Protocol for Metadata Harvesting: The Role of OAI-PMH in Digital Resource Integration

Haribararao Mojjada

MVGR College of Engineering(A) Vizianagaram-A. P, India-535005

DOI: https://doi.org/10.51584/IJRIAS.2025.100700066

Received: 15 July 2025; Accepted: 22 July 2025; Published: 11 August 2025

ABSTRACT

In the evolving digital knowledge ecosystem, metadata harvesting plays a crucial role in ensuring seamless access, interoperability, and visibility of scholarly content across distributed digital repositories. This study examines the significance of metadata and explores the architecture, functionality, and implementation of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). By facilitating standardized metadata exchange, OAI-PMH enhances federated discovery, open access, and integration of institutional and global repositories. The research employs a qualitative, literature-based approach and demonstrates successful use cases at NDLI, OpenAIRE, and DSpace. There are existing challenges in metadata harvesting at scale with OAI-PMH, including that it does not support full-text items very well and does not provide real-time updates. The research also discusses new trends in harvesting metadata, including semantic web technologies, AI-enrichment, APIs, and devices such as ResourceSync, which enables the real-time synchronization of items. The research concludes with further thoughts on the usefulness of OAI-PMH in the future and provides recommendations for hybrid approaches to metadata management for future digital libraries.

Keywords: Metadata Harvesting, OAI-PMH, Interoperability, Digital Libraries, Institutional Repositories.

INTRODUCTION

In the increasingly dynamic digital information environment, metadata serves as an essential facilitator of access, discovery, and organization of digital content. As higher education, research, and library sectors increasingly develop and publish scholarly output in digital form, the proliferation and diversity of metadata continues to provide an enormous landscape. However, given the differences in systems, OS, and platforms that have been used to store and provide access to digital content through institutional repositories, digital libraries, and open access archives, the ability to share and integrate metadata across independent or heterogeneous systems can create complexity. How, then, do we begin to achieve standardization to support integration? This is where recognized, standardized protocols such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) become critically important. OAI- PMH enables the harvesting of structured metadata from multiple sources on a large scale and automates the retrieval process. By supporting the ability for service transforms and aggregating content into a unified interface, OAI-PMH enables content providers to provide metadata, at least at a basic level, while maintaining equitable access to quality resources within a user interface. OAI-PMH also maintains the ability to access metadata in a standard way, often in XML with the Dublin Core metadata standard, which simplifies interoperability. Bridging repositories through OAI-PMH provides visibility of resources outside of a single repository, federated searching of heterogeneous metadata, and user access to knowledge. Metadata and protocols for harvesting, such as OAI-PMH, provide integrated access to the resource-rich world of digital content, contributing to collaborative information-sharing efforts and advancing the broader contexts of open access and scholarly communication.

Significance of the study:

We are witnessing a migration to the digital environment in all areas of Library and Information Science (LIS). In the current digital landscape, the possibilities for library and archival professionals and students are beyond anything we could have envisioned a few decades ago. We are in a digital environment that is increasingly shifting from a primarily catalog-based and physical resource orientation to a more metadata-rich, integrated digital world. Metadata is now central to our work in resource discovery, interoperability, and knowledge sharing. There will be a growing need for LIS professionals to be competent in metadata standards, such as Dublin Core, harvesting protocols like OAI-PMH, and utilizing digital library software, including DSpace, EPrints, and other similar platforms. There are evolving trends in utilizing Semantic Web technologies, adding API's to systems, and incorporating AI-based enrichment to enhance metadata interoperability. Institutions are focusing more on workflow processes, leveraging metadata to integrate across systems, and supporting federated access to institutional knowledge. Therefore, it is not a surprise that research on metadata harvesting and the management of harvested metadata is meaningful, relevant, and should be studied by today's LIS scholars and professionals.

Objectives of the Study

To examine the importance and value of metadata to the digital information environment as it relates to access, organization, and discovery of scholarly resources.

To describe the structure, standards, and input capacity of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to facilitate interoperable metadata exchange.

To explore the application of OAI-PMH and the use cases, at the national and international levels, that make use of OAI-PMH in digital libraries and repositories.

To evaluate the constraints and particulars of OAI-PMH, including scalability, real- time updates, complete metadata, and security.

To suggest possible future or innovative new directions of metadata harvesting, taking into account semantic web developments, AI-based enhancement, and hybrid metadata.

Research Design of Study

The study employed a qualitative research design, utilizing an exploratory and descriptive approach, to evaluate the role, implementation, benefits, and limitations of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in the integration of digital resources. The study has also incorporated secondary data sources from an extensive review of academic literature, technical documentation, standards manuals, and case studies of digital libraries and institutional repositories, including NDLI, DSpace, OpenAIRE, and BASE. The sources came from peer reviews, conference proceedings, white papers, and institutional reports that were relevant to metadata management and digital library protocols. The study has an emphasis on synthesizing contemporary practices, evaluating the technical functionality of OAI-PMH, and uncovering the scope and future directions for harvesting metadata. The goal was to establish an understanding of the core structural features, operational workflow, use cases, and integration issues for OAI-PMH. The study offers a comprehensive framework for researchers, data curators, librarians, creators, and intermediaries, facilitating discovery and interoperability within the scope of distributed digital data while addressing gaps and opportunities for development and technological advancement.

REVIEW OF LITERATURE

In their article, Baglioni et al. (2025) reported on the interoperability of scholarly repository registries. The abstract provided commentary on authoritativeness, disambiguation, and coverage as concerns due to fragmented registries. The focus, therefore, is advocating for interoperability among scholarly repository registries to facilitate full use of research products. They sought to analyze the data models of four significant registries (FAIRsharing, re3data, OpenDOAR, ROAR) and recommend a standard data model. Their findings introduce a common data model that enables the sharing of information about repositories and supports single-registration workflows.

The paper by Psyrra and Mangina (2023) proposes techniques for Moodle administrators to semi-automatically describe learning resources using LOM-based metadata, thereby improving discoverability. The theme is enhancing the reusability and discoverability of XR digital learning objects in Moodle repositories. Objectives include demonstrating two plugins for metadata creation and OAI-PMH exposure. Findings indicate that integrating IEEE-LOM and OAI-PMH standards is feasible for enhancing the discoverability of learning content.

The paper "Open Harvester Systems with Special Reference to OAI-PMH Service Providers: A Study" by Dr. Bairam Khan (2023) highlights the study of Open Harvester Systems (OHS) and OAI-PMH Service Providers. The theme revolves around the importance of OAI-PMH- supported metadata harvesting for next-generation library interfaces. Objectives include discussing the importance of OAI-PMH and identifying compliant service providers. Findings reveal that 19 OAI-PMH Service Providers are listed, with none from India, Asia, Oceania, or Antarctica. Additionally, digital resources are available in multiple languages.

This review focuses on "Reproductive health information needs and maternal literacy in the developing world: A review of the literature" by Margaret S. Zimmerman (2017). The abstract highlights the analysis of literacy and education's relationship with reproductive health in developing nations, where over 800 women die daily from pregnancy-related causes. The theme examines the crucial connection between maternal literacy and maternal and child health outcomes. Objectives include reviewing literature on health disparities and identifying successful educational interventions. Findings indicate a positive correlation between female literacy/education and improved maternal and child health outcomes, emphasizing the benefit of adult literacy acquisition.

This paper by Kurt Hornik (2017) introduces the OAI-Harvester R package for metadata harvesting using the OAI-PMH protocol. The theme focuses on facilitating efficient content dissemination and repository interoperability through OAI-PMH. Objectives include providing R functions for OAI-PMH requests and transforming XML results into functional R data structures. Findings show the package successfully performs OAI-PMH verbs, handles large lists, and transforms Dublin Core metadata for analysis.

This research aims to ingest metadata from Indian institutional digital repositories to build a single repository utilizing OAI-PMH and Dublin Core standards, thereby enhancing access and interoperability (Teli, 2015). The goals of this research are to aggregate metadata into a single searchable platform and make the research outputs freely available. The study showed that by utilizing six repositories, around 34,500 metadata records were harvested. OAI is utilized by Search Encore, a PHP-MySQL web application system that enables searching by faculty, publication title, and repository. The literature review contained studies on metadata in their micro-document, e-resources, and web resources, with OAI-PMH as the low-barrier metadata harvest protocol, and OAI-ORE can be used to aggregate different digital objects like images, rich media, and data as the aggregating mechanism (Roy et al., 2013; Winn, 2010). DSpace version 1.4 is used to set up repositories, while DOAR is used to corroborate global standards

This paper applies the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) from digital libraries to GIS for spatial metadata interoperation (Mao, 2005). It aims to provide a new method for transmitting, sharing, and interoperating spatial information. The study extends OAI-PMH to support spatial metadata standards and elements. Findings indicate the developed system is feasible for spatial metadata interoperation, offering a rapid and effective way to share metadata across various fields.

This article by Hunter and Guy (2004) discusses "Metadata for Harvesting: The Open Archives Initiative, and how to find things on the Web." The abstract highlights the potential of OAI-PMH for resource discovery beyond standard search engines. The theme is leveraging metadata harvesting to make institutional resources widely available. Objectives include illustrating how existing metadata can be repurposed for OAI-compliant repositories and addressing practical/philosophical questions. Findings indicate that OAI-PMH facilitates efficient and low-cost resource discovery; however, copyright and intellectual property issues require further attention.

Metadata Harvesting

Metadata: Metadata is commonly defined as "data about data." It provides descriptive, structural, and administrative information about digital or physical resources, such as title, author, date, format, and subject. In the context of digital libraries and repositories, metadata enables efficient organization, discovery, retrieval, and management of resources. It plays a crucial role in enhancing the accessibility and usability of digital content by serving as a standardized summary of the resource’s key attributes.

Metadata Harvesting: Metadata Harvesting is the process of systematically collecting metadata records from multiple repositories and aggregating them into a centralized platform or service. This method collects only the descriptive metadata that represents the content, not the content itself. All the harvested metadata can be indexed and searched, allowing for the discovery of distributed digital resources from a single point of access. This benefit is particularly relevant for institutional repositories, open-access archives, and digital libraries, where content is distributed across multiple systems and platforms. Due to the shared metadata, harvesting enables the centralization and integration of resources, thereby improving visibility, discoverability, and access to educational and scholarly content. An arrangement like OAI- PMH (Open Archives Initiative Protocol for Metadata Harvesting) enables standardization of this effort, supporting interoperability among heterogeneous systems, and leads to more effective knowledge sharing and access across institutions.

Benefits of Metadata Harvesting

Increased Discoverability

Metadata harvesting helps improve the discoverability of digital content by aggregating metadata from multiple repositories into a single location. In this way, users can search for and find resources from various institutions or platforms through a single interface, rather than browsing multiple databases. This not only enhances users' access to scholarly resources but also improves the visibility of content and preserves valuable research outputs, particularly those stored in open-access repositories, making them discoverable for researchers, students, and the public.

Improved Interoperability

Interoperability is crucial in various situations where systems comprise disparate software and are based on different standards, often within a highly heterogeneous digital environment. Metadata harvesting will support interoperability by implementing agreed common metadata standards like Dublin Core (as an example) and by using protocols like OAI-PMH to allow various digital libraries and repositories to communicate and share metadata. Ultimately, this means institutions can share information more quickly and broadly, pull data in from systems more easily, and participate in knowledge networks without the need to go through proprietary technology or customized implementations.

Efficient Aggregation of Resources

Metadata harvesting allows for the compilation of layered, descriptive information from multiple repositories into a single service. By bringing together all these different repositories to provide a vast index of metadata representing resources distributed across multiple locations,we significantly increase access to resources from various systems without physically moving or copying any resources. It results in a bigger universe of aggregated content for digital libraries, federated search systems, and discovery tools, which makes everything you collect or create more effective and impactful on institutional and academic collections.

Cohesion support for open access

Metadata harvesting is crucial to advancing equitable open-access efforts, which involve integrating metadata from open-access repositories into public discovery tools. This facilitates the discovery and access to freely available scholarly literature, increasing exposure for institutions that have not experienced global discovery. In harvesting efforts, open-access content is accessible for participating researchers and contributes to global efforts in making scientific and academic information available to all, improving equity in knowledge sharing.

Reduced Redundancy

Instead of being a source of full-text documents of files, metadata harvesting is simply gathering descriptive metadata, which removes the weight of redundancy, both in storage and in data integrity. The process of organizing simply by gathering stacks of metadata is an efficient method for loading large amounts of data onto digital platforms, while managing the tool's progress through the discoverability and accessibility of content.

Additionally, within the scope of minimizing duplication, this approach reduces the amount of outdated and redundant content in the harvested metadata, as only metadata is harvested and only checked periodically. Increased Efficiency and Sustainability for Resource Integration: Metadata harvesting supports federated search practices, which enable users to search across multiple repositories or databases simultaneously. By harvesting and indexing metadata, users can query a single query and retrieve responses from various centralized information stores. This saves time for users and researchers by retrieving more relevant literature or resources, while providing a more complete understanding of the long, continuous tail of literature or resources. Libraries, universities, and research sites that monitor usage can provide users with unimpeded access to a selection of academic content.

Better Resource Management

Harvesting metadata can provide organizations with the option to analyze and rationalize metadata in ways that may provide more insight into the types, sources, and usages of digital content. This may help organizations manage their collections more effectively, track usable digital assets, and enhance mail address visibility in their academic output. Suppose an organization has harvested a database of records that contains metadata. In that case, it may be useful for organization administrators to note gaps that exist and track access to digital content over time. Administrators can view only records related to items or resources that they have created and shared, and use that information to inform decisions regarding digital preservation, collection development, and communications with users or other stakeholders.

Scalability and Automation

With harvesting metadata the process is scalable and highly automated, as the usage of repositories, and/or collections, change the metadata records can be updated or new metadata records can be harvested. Typically, metadata can be harvested from multiple repositories enabling frequent and timely updates of any datasets or indexed search databases. Due to the scalability of harvesting, many thousands of metadata records may be harvested as collected with little human interaction. This is especially advantageous for larger institutions or repositories created by multiple institutions. Automated harvesting is also advantageous because it creates consistency in metadata records, and updates, therefore timestamps are accurate to the harvesting repositories that supplied the records.

Foundation for Advanced Services

Metadata harvesting can provide a solid foundation for advanced digital services, such as citation tracking, content recommendations, usage analytics, and user-generated content links. Advanced digital services can enhance user engagement, enrich academic research workflows, and increase the value of digital repositories to users. Additionally, metadata harvesting can support integrations with artificial intelligence tools, visualization tools, or scholarly communication networks, consistently extending and growing the capabilities and scope of digital libraries.

Introduction to OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard has become the widely accepted means of exchanging metadata between digital repositories. OAI- PMH provides an infrastructure for service providers (clients) to harvest metadata records from data providers (servers) in a standardized, fully automated method. OAI-PMH is designed using web standards, is reasonably simple to deploy with existing web technologies (OAI-PMH uses HTTP for communication and XML for encoding metadata), is lightweight, and is interoperable across multiple systems, while providing access to additional standard metadata formats, regardless of the details of its schema (e.g., Dublin Core, MODS). OAI-PMH includes six basic service requests (or verbs) that provide the methods for basic OAI-PMH services (e.g., Identify, List Records, Get Record) to enable the harvesting of metadata responses. OAI- PMH is a lightweight protocol that enables cross-platform metadata integration, supports interoperability and discovery of resources, and simplifies access to digital library content. The OAI-PMH has been adopted and implemented in repositories, institutional libraries, and open archives worldwide.

Core Features

Lightweight HTTP-based Protocol using RESTful Architecture:

OAI-PMH is a lightweight protocol that utilizes a RESTful (Representational State Transfer) architecture, employing HTTP to facilitate standardized and straightforward client-server communication through service requests and responses. OAI-PMH does not utilize a complex web service relying on SOAP or other technologies; instead, it uses a random client-server interaction pattern with uniform service requests, utilizing URLs and query parameters across the two systems. OAI-PMH maps methods to a simple HTTP GET or POST action (e.g., ListRecords or GetRecord) with complete stateless and scalable independence. OAI-PMH is a lightweight protocol/block that integrates with digital repository systems, utilizing scripting languages (e.g., Python, PHP) to execute code. OAI-PMH provided a method or means to exchange using only lightweight processes regarding the library's file formats for the existing data, applying the harvesting protocol.

Uses Dublin Core as Default Metadata Format

OAI-PMH utilizes Dublin Core (DC) as its default metadata schema and is well-known among libraries and digital repository staff. Dublin Core is a defined suite of 15 fundamental elements and properties, using terms such as title, creator, subject, and date, to constitute a set for consistently describing digital resources. OAI-PMH is a schema-agnostic format that is also compatible with XML, allowing repositories to exchange or share data with interoperability using any internal metadata standard between repositories. All OAI-PMH-compliant repositories must expose metadata in the Dublin Core metadata format, as well as provide metadata in other schemas (e.g., MARCXML, MODS), ensuring a baseline DC metadata schema across multiple systems and establishing a metadata floor for collections or systems harvesting metadata.

Support Incremental Harvesting Using Datestamps

The hallmark of the OAI-PMH standard is its support for incremental harvesting. Incremental requests occur when service providers or clients request metadata records that have been added or modified since the last harvest. The OAI-PMH standard provides OAI-PMH service providers and clients with the option to support incremental harvesting when making requests by utilizing date stamps. This allows the use of the two datestamp parameters (from and until) in the OAI-PMH request to retrieve metadata records added or modified since the last harvesting period. This is a helpful request to economically and operationally reduce costs, increase efficiency, and support digital repository synchronization over time, while also reducing bandwidth (if incremental user data). The incremental harvesting request is handy to aggregators or large scale repositories that only want to manage the increasing number of records, they do not want to (from a logistical perspective) to repeat harvesting whole inventory sets or groups, and with order and transaction management of the most recent records in their inventory if that is the business they wanted. Incremental harvesting also allows the provider or researcher to engage with the application as a pathway for measuring system performance and collection of metadata responsibilities when conducting ongoing background processes for harvesting records.

Facilitates Selective Harvesting by way of Sets and Metadata Formats

With OAI-PMH, selective harvesting is facilitated by allowing service providers to harvest metadata by one or more "sets" — the logical grouping of records — or by one or more metadata formats. Repositories can create sets and organize their records into sets (grouped records based on various criteria, such as subjects, departments, or collections), and expose them to harvesters using the ListSets verb. Similarly, repositories can also expose multiple metadata formats (for example, Dublin Core, MARCXML, etc.) associated with the same Record. By providing selective harvesting, OAI-PMH offers a finer level of control over the harvested metadata. With selective harvesting, harvesters can be more explicit in their queries and harvest only the records they seek, thereby limiting the amount of unnecessary data transmission.

Technical Summary of OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) was developed to provide a standard for metadata exchange between repositories. OAI-PMH operates through a client-server interaction over the

HTTP protocol, where the Data Provider represents repositories that expose metadata, and the Service Provider represents systems or services that can harvest the metadata to aggregate, index, or reuse it. The interaction between the Data Provider and the Service Provider is dictated through six verbs, or service requests, which detail how the metadata will be requested or accessed.

Identify – This verb will present general information about the repository, including the repository name, base URL, the protocol version in use, the earliest date stamp the repository will provide, and an admin contact. The Identify verb helps the harvester determine the originating source and verify whether the repository is compatible with their harvesting activities.

List Metadata Formats – This command lists the metadata schemas supported by this repository (for example, Dublin Core, MODS, MARCXML, etc.). Knowing the available metadata schemas enables the service provider to select the correct metadata format based on their preferred schema and the compatibility of their systems with the metadata. Listsets – This verb returns the hierarchical tree of sets in the repository. Sets are ways to group records into optional collections, departments, or subjects for selective harvesting.

List Identifiers – Returns a list of record headers (unique identifiers with datestamps) without the entire metadata. This is useful for synchronization and tracking changes without transferring the complete Record.

List Records – Returns the complete metadata records from the repository in the specified format. This is the primary verb used for bulk metadata harvesting.

Get Record - Returns a single metadata record by its unique identifier. It is typically used for specific queries or when you want to perform an update.

These verbs provide structured, interoperable, and scalable communication between metadata producers and harvesters, making OAI-PMH a vital tool for integrating digital resources across various environments.

Role of OAI-PMH in Digital Resource Integration:

OAI-PMH is an essential pathway for connecting digital resources across multiple platforms, systems and institutions. We are in a data -rich world for academia and research where the nature and type of digital content is often fragmented across millions of largely differently built repositories (insert diff) with different technology, metadata standards, and access protocols.

OAI-PMH provides a structured, standardized, and interoperable method for harvesting different systems' metadata sources, and allowing repositories to offer a discovery/search mechanism for users to search and/or discover 'records' of the multiple digital resources.

The most ubiquitous use case for OAI-PMH is the harvesting of metadata from institutional repositories (IRs) such as DSpace, EPrints, Fedora, etc.. Any IR may include a plurality of scholarly outputs, including but not limited to: theses, dissertations, journal articles, datasets, etc. OAI-PMH allows OAI-PMH service providers to consistently and systematically harvest metadata from these repositories, and them cluster metadata into a centralized index.

Open DOAR (Directory of Open Access Repositories) and BASE (Bielefeld Academic Search Engine) routinely harvest openly available metadata from academic repositories via OAI- PMH. Harvesting publicly available metadata and making it openly accessible increases the availability of scholarly discourse, particularly scholarly outputs produced by institutions that are not primarily contextualized with the international one.

Another use case for OAI-PMH are federated/discovery services (e.g. WorldCat, Europeana, Primo) that harvest metadata from a variety of repositories and library systems to provide individuals with a world view of searching. OAI-PMH allows federated/discovery services to be compliant, and regularly harvest so that they can stay current with the latest metadata records, contribute to the discovery of individual outputs, and engage continuously with digital repositories.

OAI-PMH supports collaborative, multi-institutional, academic work among research consortia, digital libraries, and national repositories (e.g. UK, Australia). OAI-PMH makes it easier for institutions to synchronize data in a fact-to-fact manner, which continues to alleviate the need to eliminate redundancy and establish consistent metadata for it and across systems.

In another sense, OAI-PMH serves to a connector for a range of different repositories/services that assist in making one primary digital infrastructure. OAI-PMH creates an agent for automating the sharing of structured and standards-compliant metadata. For example, OAI- PMH could enable institutions to connect their digital collections/assets to larger discovery services, making contributions to open access, interoperability, and collaboration in the research process.

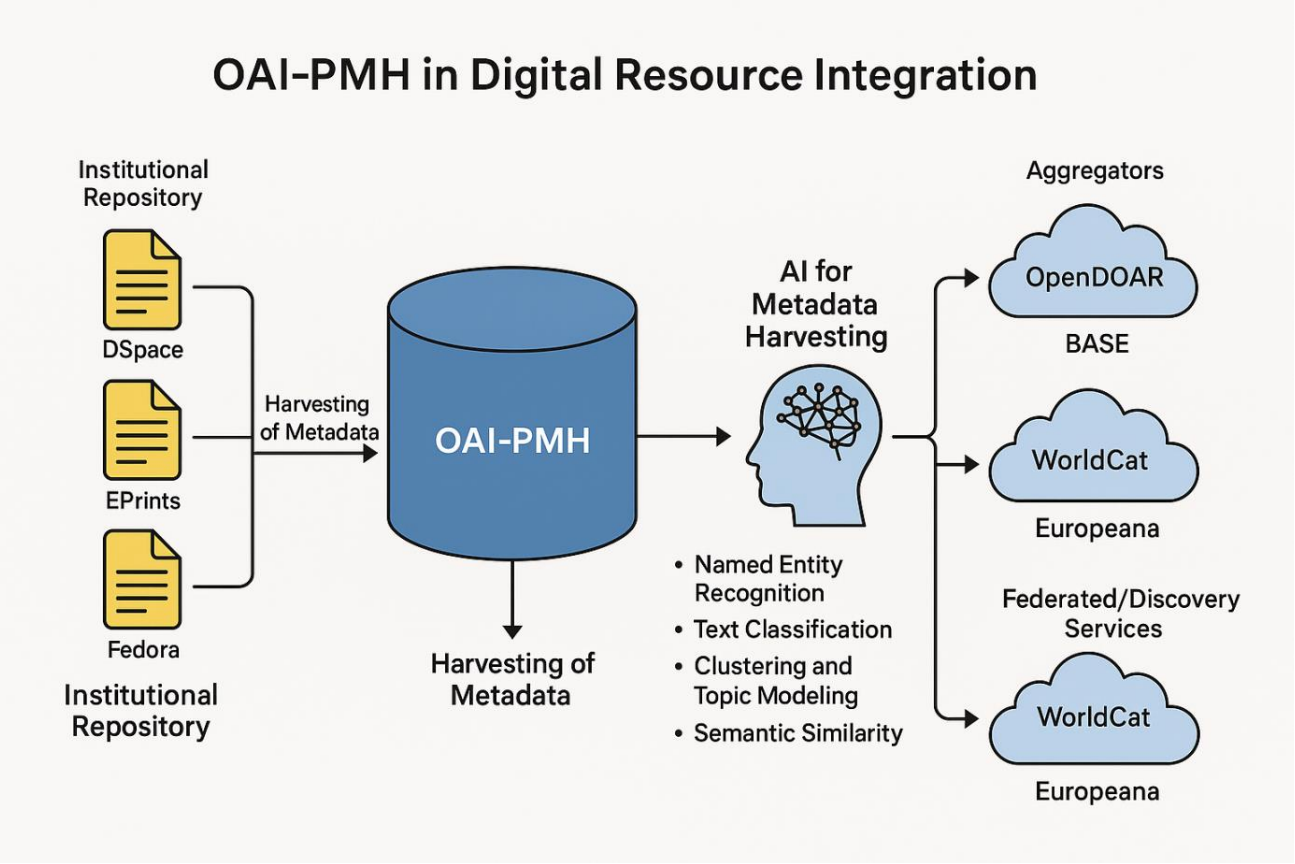

Figure 1: Role of OAI-PMH in Digital Resource Integration

Technical Exploration: AI Integration and Scalability in Metadata Harvesting

While the original study identifies future directions, such as AI-driven metadata enrichment, and mentions protocols like ResourceSync to address real-time updates, it does not thoroughly investigate **specific AI techniques or practical implementations**. This section addresses those omissions.

Artificial Intelligence (AI) Techniques for Metadata Enrichment and Classification

AI and Machine Learning (ML) have practical implementations in improving metadata quality, consistency, and scalability:

Named Entity Recognition (NER): NLP-based NER applications (e.g., SpaCy, BERT-NER) could also value/add by automatically extracting entities (author names, affiliations, location, keywords) from unstructured documents to enrich the fields in the metadata.

Text Classification Models: Deep learning models (e.g., Convolutional Neural Networks (CNNs), BERT-RoBERTa) can be trained to classify documents by subject category (e.g., based on title, abstract, or full-text) and add more descriptors (e.g., dc: subject) to the metadata.

Clustering and Topic Modeling: Algorithms (e.g., Latent Dirichlet Allocation (LDA), BERTopic) enable unsupervised discovery of themes from large datasets that could automatically populate metadata for theme browsing or faceted search.

Semantic Similarity and Linking: Applications (e.g., Sentence-BERT, Knowledge Graph Embeddings (e.g., RDF2vec) could determine semantic similarity between detailed metadata attributes across two metadata data sources that could provide for deduplication or similarity linking to documents across repositories.

Scalability solutions with AI and Big Data Techniques

Several technical approaches will be available to overcome the scalability issues that were inherent to OAI-PMH with large datasets:

Distributed harvesting pipelines: Unless you are harvesting metadata from very, very few repositories (a few GBs of metadata), you will need to implement distributed harvesting each time metadata is harvested. Distributed harvesting could leverage Apache Kafka for ingestion and then utilize Apache Spark, or other equally viable distributed systems, to process and write metadata from various repositories while simultaneously enabling Influx with real-time distributed computing.

Incremental Metadata Detection via AI: ML Models can be trained to recognize significant metadata changes that necessitate re-harvesting. This can eliminate unnecessary repeated harvesting and save system overhead.

Intelligent Scheduling Algorithms: Reinforcement Learning or heuristic-based scheduling (e.g., Ant Colony Optimization) can be used to determine the most efficient time and frequency to poll each repository based on updates and prioritization.

Security Considerations and AI Applications

OAI-PMH and ResourceSync are not built with secure authentication or encryption in mind. Security will not be an issue for AI augmentation if applied in an auxiliary security model:

Anomaly Detection for Metadata Access Logs: Metadata and harvesting access logs can be constructed as time-series data with unsupervised ML models such as Isolation Forests or Autoencoders to find abnormal user behavior and patterns in metadata access that could indicate misuse or cyber forensics.

Federated Learning for Sensitive Metadata: Rather than centralizing sensitive metadata, federated learning enables local models to be trained on institutional nodes and shared in an encrypted form of the model weights, thereby obfuscating the raw metadata and providing a secure and private instance of user data.

Integration with OAuth 2.0 and API Gateways: Modern APIs for harvesting metadata can be secured with recruitment resources such as the OAuth 2.0 protocol, and AI systems can monitor and throttle anomalous API behavior in real-time.

Current Implementations and Research Prototypes

OpenAIRE Research Graph: Incorporates AI techniques to deduplicate, classify, and enrich over 100 million harvested metadata records via OAI-PMH and REST API. Uses ML to build disambiguation models on authorship and affiliation.

CORE Aggregator (UK): Uses deep learning models to provision metadata enrichment and document classification. The pipeline is done by detecting duplicates, identifying language, and normalizing metadata at scale.

ROER (Repository of Open Educational Resources): NLP pipeline using BERT generates metadata descriptors on OER content harvested via OAI-PMH.

To develop AI-enhanced metadata harvesting ecosystems in OAI-PMH and ResourceSync to improve scalability, enrichment, and security is very possible. However, it will likely require functionality and performance reliant on a hybrid architecture. This will offer ease of current protocols with the standardization and detail of AI and API systems. For next-generation digital libraries, the following steps will need to be to develop an interoperable framework consisting of OAI-PMH, REST APIs, AI-based enrichment layers, and security modules.

Use Cases in Libraries and Repositories

Libraries and repositories globally have initiated the use of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to share metadata and build-up connections/integration capabilities and access. Key platforms and institutional repositories globally have implementations of OAI-PMH, and in many instances, it is obvious that OAI- PMH has enhanced services in library service, digital library service, and dissemination of academic information.

One of the best examples exists in the National Digital Library of India (NDLI). NDLI is harvesting metadata using OAI-PMH from a range of learning resources from Indian academic institutions, including IITs, universities, and colleges. NDLI is harvesting metadata from institutional repositories, and in turn, NDLI is in the process of creating a central online hub for educational resources for learners, teachers, and researchers in India to seamlessly and swiftly access an extensive range of learning resources such as books, theses, conference papers, video lectures, etc. Many universities worldwide utilize DSpace, a leading open-source institutional repository platform that can be configured to integrate with OAI-PMH. Most of these repositories are configured to expose records relevant to theses, dissertations, research papers, and datasets. Enabling OAI-PMH ensures that academic outputs from institutions are visible to harvesters or aggregators, such as BASE, CORE, and OpenDOAR, thereby enhancing the global discoverability of their scholarly work.

In Europe, OpenAIRE (Open Access Infrastructure for Research in Europe) has embraced OAI-PMH's mechanisms and functionality. OpenAIRE provides extensive support for the development of the open science movement by collecting metadata from European Universities, research organisations, and national repositories to expand a wealth of interrelated open-access scholarly publications. OpenAIRE not only tracks the dissemination of publicly funded research but also assists with determining compliance with funders' expectations related to publications under open access.

These examples demonstrate how OAI-PMH can be utilized to enhance collaboration, visibility, and access to knowledge, whether for national-level digital libraries such as NDLI, institutional repositories like DSpace, or continental-level networking initiatives like OpenAIRE. OAI-PMH, therefore, can enable the effective packaging, scaling, and standards- driven harvesting and processing of metadata, thereby augmenting the available infrastructure for modern digital scholarship to ensure that valuable research reaches as many people as possible.

CHALLENGES AND LIMITATIONS

Limited Support for Full-Text Content:

OAI-PMH is designed for metadata harvesting and is limited in its ability to transfer full-text or full-content documents as a single unit. In essence, a user can potentially find resources using the metadata provided by OAI-PMH; however, the link to the full-text content is outside of that context and is currently being established, which may later fail to be established, meaning reliable access to that resource cannot be assured. This type of limitation further diminishes its potential in terms of utility for full-content aggregation and systems designed to enable integrated content access.

No Real-Time Updates

As an OAI-PMH provider (service provider), metadata harvesting operates on a polling model and, as discussed elsewhere, the harvesting of metadata at scheduled intervals without an event- driven mechanism for updates allows for time delays in capturing added or changed records, which mean that that there will be time delays in the time-sensitive currency of records reflected on aggregated platforms. In highly dynamic and fast-growing repositories, these time delays may make OAI-PHM less effective in ensuring the freshness of search results and that users have access to the most current record entries.

Issues with Scalability with Large Repositories

OAI-PMH was designed to work efficiently with repositories of small to medium size; however, repositories with millions of records add a level of intractability not contemplated in OAI-PMH. Depending on the number of records, the protocol's linear request-response model, combined with excessive requests, can overload the data repository service, whose primary function is to respond to those requests. When too many requests for records occur in open- access systems, it is likely to cause undesirable delays in the response, which is not suited for use where each side is anticipating and expecting efficiency in providing timely and accurate answers. The request-response interaction is not intended, without additional load balancing or optimization, to store large content items, and it becomes inefficient for large-scale applications, such as real-time extraction applications.

Minimal Security and Authentication

OAI-PMH does not provide any means for secure transfer of data and no means for user authentication, which makes it untenable in situations that require safeguarding data and providing controlled access. The lack of such mechanisms limits the protocol's adoption in repositories that include sensitive or proprietary content, which requires graduated restrictions on accessing that content.

Emerging Technologies

One way to counter these limitations is to look to emerging technologies - things like ResourceSync (an extension to OAI-PMH specifically intended for synchronization solutions), and Linked Data, which a lot of the library community is studying as ways to scale better, consume real-time data, and provide semantic interoperability in new digital library infrastructure.

The Future of Metadata Harvesting and OAI-PMH

The Continued Relevance of OAI-PMH:

OAI-PMH can continue to be most useful as an underpinning protocol for metadata harvesting, particularly for digital libraries, institutional repositories, and open-access archives. OAI-PMH is a simple yet effective protocol that provides a REST-based standard, sufficiently widespread to maintain a reliable underlying standard for the exchange of basic metadata records. As the amount and complexity of digital content increase, OAI-PMH alone may not be sufficient to meet the metadata harvesting needs of the years to come.

The Adoption of Semantic Web Technologies

The future of OAI-PMH lies more with the adoption of semantic web technologies, including RDF (Resource Description Framework), and SPARQL (SPARQL Protocol and RDF Query Language). These technologies can handle machine-readable linked data, enabling more dynamic and meaningful connections between resources. By applying Linked Data principles and standards, repositories can better facilitate discoverability, data reuse, and the integration of contextually linked resources across different repositories, knowledge graphs, and digital experiences.

AI-Driven Metadata Enrichment

Artificial Intelligence (AI) and Machine Learning (ML) are set to change the landscape of metadata creation and enrichment. AI tools can also utilize machine learning to automatically generate descriptive metadata, perform content classification, and identify relationships between documents. This has the potential to reduce the manual input required from humans to create meaningfully descriptive metadata, improve the level of objectivity in the creation of descriptive metadata, and expand the semantic qualities of a metadata record, making analyses and retrieval of those records more informative and valuable.

API-based harvesting models and standards

OAI-PMH specifies verb types and request types, while API (Application Programming Interface) driven harvesting models based on RESTful or GraphQL APIs are inherent to modern web services. APIs lend themselves to convenience and flexibility, real-time data sharing or pulling actions, and finer-grained control of data exchange. Additionally, APIs are particularly useful for implementing complex systems, interactive applications, and providing users with immediate and direct access to information.

Getting Real-time Updates with ResourceSync

ResourceSync (NISO) may overcome some of the limitations of OAI-PMH, especially since it was developed specifically to accommodate OAI-PMH, allowing metadata and content to be synchronized between repositories in near real-time. Metadata content can be shared so that additions, deletions, and updates can be notified, thereby negating the reliance on OAI-PMH's periodicity. ResourceSync can be very useful when repositories require frequent updates to their information.

Revisiting Proprietary Systems, Legacy Systems, and Standards

Moving forward, libraries and digital repositories will need to balance OAI-PMH with the guidance from this study about more modern interoperable technologies. Hybrid strategies— using OAI-PMH in conjunction with Application Programming Interfaces, ResourceSync, or a Semantic Web tool—to develop controlled access to content are necessary to ensure continued interoperability while experimenting with innovations in both metadata harvesting and the digital publication ecosystem.

CONCLUSION

The current digital scholarly publishing environment fosters a robust ecology of interoperability among diverse offerings—scholarly communication, marketing, grants, big and small data, learning management systems, institutional repositories, open access, and public-facing scholarly publishing. However, this environment ultimately leads to a disconnect between producers and consumers of scholarly research. The impetus for technological advancements is creating the necessary infrastructure to ensure that institutional repositories remain openly available, integrating, organizing, and providing seamless access to the expanding body of scholarly content that continues to be housed in an ever-increasing number of disparate repositories that comprise the current digital scholarly publishing ecosystem. This study suggests that metadata, particularly metadata generated through the OAI-PMH protocols, will remain essential for supporting unified access to content, interoperability, and resource integration in digital libraries and institutional repositories. The lightweight architecture of OAI-PMH, with clearly defined service verbs and resource description vocabularies including Dublin Core, is a proven instrument for enhancing metadata interoperability among tool vendors and the broad dissemination of metadata.

This protocol has the practical potential to unify scholarly outputs that currently exist separately and to promote open access engagement without restriction through engagements like NDLI, OpenAIRE, and DSpace. It has proved its ability to publish external scholarly outputs. Less helpful are the areas where OAI-PMH has limitations, including non-synchronistic operation, lack of full-text information, and scalability issues. Looking forward, areas for additional development and improvement have emerged from this document, namely the Semantic Web and the use of other approaches, such as APIs or ResourceSync, which require the implementation of AI and metadata enrichment for storage.

The landscape of metadata management is evolving. While OAI-PMH is at the foundation of reading and harvesting metadata, digital libraries should be prepared to adopt a hybrid approach that combines legacy compatibility with innovative technology. A commitment to the solid and consistent use of metadata interoperability may ultimately enhance the discoverability of scholarly knowledge while supporting new kinds of global academic collaborations with an equitable distribution of scholarly digital access.

REFERENCES

1. Aguado López, E., Rogel Salazar, R., Becerril García, A., & García Flores, H. (n.d.). Redalyc OAI–PMH: The Open Archives Initiative Protocol for Metadata Harvesting. Universidad Autónoma del Estado de México.
2. Digital Commons. (n.d.). Digital Commons and OAI-PMH: Outbound harvesting of repository records.
3. Guy, M., & Hunter, P. (2004). Metadata for harvesting: The Open Archives Initiative, and how to find things on the web. The Electronic Library, 22(2), 168–174.
4. Mao, H. (2015). Research on metadata interoperability in digital libraries. In Proceedings of the 2015 International Conference on Education Technology and Economic Management (pp. 77–82). Atlantis Press.
5. Chumbe, S. (2015). OAI-PMH interoperability issues in institutional repositories: An assessment using the BASE test tool. ATINER Conference Paper Series LIS2015-1550.
6. Hofman, J., & de Boer, V. (2023). Metadata harvesting with R and OAI-PMH: A scalable solution for data integration. In Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2023) (pp. 541–552). Springer.
7. Zhang, X., & Liu, Z. (2022). Linked data-driven metadata harvesting for digital libraries. International Journal on Digital Libraries, 23, 111–125.
8. Heery, R., & Powell, A. (2006). Metadata harvesting and interoperability in digital repositories. IFLA Journal, 43(3), 217–229.
9. Teli, S. (2015). Metadata harvesting from selected institutional digital repositories in India: A model to build a central repository. International Journal of Innovative Research in Science, Engineering and Technology, 4(4), 1935–1942. https://doi.org/10.15680/IJIRSET.2015.0404018
10. Mihajlovic, J., & Stoimenov, L. (2004). Achieving OAI-PMH compliance for CDS/ISIS databases. The Electronic Library, 22(2), 146–150.
11. Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved from
12. Candela, L., Castelli, D., & Pagano, P. (2007). On integrating heterogeneous digital libraries: The DELOS approach. Information Systems, 32(5), 367–384. https://doi.org/10.1016/j.is.2005.11.003
13. Suber, P. (2012). Open Access. MIT Press.
14. Van de Sompel, H., & Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 6(2).
15. Figure.1 Role of OAI-PMH in Digital Resource Integration retrieved from chatgpt by giving prompt.