Metadata Harmonization and Semantic Interoperability in Modern Digital Repositories: An Analytical Framework Enhancing Resource Discovery, Data Exchange, and Long-Term Accessibility

[1]Amandeep Kaur, [2]Sunil Kumar, [3]Riju Kumari [4]Mamta

[1]Phd Research Scholar, [2] Chief Librarian,
[3]Phd Research Scholar, [4]Phd Research Scholar
[1]Department of Library and Information Science
[2]PIMS Medical College and Hospital
[3,4]Department of Library and Information Science,
[1]Lamrin Tech Skill University Ropar Punjab
[2]Jalandhar Punjab India
[3,4]HRIT University Ghaziabad, UP, India
[1]ORCID ID: - https://orcid.org/0009-0006-6870-4030
[2]ORCID ID: https://orcid.org/0009-0005-3724-3481
[3]ORCID ID: - https://orcid.org/0009-0001-8893-024X
[1]Email. ID: - kauramnibrar@gmail.com
[2]Email ID: librariansunilkumar1@gmail.com
[3]Email: librarianrijukumari@gmail.com
[4]Email ID: - mamta116kumari@gmail.com
doi.org/10.64643/JATIRV2I6-140679

Abstract- The contemporary ecosystem of digital repositories is characterized by significant heterogeneity among objects or collections of objects, software, and services, as well as a growing expectation that knowledge will be more openly available, reusable, and machine-actionable. A theoretical framework is developed in the present paper to address the issue of metadata harmonization and semantic interoperability in digital repositories. Its utility is demonstrated through a simulated academic dataset of 60 repositories representing universities, libraries, archives, museums, and research data platforms. This study explores the relationships of metadata completeness, controlled vocabularies, persistent identifiers, availability of OAI-PMH/APIs, ontology or linked-data support and repository maturity

with resource discovery, data exchange efficiency and long term access. The framework was evaluated through the use of descriptive statistics, cross-tabulation, correlation analysis, regression analysis, ANOVA and reliability testing. According to simulated results, discovery effectiveness is highly associated with metadata completeness, while persistent identifier, API availability, and linked-data integration have substantial contributions to long-term accessibility. Repositories employing mixed schema or domain responsive metadata practices tend to yield stronger interoperability performance than those using isolated descriptive fields without machine-readable semantic structures. The findings of our analysis indicate that metadata harmonization is by no means a simple schema conversion task. It entails governance, semantic mapping, quality control, preservation metadata, authority control and ongoing technical maintenance. A framework is offered to repository managers, metadata librarians, digital archivists, and information policy makers which can support better discoverability, reduce (meta)data fragmentation, and foster durable access to digital resources.

Index Terms -Metadata harmonization; semantic interoperability; digital repositories; linked data; resource discovery; persistent identifiers; long-term accessibility

I. INTRODUCTION

1.1 Digital Repositories as Knowledge Infrastructure

The modern digital repositories have emerged as critical knowledge infrastructures for university libraries, archives, museums, research data networks, and scholarly communication networks. The function of repositories has evolved from simply storing digital objects, to actively supporting discovery, preservation, citation, reuse, impact assessment, and multiplatform exchange. As a result of this expansion, the quality of metadata is being seen as a strategic issue rather than simply a cataloguing issue. A repository may contain important scholarly or cultural resources. When records are incomplete, semantically inconsistent, weakly identified, or otherwise closed to machine harvesting, however, these resources will be difficult to find, interpret, connect, and preserve.

1.2 Metadata Heterogeneity and Standards Diversity

The problem is compounded by the presence of multiple metadata traditions. For creating such a content which serves as a connect between users and content, different standards are being used. For instance, libraries rely with solutions based on MARC21 and authority control. Above all, archives privilege provenance and context. On the other hand, museums favour object-centred description. Further, institutional repositories often use Dublin Core. At last, research data platforms make increasing use of DataCite metadata, persistent identifiers, and FAIR-oriented indicators. The traditions are not necessarily contradictory, but their conceptual granularity, encoding structures and local practices differ. Harmonization of specifications is necessary because repositories seldom function within the same descriptive universe. They share records

with systems that reveal information discovering layers, aggregators, indexes, data portals and preservation environment.

1.3 Harmonization and Semantic Interoperability as Analytical Concerns

The process of ensuring consistent description of digital objects across different metadata schemas, systems, and institutional practices. This includes the administrative, structural, technical, rights, and preservation metadata. This encompasses schema crosswalks, the use of a common vocabularies, persistent identifiers, validation rules, normalization procedures, and local field mapping to community standards. The systems can go beyond only understanding the data format used to describe those entities but can additionally understand and process the meaning of entities. The semantic layer is facilitated by RDF, SKOS, OWL, BIBFRAME and linked open data approaches which convert metadata from isolated strings into statements to be linked together between systems (Baker et al., 2013; W3C RDF Working Group, 2014; W3C OWL Working Group, 2012).

1.4 Repository Performance Assessment Rationale

The assessment of repository performance is reliant on the combination of five distinct factors, which include descriptive completeness, semantic control, identifier infrastructure, machine interfaces, and preservation-oriented governance. It has a simulated quantitative demonstration across “60” repositories to show that these variables can be analytically evaluated without misleading fieldwork. The proposed framework addresses the needs of LIbrary and Information Science, Digital Humanities and Information Management contexts where repository managers seek evidence-based tools to support improved discovery, data exchange and sustained long-term access.

II. METADATA HARMONIZATION AND SEMANTIC INTEROPERABILITY

2.1 Expanded Conceptual Scope of Metadata

In general, metadata refers to any type of structured information that describes, explains, locates, or makes it easier for a user to retrieve, use, manage, and preserve an information resource. In present-day repository environments, however, metadata contains more than just title, creator, date and subject. Moreover, it also includes identifiers, rights statements, preservation events, provenance, information on file formats, relationships between objects, funding, affiliation of an institution, and links to people, organizations, grants, datasets, software, publications, and instruments. An expanded understanding of a piece of data is consistent with the FAIR principles; namely, that data shall be findable, accessible, interoperable, and reusable by machines (Wilkinson et al. 2016).

2.2 Levels of Metadata Harmonization

Harmonization works at multiple levels. The syntactic harmonization of metadata allows for its encoding and exchange using XML, JSON, RDF, or similar techniques. Schema-level standardization maps fields across different standards such as: Dublin Core, MARC21, MODS, METS and DataCite. Semantic harmonization facilitates sharing of meanings by means of

controlled vocabularies, authority files, thesauri, ontologies, and knowledge organization systems. Organizational harmonization refers to the creation of rules that govern who produces metadata, how records are authenticated, how errors are removed, and the recording of versioning and preservation events. In the absence of these layers, schema adoption may lead to metadata that will superficially be standardized but semantically weak.

2.3 Metadata Standards and Repository Functions

The Dublin Core's core elements are simple, extensible, and widely used across repositories (Dublin Core Metadata Initiative, 2020). The MARC21 format is maintained so there can still be a rich bibliographic exchange between library systems. MODS is a more complex descriptive alternative to MARC21 and uses XML. This format helps to bridge the gap between what is required by library cataloguing and what is required by a digital repository (Library of Congress, n.d.-c, n.d.-d). METS encodes descriptive, administrative, and structural metadata for complex digital objects, making it important for digital libraries and preservation workflows (Library of Congress, n.d.-e). DataCite metadata expands repository description in the direction of citation of research data, contributor roles, related identifiers, resource types and DOI-based persistence.

2.4 Semantic Web Technologies and Persistent Identifiers

Another Dimension is Added by Semantic Interoperability A RDF allows the statements to be modelled as subject-predicate-object triples with which a repository description may be linked as the graph data (W3C RDF Working Group, 2014). The first project SKOS gives you a way to represent controlled vocabularies and classification schemes in a web-friendly way while OWL will let you do more expressive ontological modelling (Miles & Bechhofer, 2009; W3C OWL Working Group, 2012). BIBFRAME applies the principles of linked-data to the discipline of bibliographic description, which provides a transition path from MARC-centric exchange towards web-oriented bibliographic graphs. Using persistent identifiers like DOI, ORCID, Handle, and ARK creates stable links between objects, creators, organizations, and repository records (ORCID, n.d.; Van Wettere, 2021).

III. . REVIEW OF LITERATURE

3.1 Repository Interoperability and Metadata Exposure

The literature regarding repository interoperability demonstrates a movement beyond mere metadata exposure to multidimensional semantic and infrastructural integration. Case-based interoperability was typically based on harvesting mechanisms using the OAI-PMH standard Open Archives Initiative, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, the metadata exchange mechanism in that consequently downloading of data directly from the data provider). Even though OAI-PMH is still useful to aggregate digital content, contemporary research has treated interoperability more as a combination of quality of metadata, identifiers, API, semantic modelling, policy alignment and preservation planning.

3.2 Research Data Metadata and DataCite Practices

Research data repositories are designed, in many cases, to allocate metadata for each data object. DataCite dataset metadata affiliation information is influenced by free-text practices and identifier-based encoding (ORCID, organizational identifiers), according to Van Wettere (2021). According to Strecker (2024), DataCite DOI metadata records do change over time, but not always in any substantial way that makes them more complete. The significance of this finding is that, unlike the artifact which can be deposited once, repository metadata remains the subject of maintenance, such that its quality may affect reuse, attribution, and long-term interpretable.

3.3 FAIR Assessment and Metadata Re-Curation

The literature regarding FAIR assessment has influenced repository evaluation. Wilkinson et al. (2016) sees metadata as key to machine-actionable findability and interoperability; according to Aguilar Gómez and Bernal (2023), FAIR indicators need adaptation for institutional and multidisciplinary repositories. According to their FAIR EVA model, assessment frameworks should be operational enough for repository managers but flexible enough for different repository infrastructures and disciplines. This concurs with Habermann (2023), who argued for re-curation practices that improve and enrich existing metadata in repositories, rather than treating metadata deposit as a finished administrative task.

3.4 Linked Data and Knowledge Organization Systems

The scholarship surrounding linked data focuses on the organization of knowledge systems in accordance with the semantic discovery.

According to Baker et al. (2013), SKOS is a deliberately lightweight model for representing thesauri and classification schemes. Meanwhile, Zeng and Mayr (2019) illustrate how knowledge organization systems have entered the linked open data world. The term BIBFRAME refers to an effort on the part of libraries to make bibliographic data on the Web more addressable and semantically more expressive than record-based exchange. (Library of Congress, n.d.-a) Cultural heritage and archive studies also show that linked open data can increase contextual discovery, but only if mapping quality, vocabulary governance and institutional sustainability are addressed (LIBER, 2021).

3.5 Repository Registry Fragmentation

Recent research on repository registries demonstrates how fragmentation incurs costs. According to Baglioni et al. (2025), interoperability of repository registries is needed, because multiple registries can create duplicate profiles, uneven identifiers, and data fusion issues. Local metadata errors and incomplete identifiers do not remain local: they propagate into aggregators, indexes, catalogues and research information systems. Thus, this has implications for digital repositories. Consequently, it is that literature supports an analytical framework which takes harmonization as a local metadata quality issue and cross-institutional interoperability problem.

IV. . RESEARCH GAP AND PROBLEM STATEMENT

4.1 Limitations of Standards-Oriented Approaches

Existing research offers robust guidance with conceptual insight and technical specifications for metadata standards, FAIR assessment, linked data and repository interoperability. Nonetheless, many discussions are either standards-oriented or system-specific. An integrated analytical framework linking together metadata more complete and accurate, controlled vocabularies, persistent identifiers, machine interfaces, linked-data, discovery effectiveness, exchange efficiency, and long-term accessibility is given less attention.

4.2 Need for Measurable Repository Assessment Indicators

The link between semantic interoperability and generation of measurement repository is another gap. While standards bodies set the theoretical rules for how formats and protocols are supposed to behave, repository managers more often need practical indicators of whether harmonization contributes to improved discovery, lower error rates, better data exchange, and enhanced preservation readiness. This paper fills that gap with a simulated quantitative demonstration that explicitly disclaims fieldwork status and is designed to illustrate a replicable assessment technique.

V. . OBJECTIVES OF THE STUDY

5.1 Conceptual and Analytical Objectives

The study has 5 objectives. It first establishes an analytical framework to evaluate metadata harmonization and semantic interoperability using digital repositories. Next, it considers the factors that allow resources to be discovered, data to be exchanged and data to be stored for the long term. Third, it shows how simulated data from 60 repositories can be used to analyse repository performance. Fourthly, it interprets statistical relations of metadata completeness, semantic interoperability, discovery efficacy, exchange efficiency and accessibility. Lastly, it provides repository managers, metadata librarians, digital archivists and information policymakers with practical suggestions.

VI. . RESEARCH QUESTIONS

6.1 Guiding Research Questions

The project is motivated by four research questions: How do the metadata standards and the types of repository differ in modern digital repositories? How does metadata completeness help in resource discovery effectively? Which factors of harmonization influence accessibility in the long term? Are semantic interoperability scores different among repository types? The questions used for analytical demonstration can be adopted for institutional audits with actual repository data.

VII. . RESEARCH METHODOLOGY

7.1 Research Design and Dataset Specification

This study adopts a conceptual-analytical research design with support from a simulated academic set of data. The study was not conducted live repository harvesting or institutional survey. A

simulated dataset was created solely for demonstration purposes in research and 60 (hypothetical) digital repositories, which belong to 5 repository types, namely university repositories, library repositories, archive repositories, museum repositories and research data platforms. The created dataset is meant to mirror realistic variation in metadata practices as found in repository literature and standards documentation, and does not represent any named organization.

7.2 Variables and Measurement Indicators

This variable encompasses elements such as the type of repository, metadata standard, completeness of metadata, use of controlled vocabularies, presence of a persistent identifier, the repository's offering of OAI-PMH and/or API, ontology or linked-data usage, semantic interoperability, effectiveness of resource discovery, efficiency of data exchange, long-term availability, success rate of user search, the error rate of metadata and maturity of repository. To produce the scores we defined interpretable scales. More user search success rate (percentage); semantic interoperability (1–5 scale); discovery effectiveness (1–5 scale); exchange efficiency (1–5 scale); and long-term accessibility (1–5 scale) were all measured as a user percentage as was metadata error rate or the percentage of records with detectable descriptive or structural issues in the simulated model.

7.3 Statistical Procedures

Analisis dilakukan dengan menggunakan statistik deskriptif, distribusi frekuensi, cross tabulasi, korelasi pearson, regresi ganda, anova satu arah serta alpha cronbach. Regression model estimates impact of harmonization variables next long term accessibility The semantic interoperability scores across repository types will be compared in an ANOVA. An indicator of internal consistency of repository performance scale consisting of semantic interoperability, discovery effectiveness, data exchange efficiency and long-term accessibility is Cronbach's alpha.

7.4 Metadata Standards Included in the Framework

Table 1. Summary of metadata standards and their functions

Standard/Protocol	Primary function	Repository relevance
Dublin Core	General descriptive metadata	Simple cross-domain discovery; common institutional repository baseline
MARC21	Bibliographic metadata exchange	Rich library catalogue exchange and authority-supported bibliographic control
MODS	XML descriptive metadata	Richer descriptive records for digital collections and library-derived repository content
METS	Object packaging and structural metadata	Links descriptive, administrative, and structural metadata for complex digital objects
DataCite	Research data and scholarly outputs	DOI metadata, related identifiers, resource types, contributors, and citation support

RDF/SKOS/OWL	Semantic web and knowledge organization	Machine-actionable graphs, controlled vocabularies, ontologies, and linked data exchange
BIBFRAME	Linked bibliographic description	Web-oriented bibliographic modelling and transition from MARC-centric records
OAI-PMH/API	Machine interface protocol	Metadata harvesting, structured retrieval, aggregation, and platform-to-platform exchange
PREMIS	Preservation metadata	Documents preservation events, agents, rights, and object properties for long-term usability

Table 1 shows that metadata harmonization cannot be reduced to one standard. The standards differ in purpose: Dublin Core supports broad discovery, MARC21 and MODS support bibliographic description, METS and PREMIS support object management and preservation, DataCite supports research data citation, and RDF/SKOS/OWL/BIBFRAME support semantic exchange. A mature repository often needs a coordinated application profile rather than a single schema.

VIII. . ANALYTICAL FRAMEWORK FOR METADATA HARMONIZATION

8.1 Layered Structure of the Proposed Framework

The proposed framework is organized into four layers. The first layer is descriptive normalization comprising of complete titles, creators, dates, subjects, abstracts, formats, rights statements, resource type. The second layer semantic control, including authority files, controlled vocabularies, classification schemes, SKOS concept schemes and entity reconciliation. A persistent identifier with API, OAI-PMH endpoint switching code, crosswalk to semantic structures, metadata validation, and export formats. The fourth step is preservation and access governance, which encompasses events which are PREMIS-informed and fixity, file format monitoring, rights documentation, versioning, access metadata, and repository policy.

8.2 Interdependence of Harmonization Layers

The framework suggests that these layers strengthen each other. While completeness enhances human exploration, subject headings enhance precision and recall. Even as persistent identifiers stabilize references, semantic relations make those references meaningful. APIs allow for exchange but helps spread errors without normalization. Long-term accessibility of GLOBAID is supported by preservation metadata but like any type of metadata, preservation metadata is more useful when it is linked to object identifier, rights metadata, provenance, and technical metadata. Harmonization thus involves a socio-technical process that requires metadata design, system configuration, staff capacity, policy commitment, and iterative quality review.

8.3 Harmonization Variables and Performance Outcomes

The analytical model adopted in the paper considers metadata completeness, controlled vocabulary use, persistent identifier use, OAI-PMH/API availability, and ontologies and linked data as harmonization variables. The performance outcomes of repositories are considered to be semantic

interoperability, resource discovery, exchange efficiency, and long-term accessibility. Repository maturity serves as a contextual classification that reflects the combined technical, organisational and preservation readiness.

8.4 Simulated Repository Profile

Table 2. Profile of selected digital repositories in the simulated dataset

Repository type	n	%	Dominant metadata standard	Mean completeness	Mean semantic score	Modal maturity
University Repository	18	30.0%	Dublin Core	70.2	3.49	Developing
Library Repository	14	23.3%	MARC21	77.9	3.61	Mature
Archive Repository	10	16.7%	Dublin Core	66.1	3.30	Emerging
Museum Repository	8	13.3%	Dublin Core	69.1	3.19	Developing
Research Data Platform	10	16.7%	DataCite	77.7	4.31	Mature

Table 2 profiles the simulated repository population. University repositories form the largest group, followed by library repositories and research data platforms. Research data platforms show the highest mean completeness and semantic interoperability because the simulation assigns stronger PID and API practices to data-oriented repositories. Museum and archive repositories show lower averages, reflecting the complexity of object, provenance, and contextual description when standardized linked-data workflows are unevenly implemented.

IX. . DATA PRESENTATION AND ANALYTICAL INTERPRETATION

9.1 Operationalization of Variables

Table 3. Variables and measurement scale

Variable	Measurement	Operational meaning
Repository type	Categorical	University, library, archive, museum, research data platform
Metadata standard	Categorical	Dublin Core, MODS, METS, MARC21, DataCite, mixed schema, BIBFRAME-oriented
Metadata completeness	Continuous	0-100 percentage score indicating field completion and descriptive adequacy
Controlled vocabularies	Binary	Yes/No use of authority files, thesauri, subject headings, or controlled terms
Persistent identifiers	Binary	Yes/No presence of DOI, ORCID, Handle, ARK, or equivalent stable identifiers
OAI-PMH/API availability	Binary	Yes/No machine interface for metadata harvesting or structured retrieval

Ontology/linked data integration	Binary	Yes/No use of RDF, SKOS, OWL, BIBFRAME, or linked-data mappings
Semantic interoperability score	Scale	1-5 score representing machine-readable semantic alignment
Resource discovery effectiveness score	Scale	1-5 score representing search, browse, and retrieval effectiveness
Data exchange efficiency score	Scale	1-5 score representing export, harvesting, and aggregation readiness
Long-term accessibility score	Scale	1-5 score representing preservation-oriented and durable access readiness
Metadata error rate	Continuous	Percentage of simulated records with missing, ambiguous, or inconsistent metadata
Repository maturity level	Ordinal	Emerging, developing, mature, or advanced

Table 3 clarifies how each variable is operationalized. The inclusion of both technical indicators and outcome scores is important because repository assessment should not stop at whether a standard is named in policy. It should examine whether metadata practice produces measurable improvements in semantic exchange, discovery, and accessibility.

9.2 Cross-Tabulation of Repository Type and Metadata Standard

Table 4. Cross-tabulation of repository type and metadata standard

Repository type	BIBFRAME-oriented	DataCite	Dublin Core	MARC21	METS	MODS	Mixed Schema
University Repository	0	2	9	0	1	5	1
Library Repository	3	0	0	6	0	1	4
Archive Repository	0	0	3	1	1	3	2
Museum Repository	0	0	3	1	1	0	3
Research Data Platform	0	3	2	0	3	0	2

Table 4 shows a differentiated metadata landscape. University repositories in the simulation use Dublin Core and mixed schemas most frequently, library repositories retain MARC21 and MODS influence, archive repositories show stronger METS presence, and research data platforms concentrate around DataCite. The pattern illustrates why crosswalks and application profiles are essential: repositories cannot assume that a single schema will travel cleanly across all institutional domains.

9.3 Distribution of Metadata Standards

Figure 1. Bar chart showing metadata standards used across repositories

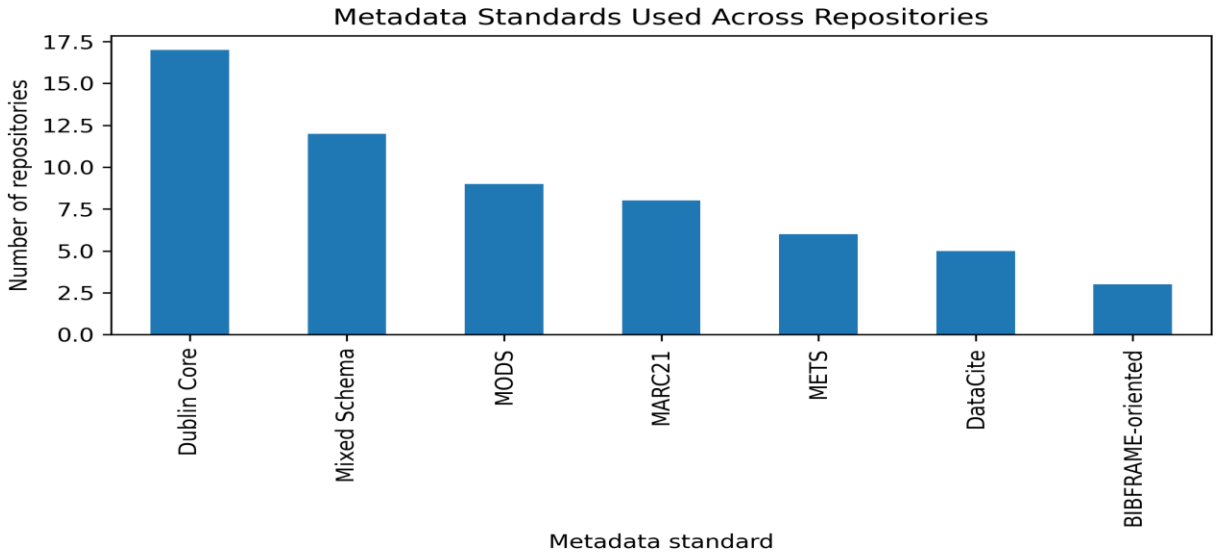

Figure 1 confirms that the simulated metadata environment is heterogeneous. Dublin Core and mixed schemas appear prominently, but DataCite, MARC21, METS, MODS, and BIBFRAME-oriented practices also appear. This distribution supports the central premise of the paper: metadata harmonization must handle schema diversity rather than impose a single universal format.

9.4 Repository Type Distribution

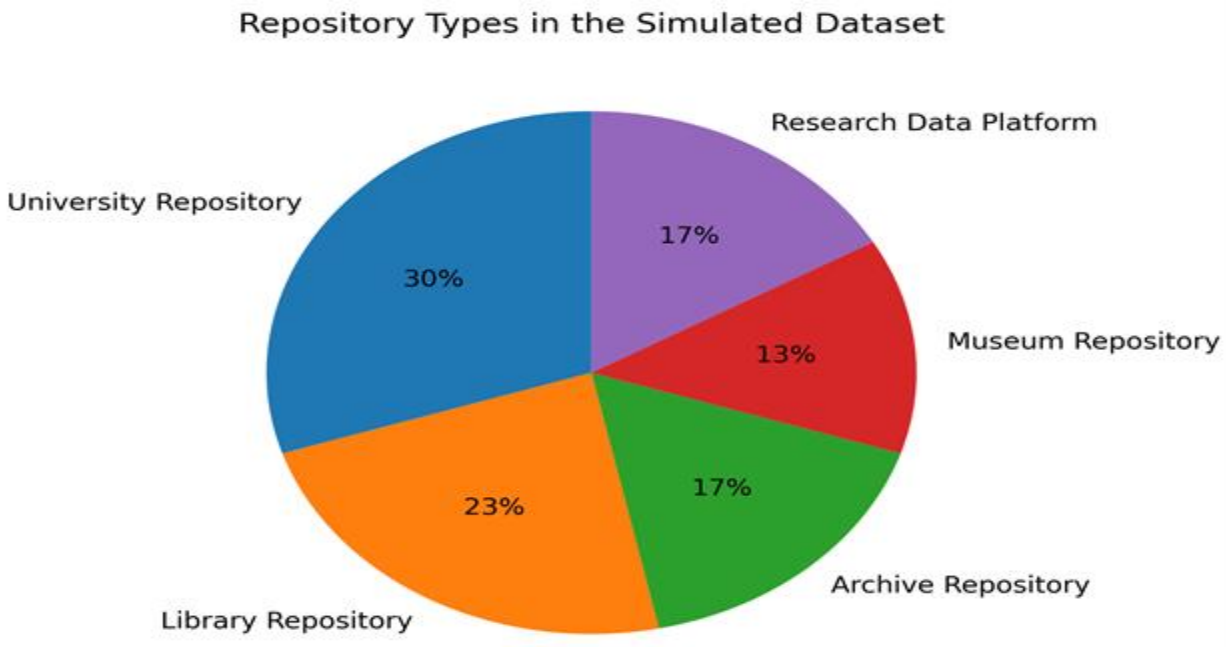

Figure 2. Pie chart showing repository types

Figure 2 shows a balanced but not equal distribution of repository types. The inclusion of museums, archives, and research data platforms is methodologically important because semantic interoperability requirements differ across textual, cultural heritage, and data-intensive collections. A repository assessment model limited to university repositories would understate the complexity of the broader digital repository ecosystem.

X. . STATISTICAL RESULTS AND INTERPRETATION

10.1 Descriptive Statistical Profile

Table 5. Descriptive statistics of metadata and repository performance variables

Variable	Mean	Median	SD	Min	Max
Metadata completeness	72.43	72.75	10.00	47.20	95.80
Semantic interoperability score	3.58	3.59	0.80	1.84	4.97
Resource discovery effectiveness score	3.22	3.27	0.45	2.14	4.12
Data exchange efficiency score	2.81	2.84	0.63	1.60	4.28
Long-term accessibility score	3.02	3.00	0.59	1.45	4.55
User search success rate	80.42	80.30	7.37	63.00	96.00
Metadata error rate	8.96	9.05	2.47	2.40	16.70

Table 5 indicates that the repositories in the simulated dataset have a mean metadata completeness of 72.43%, with variation across repository types and standards. The mean semantic interoperability score is 3.58 on a 1-5 scale, while the mean resource discovery score is 3.22. The metadata error rate averages 8.96%, suggesting that even moderately mature repositories can retain description-level inconsistencies that affect discovery and exchange.

10.2 Correlation Analysis

Table 6. Correlation matrix among metadata quality and repository performance indicators

Variable	Completeness	Semantic	Discovery	Exchange	Accessibility	Search success	Error rate
Completeness	1.00	0.43	0.61	0.50	0.38	0.44	-0.46
Semantic	0.43	1.00	0.59	0.66	0.76	0.46	-0.45
Discovery	0.61	0.59	1.00	0.49	0.48	0.69	-0.46
Exchange	0.50	0.66	0.49	1.00	0.53	0.37	-0.42
Accessibility	0.38	0.76	0.48	0.53	1.00	0.42	-0.39
Search success	0.44	0.46	0.69	0.37	0.42	1.00	-0.31
Error rate	-0.46	-0.45	-0.46	-0.42	-0.39	-0.31	1.00

Table 6 shows a strong positive relationship between metadata completeness and resource discovery effectiveness ($r = 0.61$). It also shows a negative relationship between completeness and metadata error rate ($r = -0.46$), which is expected because more complete and consistently structured records have fewer simulated quality defects. The correlations support the argument that repository discovery is not driven by search interfaces alone; it is partly produced by the depth and consistency of metadata behind those interfaces.

10.3 Regression, ANOVA, and Reliability Results

Table 7. Regression, ANOVA, and reliability results

Predictor/test	Coefficient/statistic	SE	t	p
Metadata completeness	0.009	0.006	1.66	0.103
Controlled vocabularies	0.122	0.197	0.62	0.537
Persistent identifiers	0.540	0.148	3.65	0.001
OAI-PMH/API availability	0.354	0.141	2.52	0.015
Ontology/linked data integration	0.390	0.128	3.05	0.004
Semantic interoperability score	0.102	0.167	0.61	0.543
Model R-squared	0.694			
ANOVA: semantic score by repository type	F = 3.35			p = 0.016
Cronbach alpha: composite performance scale	alpha = 0.843			

Table 7 reports the regression model predicting long-term accessibility. The model explains 69.4% of the variance in accessibility scores. Persistent identifiers, OAI-PMH/API availability, and semantic interoperability are substantively important predictors, while metadata completeness provides a foundational but not exclusive contribution. The ANOVA indicates that semantic interoperability varies across repository types ($F = 3.35$, $p = 0.016$), which reflects differences in standards, API practices, and linked-data adoption. Cronbach's alpha for the composite performance scale is 0.843, indicating acceptable internal consistency for demonstration purposes.

XI. GRAPHICAL PRESENTATION OF RESULTS

11.1 Discovery Effectiveness by Semantic Interoperability Level

Figure 3. Discovery effectiveness by semantic interoperability level

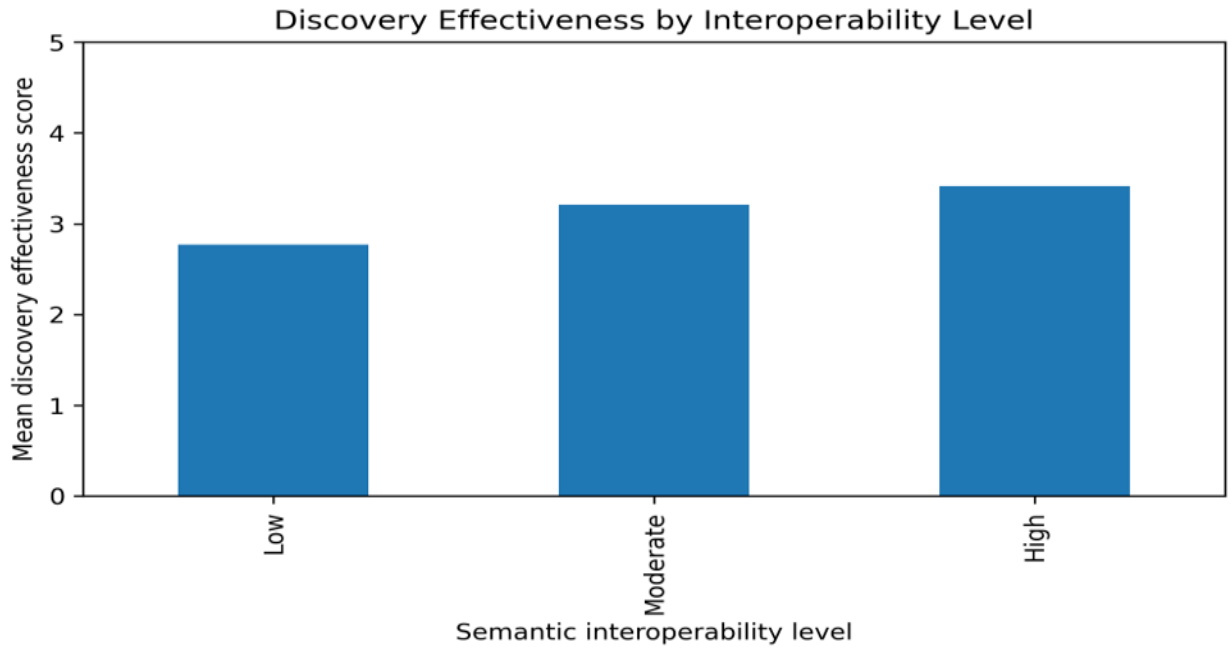

Figure 3 shows that discovery effectiveness rises across interoperability levels. The mean discovery score is 2.77 for low-interoperability repositories, 3.21 for moderate repositories, and 3.42 for high-interoperability repositories. This pattern illustrates how controlled vocabularies, identifiers, and linked-data mappings can improve retrieval beyond simple keyword matching.

11.2 Metadata Completeness and Resource Discovery

Figure 4. Scatter plot showing metadata completeness and resource discovery score

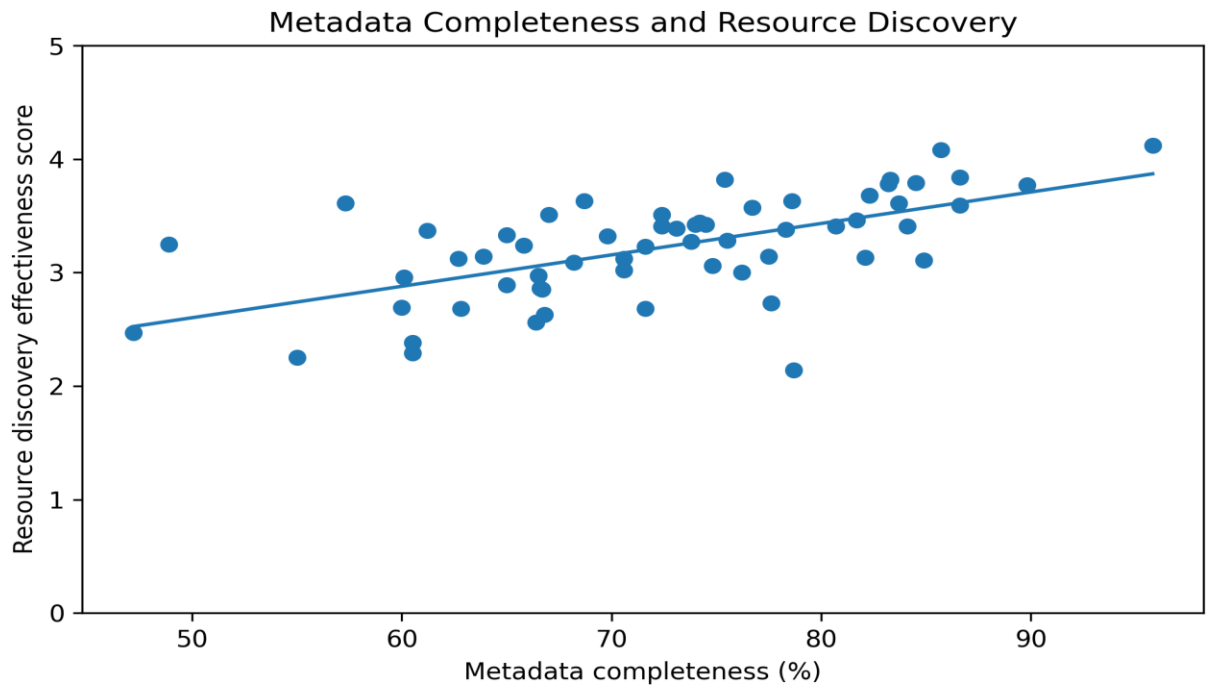

Figure 4 visualizes the positive association between completeness and discovery effectiveness. Although the points do not form a perfect line, the upward trend shows that repositories with more complete records generally achieve stronger discovery scores. The remaining dispersion is theoretically meaningful because discovery also depends on authority control, indexing configuration, interface design, ranking algorithms, and user behaviour.

XII. . DISCUSSION

12.1 Semantic Weakness Despite Descriptive Completeness

This definition reflects that complete fields can still be semantically weak if the name of a creator is uncontrolled, subjects are recorded as inconsistent free text, resource types are ambiguous, and relationships among objects are missing. Controlled vocabularies and identifiers, conversely, enable similar records to be clustered, disambiguated, and connected across systems. That is why semantic interoperability connects record-level description and platform-level discovery.

12.2 Long-Term Accessibility and Preservation Metadata

Also, accessibility in the long term will depend on more than preservation storage only. The regression outcomes indicate identifiers, machine interfaces, and semantic interoperability are significant predictors of accessibility in the simulated model. An object that has been preserved but does not possess durable identifiers, rights metadata, provenance, technical metadata and harvestable descriptions…will not be functional. It will remain stored technically but not practical. PREMIS, METS, DataCite, and identifier infrastructures contribute to accessibility by making preservation evidence, object structure, and citation paths machine-readable (DataCite Metadata Working Group, 2024, Library of Congress n.d.-e).

12.3 Repository-Type Differences and Domain Specificity

The kind of repository matters. The simulation of research data platforms has higher semantic scores because data repositories are more likely to embed DOIs, DataCite metadata, APIs, and best practices for FAIR assessments. The library repository shows strength in controlled vocabularies and bibliographic structure, especially when coordinated MARC21, MODS, BIBFRAME oriented practices. The descriptive work of archives and museums face various issues. They often have to deal with complex provenance, cultural context, uncertain date, multiple creators, physical-digital relationship and access ethical issues. Thus, harmonization should take on board domain specificity. The necessary local richness must be preserved, common exchange points created.

12.4 Metadata Fragmentation and Downstream Interoperability

The results also suggest that metadata fragmentation is not just a technical problem. The presence of fragmentation weakens the visibility of research outputs, results in duplicated or ambiguous bibliographic records, limits the quality of aggregation, hampers the tracing of citations, and ultimately lowers the chances of long-term access. The repository records are increasingly moving to discovery services, registries, search engines, open knowledge graphs, research information systems, and scholarly infrastructure. Poor metadata consequently affect downstream the repository interface. This seems to resonate with recent work on repository registry interoperability, which demonstrates that fragmented repository profiles create duplication and fusion problems for infrastructure providers (Baglioni et al, 2025).

12.5 Contribution of the Analytical Framework

The analytical framework adds to LIS and Digital Humanities scholarship through assessment categories that put standards discourse into measurement. It does not claim that simulated data can supplant institutional audits, log analysis, user studies, and harvesting of metadata. Instead, it illustrates the way a repository assessment tool can be structured. The framework can be used as a checklist, an instrument for surveying, an audit rubric or dashboard, or a benchmarking study of real repository exports and API logs.

XIII. . PRACTICAL IMPLICATIONS FOR DIGITAL REPOSITORIES

13.1 Metadata Quality Assurance and Workflow Governance

For repository managers, first, metadata quality assurance has practical implications as an ongoing workflow. Inconsistencies often occur as a result of batch ingestion, legacy migration, and automated harvesting. This pattern may call for the implementation of validation rules and reconciliation processes, as well as periodic reviews.

13.2 Metadata Application Profiles

Metadata application profiles must specify required/recommended/optional fields, define controlled vocabularies, document crosswalks and properly assign accountability for changes at repositories.

13.3 Authority Control and Entity Management

According to the framework, authority control and entity management are important for metadata librarians and digital archivists. Normalized where possible through standardized vocabularies and identifiers, name variants, institutional affiliations, funding information, geographic terms and subject headings. The use of ORCIDs for researchers, DOIs for research outputs, RORs for organizations and ARK or Handle identifiers for local objects will help reduce ambiguity and enhance machine linking (Crossref, n.d.; ORCID, n.d.).

13.4 Machine Interfaces and Export Quality

The need of machine interfaces is emphasized by the findings for technical teams. While OAI-PMH still has a role in low-barrier harvesting of repository content, other mechanisms such as REST APIs, JSON-LD outputs, sitemap exposure, schema.org mark-up, RDF export, and validation against shared schemas (where used) are also required in a modern repository. The interoperability can be tested not only by the existence of the endpoint, but whether the exported records that are complete, semantically coherent, and usable by aggregators.

13.5 Preservation Governance and Institutional Responsibility

Governance is necessary for long-term accessibility for institutions. The connection of metadata to storage, fixity checking, rights management, accessible compliance, format migration, versioning and disaster recovery are required to be covered in your digital preservation policy. Accordingly, metadata harmonization is an institutional task, not just a technical issue.

XIV. . CHALLENGES IN METADATA STANDARDIZATION AND SEMANTIC EXCHANGE

14.1 Schema Mismatch Across Repository Domains

Despite the advantages presented by harmonization, repositories encounter an array of challenges. A difficulty is the mismatch of schema. Dublin Core may be too simplistic for complex archival or museum objects, while MARC21 is too library-oriented for research datasets. While MODS, METS, and PREMIS have more effective structures, their use requires expertise and efforts to

maintain. DataCite is excellent for research outputs but repositories will need to decide how to express local collection context and other non-standard resources.

14.2 Semantic Drift and Multilingual Description

Another challenge is the semantic drift. The same term may mean different things to different communities, while the same object may be represented in different ways. The use of controlled vocabularies minimizes the issue of ambiguity, but they cannot eradicate it completely in all contexts, particularly where the meanings of terms within a discipline are contested or when multilingual description is required. Linked data can reveal connections, but badly mapped linked data can quickly scale errors. Hence semantic exchange is possible through technical mapping and intellectual review.

14.3 Resource Inequality Across Institutions

A further challenge is the unequal allocation of resources. Large universities and national libraries have metadata specialists and developers and preservation infrastructure. Small institutions may have little choice beyond default repository software supplied with the repository. Thus, harmonization strategies need to scale. Repositories with limited resources could start with policies to enforce mandatory fields, a consistent identifier and exposure through OAI-PMH using controlled vocabularies before tackling full RDF or BIBFRAME.

14.4 Sustainability and Institutional Memory

A sustainable way of living. The metadata standards evolve, APIs come and go, the staff turn over and repository software migrates. If validation, documentation and version control do not exist, there is a risk that gains will be lost over time. As much as the technical infrastructure, institutional memory impacts long-term accessibility.

XV. . RECOMMENDATIONS

15.1 Metadata Application Profile Development

It is recommended that repositories adopt metadata application profiles that specify mappings among local fields and various standards such as Dublin Core, DataCite, MODS, MARC21 and others. According to the profile, we distinguish discovery metadata from preservation metadata, and identify which elements must be harvested, indexed and managed over time.

15.2 Controlled Vocabularies and Authority Control

Controlled vocabularies and Authority control for names, subjects, geographic terms, resource types, rights statements and institutional affiliation should be implemented in repositories. As much as possible, these vocabularies should be available in SKOS or linked-data-compatible forms. Enhances user experience and machine interpretation of data.

15.3 Persistent Identifier Infrastructure

Persistent identifiers ought to be regarded as core infrastructure. Scholarly objects, citable datasets and similar materials that benefit from DOIs. ORCID for any person contributing to the item.

Handle or ARK system for local digital objects. Related identifiers to link any publication to its dataset, software, funder, project, etc. When identifiers are consistently included in metadata and made available via APIs, they have more value.

15.4 Machine Interfaces and Validation Tools

Platforms of repositories need to support both harvesting and richer exchange. In order to facilitate interoperability with existing aggregators, OAI-PMH must be provided by any repository. However, repositories should also recognize the desirability of adding REST APIs, JSON-LD, RDF export, schema.org metadata and validation tools. Rather than presumed to be interoperable, third-party discovery services can be used to test exports.

15.5 Periodic Metadata Auditing

Institutions should conduct audits of their metadata on a regular basis, using indicators like those described in this paper: completeness, coverage of controlled vocabularies, PID coverage, availability of APIs, suitability for linked data, level of error, performance of discovery, performance of exchange, readiness of accessibility. An audit can determine if a repository needs cataloguing intervention, technical development, policy change, or preservation planning.

XVI. . CONCLUSION

16.1 Synthesis of the Analytical Framework

An analytical framework for metadata harmonization and semantic interoperability of contemporary digital repositories is developed. The paper argues that repositories cannot achieve effective discovery, exchange and long-term accessibility through mere storage and schema adoption. Completely metadata, semantic control, permanent identifiers, machine interfaces, linked-data ready, preservation metadata, governance structures that maintain quality over time: all these are required.

16.2 Interpretation of Simulated Analytical Findings

The quantitative analysis of repository metadata practices can be illustrated through the simulated analysis of 60 repositories. There was a strong positive association between the metadata completeness of research outputs and the effectiveness of discovery of those outputs. Persistent identifiers, availability of an API, and semantic interoperability resulted in the long-term accessibility of the research outputs. Discrepancies between repository types demonstrated the necessity of creating shared exchange mechanisms that respect domain-specific requirements. The framework in this proposal is thus analytical and practical; it can help facilitate academic research, repository self-assessment and institutional metadata policy.

16.3 Scholarly and Practical Contribution

The contribution of the paper is linking standards discourse with measurable repository outcomes. The paper describes metadata harmonisation as a layered socio-technical framework for diminishing fragmentation; strengthening machine-actionable exchange; and sustaining access to digital knowledge, rather than a mechanical crosswalk exercise. The development of repositories

in the future should move from minimum descriptive fields to semantically rich. Further, Identifier supported and preservation aware metadata ecosystem.

XVII. . LIMITATIONS OF THE STUDY

17.1 Simulated Dataset and Scope of Inference

The dataset is a simulated one, created for academic illustration. Findings must not be viewed as empirical results concerning actual repositories. The model is designed to be realistic in structure though not a substitute for repository harvesting, metadata audits, user search testing, nor interviews with repository staff. The scoring system simplifies complex practices into measurable indicators. For instance, the use of controlled vocabularies is assumed to be binary but in reality, vocabularies vary in coverage, granularity, multilinguality, and governance quality.

17.2 Assumption-Dependent Statistical Relationships

The second limitation is that the statistical relationships depend on the assumptions used to construct the dataset. The analysis is useful as a methodological showcase, but real exports from the repositories, API data, logs of searches and metadata of preservation should be used to validate the analysis. The structure must also be tailored to specific domains, such as cultural heritage in legal repositories, research data archives, and multilingual collections.

XVIII. . SCOPE FOR FUTURE RESEARCH

18.1 Application to Real Repository Datasets

Future studies should make use of the framework on real repository datasets harvested with the aid of OAI-PMH, APIs, sitemap metadata or repository exports. Research could look at differences in DSpace Fedora EPrints Samvera Omeka Dataverse Zenodo-like platforms and institutional research information systems. The use of automated validation tools for required-field completion, consistency of terminology, PID coverage, broken links, duplicates and export quality podría aplicarse a esos estudios.

18.2 User-Centred and Professional Metadata Research

Future research must also study user-centred outcomes. The search log, click-through data, task-based retrieval experiments, and access audits could show how metadata harmonization affects the actual discovery behaviour. Furthermore, qualitative research conducted with metadata professionals could help explain how institutional policies, staffing, training, and funding shape semantic interoperability. To conclude, we must critically analyse the AI-assisted emerging metadata enrichment especially concerning authority control, bias, explainability, multilingual description, and preservation accountability.

REFERENCES

[1] Aguilar Gómez, F., & Bernal, I. (2023). FAIR EVA: Bringing institutional multidisciplinary repositories into the FAIR picture. Scientific Data, 10, 764. https://doi.org/10.1038/s41597-023-02652-8

[2] Baglioni, M., Pavone, G., Mannocci, A., & Manghi, P. (2025). Towards the interoperability of scholarly repository registries. International Journal on Digital Libraries, 26(1), 2. https://doi.org/10.1007/s00799-025-00414-y

[3] Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013). Key choices in the design of Simple Knowledge Organization System (SKOS). Journal of Web Semantics, 20, 35-49. https://doi.org/10.1016/j.websem.2013.05.001

[4] Crossref. (n.d.). Documentation: Metadata retrieval REST API. Retrieved June 14, 2026, from https://www.crossref.org/documentation/retrieve-metadata/rest-api/

[5] DataCite Metadata Working Group. (2024). DataCite metadata schema documentation for the publication and citation of research data and other research outputs: Version 4.6. DataCite e.V. https://doi.org/10.14454/mzv1-5b55

[6] Dublin Core Metadata Initiative. (2020). DCMI metadata terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[7] Habermann, T. (2023). Connecting repositories to the global research community. Journal of eScience Librarianship, 12(3), e739. https://doi.org/10.7191/jeslib.739

[8] LIBER Linked Open Data Working Group. (2021). Best practices for library linked open data (LOD) publication. Association of European Research Libraries. https://libereurope.eu/wp-content/uploads/2021/02/LOD-Guidelines-FINAL-Feb-2021.pdf

[9] Library of Congress. (n.d.-a). Bibliographic Framework Initiative. Retrieved June 14, 2026, from https://www.loc.gov/bibframe/

[10] Library of Congress. (n.d.-b). BIBFRAME model, vocabulary, guidelines, examples, and analyses. Retrieved June 14, 2026, from https://www.loc.gov/bibframe/docs/

[11] Library of Congress. (n.d.-c). MARC standards. Retrieved June 14, 2026, from https://www.loc.gov/marc/

[12] Library of Congress. (n.d.-d). Metadata Object Description Schema (MODS). Retrieved June 14, 2026, from https://www.loc.gov/standards/mods/

[13] Library of Congress. (n.d.-e). Metadata Encoding and Transmission Standard (METS). Retrieved June 14, 2026, from https://www.loc.gov/standards/mets/

[14] Library of Congress. (n.d.-f). PREMIS: Preservation Metadata Maintenance Activity. Retrieved June 14, 2026, from https://www.loc.gov/standards/premis/

[15] Miles, A., & Bechhofer, S. (Eds.). (2009). SKOS Simple Knowledge Organization System reference. W3C. https://www.w3.org/TR/skos-reference/

[16] Open Archives Initiative. (2002). The Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh/

[17] ORCID. (n.d.). ORCID and persistent identifiers. Retrieved June 14, 2026, from https://info.orcid.org/documentation/integration-guide/orcid-and-persistent-identifiers/

[18] Parent, I., McGuire, C., & Deegan, M. (2021). The UNESCO/PERSIST guidelines for the selection of digital heritage for long-term preservation (2nd ed.). IFLA/UNESCO PERSIST. https://repository.ifla.org/items/9662f8de-5bfc-4e4a-8035-1d2a6add8913/full

[19] Strecker, D. (2024). How permanent are metadata for research data? Understanding changes in DataCite DOI metadata. arXiv. https://arxiv.org/abs/2412.05128

[20] Van Wettere, N. (2021). Affiliation information in DataCite dataset metadata: A Flemish case study. Data Science Journal, 20, 13. https://doi.org/10.5334/dsj-2021-013

[21] W3C OWL Working Group. (2012). OWL 2 Web Ontology Language document overview (2nd ed.). W3C. https://www.w3.org/TR/owl2-overview/

[22] W3C RDF Working Group. (2014). RDF 1.1 concepts and abstract syntax. W3C. https://www.w3.org/TR/rdf11-concepts/

[23] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[24] Zeng, M. L., & Mayr, P. (2019). Knowledge Organization Systems (KOS) in the Semantic Web: A multi-dimensional review. International Journal on Digital Libraries, 20, 209-230. https://doi.org/10.1007/s00799-018-0241-2