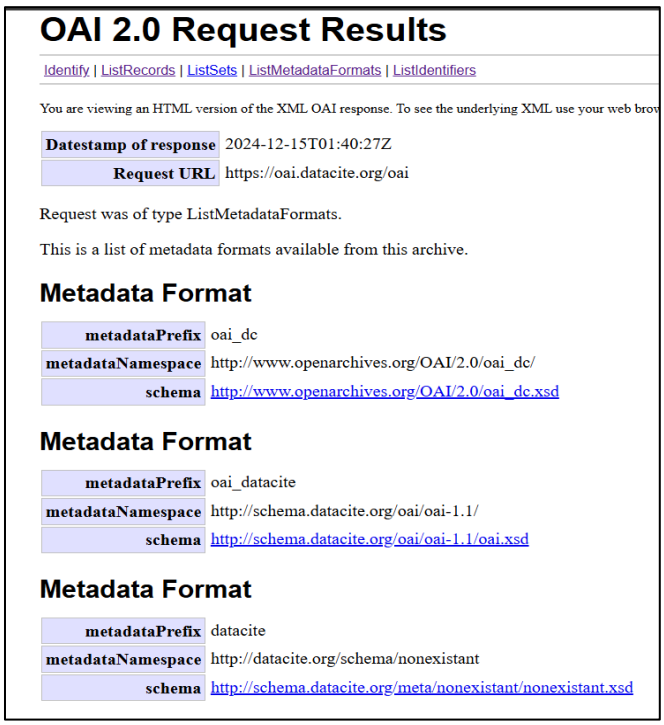

Fig. 3 ListMetadata formats extracted

To summarize the Metadata Extraction Specifications extracted in this analysis,

Data Capture Parameters

Source Platform: DataCite.org
Extraction Method: OAI-PMH Protocol
Metadata Format: Dublin Core
Key Captured Fields: Title, Author/Creator, Publisher, Date, Resource Identifier

Utilizing these OAI-PMH services, the sample metadata to be analyzed was extracted in XML format, including fields like title, author or creator, publisher, date, resource identifier, etc., in the Dublin Core Metadata format as prescribed by the ODL. At ODN, the recommendation is to use the Rights Statements over Creative Commons licenses for most items. In most cases, the library providing the resource is not the creator of the resource and cannot provide a license for the material in the same way a creator is.

The metadata harvesting exercise conducted using the OAI-PMH protocol successfully demonstrated the process of extracting, analyzing, and validating metadata for oral history records. By leveraging the DataCite OAI-PMH [6],[7] endpoint (https://oai.datacite.org/oai), the metadata was systematically retrieved using key verbs like ListMetadataFormats and ListSets to identify available metadata formats and subsets of records. Through the step-by-step approach, critical metadata gaps such as missing fields, inconsistent formats, and redundant entries in the metadata will be analyzed. A core example of an oral history interview was analyzed in-depth to highlight these challenges, and targeted solutions were implemented, including automated validation, standardized templates, and metadata enrichment using NLP techniques.

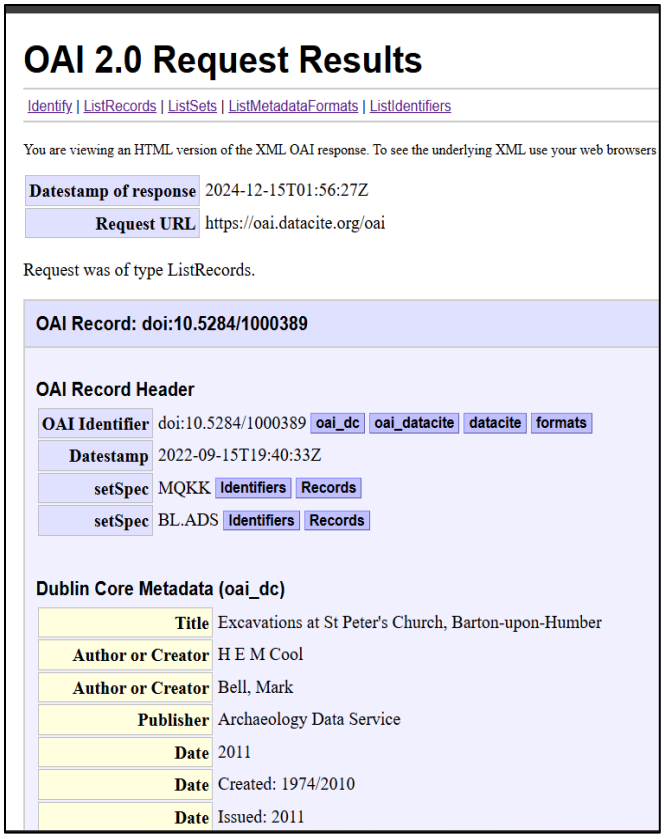

Fig. 4 ListRecords extracted

The findings emphasize the importance of real-time validation and structured metadata management to ensure consistency, completeness, and usability of harvested data. Once the metadata was harvested, the next step was to conduct an initial review to identify any obvious issues with the data. This involved a manual check of the metadata for completeness and consistency.

Table 1. Metadata findings and review

Metadata field	Example Data	Potential Gap	OAI-PMH Issue
Title	Excavations at St Peter's Church, Barton-upon-Humber	Missing keywords in the title (e.g., "Digital Humanities")	Missing or inconsistent themes limit the ability to categorize or search data.
Author Or Creator	H E M Cool Bell, Mark	misspelling of the first name, inconsistent name format (e.g., "H E M Cool." vs. "Hem, Cool")	Inconsistent name formatting may prevent accurate aggregation
Publisher	Archaeology Data Service	None	OK
Date	Created: 1974/2010	Format inconsistencies (e.g., 1974/2010 vs. 01-01-1974: 01-01-2010)	Date format inconsistencies may prevent proper sorting and searching.
Resource Type	Dataset	None	OK
Format	text/csv	None	OK

4. Results and Discussion

In the metadata extraction process for a collection of digital humanities histories, several key issues were identified, affecting the overall quality and accessibility of the data. Below are the major problems observed during the analysis:

4.1. Missing Key Information

20% of the metadata records were missing key attributes such as the "Title" and "date". The missing fields hindered the ability to index and search the histories accurately. It also made it difficult to cite and contextualize published articles properly. In one scenario - a record for an interview only included a vague title like "Oral History Interview," without specifying the interviewee's name or the interview date. This creates ambiguity and limits searchability.

4.2. Inconsistent Formats

The metadata for audio files showed inconsistency in how file formats were labeled. Some records used terms like "MP3," while others used "audio/MPEG". This inconsistency could confuse systems that rely on standardized file format labels for organizing and retrieving media files. It also complicates processes like media conversion and archival management.

4.3. Redundant Entries

Duplicate metadata entries [8] for the same published articles were observed due to overlapping harvesting schedules from multiple repositories. Redundant records can lead to confusion, inaccurate reporting, and inefficiencies in storage. Users might encounter the same article or collection multiple times, leading to redundant transcription, review, and analysis efforts.

4.4. Proposed Improvements

Several strategies can be implemented to address the metadata issues identified in the case study. These solutions aim to enhance metadata quality, standardize practices, and automate the validation and correction process.

4.5. Real-Time Validation

Implement automated validation scripts to flag missing or inconsistent fields during metadata harvesting. These scripts can run as part of the metadata harvesting process to identify and address real-time issues, ensuring that incomplete or incorrectly formatted records are not added to the database. Custom Python Scripts [9], can be written to check for missing fields, such as "author name" or "recording date," and generate reports on the frequency and type of missing data. It can validate and standardize metadata during harvesting, flagging incomplete or inconsistent records.

4.6. Standardized Templates

Utilizing standardized templates can enforce consistency across all records. By adopting a uniform structure for metadata (e.g., Dublin Core, MODS), [10] repositories can ensure that all required fields are included and that data is formatted consistently. The good news is that ODL has already adopted standardized templates and checks for two required submission fields. However, there is an opportunity to accommodate additional attributes in the template.

4.7. Metadata Enrichment

Machine learning techniques, such as Natural Language Processing (NLP) [11],[12], can be employed to enrich metadata by filling in missing fields based on contextual analysis. For example, NLP models can extract author names, locations, and key attributes from the interview transcripts or audio files. Pre-trained models such as BERT or GPT can be fine-tuned to extract metadata from oral history transcripts or audio recordings. Audio Processing Tools [13],[14], such as speech-to-text, can be used to transcribe audio interviews, and then NLP techniques can be applied to extract additional metadata.

5. Conclusion

This study has explored the metadata harvesting practices of the Ohio Digital Library (ODL), providing insights into its methodologies, challenges, and contributions to digital humanities research. The analysis reveals that the ODL's adoption of standardized metadata schemas and protocols, such as Dublin Core, MODS, and OAI-PMH, is central to ensuring interoperability and efficient data exchange. By adhering to these standards, [15] the ODL enables seamless integration and accessibility of its extensive collections, serving as a model for other digital libraries. The investigative analysis of metadata harvesting for the digital humanities collection revealed quite a few issues with missing fields, inconsistent formats, and redundant entries. By implementing real-time validation, standardizing metadata templates, and enriching the data using NLP techniques, these problems can be addressed. This case study serves as an example of how metadata quality can be improved to enhance the usability, searchability, and accessibility of oral history collections.

References

[1] Harsh Aijt Khajgiwale et al., "A Study on Harvester for Oai-Pmh Compliant Institutional Repositories for Academic Institutions," *International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 63-66, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[2] Runsha Dong et al., "Design and Application on Metadata Management for Information Supply Chain," *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, Qingdao, China, pp. 393-396, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ohio Digital Library Platform Documentation, OverDrive, 2023. [Online]. Available: https://ohdbks.overdrive.com/

[4] Timothy W. Cole, and Muriel Foulonneau, *Using the Open Archives Initiative Protocol for Metadata Harvesting*, Westport, Libraries Unlimited, 2007. [Google Scholar] [Publisher Link]

[5] LibGuides: Metadata Basics: Harvesting, University of Texas Libraries, 2025. [Online]. Available: https://guides.lib.utexas.edu/metadata-basics/harvesting

[6] Malaka Friedman, "Matthew K. Gold and Lauren Klein, Eds., Debates in the Digital Humanities 2019," *Hyperrhiz: New Media Cultures*, no. 24, 2021. [CrossRef] [Publisher Link]

[7] Ohio Digital Network, ODN Metadata Manual: Best Practices for Creating and Contributing Metadata, 2023. [Online]. Available: https://ohiodigitalnetwork.org/.

[8] Jung-ran Park, and Eric Childress, "Dublin Core Metadata Semantics: An Analysis of the Perspectives of Information Professionals," *Journal of Information Science*, vol. 35, no. 6, pp. 727-739, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[9] Alon Kadury, and Ariel J. Frank, "Harvesting and Aggregation of Digital Libraries using the OAI Framework," *In Proceedings of the Third International Conference on Web Information Systems and Technologies WEBIST*, Barcelona, Spain, vol. 1, pp. 441-446, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[10] Robert R. Downs et al., "Harvestable Metadata Services Development: Analysis of Use Cases from the World Data System," *Data Science Journal*, vol. 22, no.1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] OCLC, Descriptive Metadata for Web Archiving: Review of Harvesting Tools, OCLC Research, 2018. [Online]. Available: https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata/harvesting-tools.html

[12] Oral History Association, Oral History Metadata Survey, Oral History Association, 2019. [Online]. Available: https://www.oralhistory.org/research/

[13] Ian H. Witten, David Bainbridge, and David M. Nichols, *How to build a digital library*, Morgan Kaufmann, 2nd ed., 2009. [Google Scholar] [Publisher Link]

[14] Xavier Ochoa, and Erik Duval, "Towards Automatic Evaluation of Metadata Quality in Digital Repositories," *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, vol. 4231, pp. 372-381, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[15] Katrina Fenlon, "Modeling Digital Humanities Collections as Research Objects," *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Champaign, IL, USA, pp. 138-147, 2019. [CrossRef] [Google Scholar] [Publisher Link]