MetaFAIR Ecosystem for SSH Digital Libraries: Dataset Packaging, Validation, and SPARQL Publication

Emiliano Degl'Innocenti[1,†], Francesco Pinna[1,†], Alessia Spadi[1,†] and Federica Spinelli[1,†]

[1]National Research Council of Italy, Opera del Vocabolario Italiano (CNR-OVI), Via di Castello 46, 50141, Florence, Italy

Abstract

Data management in the Social Sciences and Humanities (SSH) domain is challenged by the heterogeneity of data, standards, and practices across disciplines. This contribution presents a platform-oriented approach designed to support data curation, interoperability, and reuse through structured workflows and shared architectural principles. Rather than imposing a single data model, MFE allows different community standards and institutional practices to coexist, while ensuring that datasets meet common requirements for quality, documentation, and reuse. The platform supports the entire data lifecycle, from ingestion and validation to publication and long-term access, and connects dataset documentation, metadata, and digital assets into a coherent management framework. MFE is designed as a modular and scalable system. By providing standard interfaces and clear governance mechanisms, it facilitates data management across projects, institutions, and disciplines. The paper discusses how this approach contributes to strategic goals such as sustainability, interoperability, and responsible data stewardship, and illustrates how digital library infrastructures can evolve into data-centric platforms that support reuse, collaboration, and long-term value creation in the SSH domain.

Keywords

data management, FAIR data, data architecture, knowledge organization

1. Introduction

1.1. Context

Over the last decades, the SSH (Social Science and Humanities) and the GLAM (Galleries, Libraries, Archives and Museums) sectors have undergone an intense process of "datization".[1] Large-scale programmes and project-driven initiatives have produced a growing number of digital collections, including digitised manuscripts, archival records, museum objects as well as born-digital corpora and research datasets. Scientific digital libraries in the DH field have emerged as a way to share these outcomes and to create linked information spaces that support complex forms of exploration and analysis [1].

However, the mere availability of digital collections online does not automatically translate into effective reuse. Aggregator-type digital libraries that harvest content from multiple repositories routinely face issues such as heterogeneous formats, incompatible descriptive practices, and inconsistent metadata quality, all of which hinder cross-collection discovery and interoperability [2], thus limiting their potential for interdisciplinary research. In SSH, this situation is exacerbated by the structural fragmentation of data formats and schemas, semantic heterogeneity across disciplines, and uneven adoption of community standards, which collectively undermine full adherence to the principles of findability, accessibility, interoperability and reusability expressed by the acronym FAIR.[2] The FAIR

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

[†]These authors contributed equally.

emiliano.deglinnocenti@cnr.it (E. Degl'Innocenti); francesco.pinna@cnr.it (F. Pinna); alessia.spadi@cnr.it (A. Spadi); federica.spinelli@cnr.it (F. Spinelli)

0000-0002-3839-9024 (E. Degl'Innocenti); 0009-0009-4925-3467 (F. Pinna); 0000-0002-9670-3426 (A. Spadi); 0000-0002-2195-3930 (F. Spinelli)

[1]On the concept of datization see Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

[2]https://www.go-fair.org/fair-principles/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-08-20

guidelines themselves explicitly stress that data should be not only accessible to humans but also machine-actionable, enabling automated discovery and integration in computational workflows [3].

In response, digital libraries are designed not as static portals exposing "digital surrogates", but as infrastructures that support data-intensive research and computational access to collections. The "Collections as Data" movement exemplifies this shift, reframing cultural heritage holdings and scholarly corpora as structured, well-documented datasets intended for computational reuse rather than purely for human viewing [4]. This evolution places new demands on the underlying architectures of digital libraries, which must now orchestrate workflows for ingestion, validation, semantic enrichment and publication of heterogeneous SSH resources.

Within this broader landscape, research infrastructures such as the Digital Research Infrastructure for the Arts and the Humanities (DARIAH ERIC[3]) and its national node DARIAH-IT[4] play a key role in providing distributed, standards-based environments that bring together data, tools and workflows for the arts and humanities [5]. The MetaFAIR Ecosystem (MFE), developed for DARIAH-IT within the Humanities and Cultural Heritage Italian Open Science Cloud (H2IOSC) project,[5] is situated precisely at this intersection: it is conceived as a digital ecosystem that treats SSH datasets as structured collections, supports their FAIRification through semantic workflows, and exposes them as interoperable resources, also allowing their reuse in other digital library and open science environments.

1.2. Problem statement

Despite the growing availability of digital collections in the SSH domain, the underlying data remain highly heterogeneous. In fact, Cultural Heritage Institutions, as well as many research projects and working groups, employ a wide range of metadata schemas and formats, such as Dublin Core (DC), Encoded Archival Description (EAD), Metadata Object Description Schema (MODS), the Europeana Data Model (EDM), or domain-specific XML schemas, often combined with extensive local customisations [6, 7]. The result is frequently a fragmented landscape in which similar entities (*i.e.*, persons, places, dates, titles of literary works or other artefacts) are described in incompatible ways, identifiers are reused or omitted inconsistently, and relations between digital objects, their physical counterparts and contextual entities are encoded with varying levels of granularity. This structural and semantic heterogeneity complicates even basic operations such as cross-collection search and filtering, with the consequence that often the machine-actionability required by the FAIR principles is undermined [8].

At the same time, research infrastructures and open science initiatives increasingly rely on federated environments, most notably the European Open Science Cloud (EOSC)[6] and its disciplinary clusters, such as the SSH Open Cloud (SSHOC)[7] for Social Science and Humanities, to provide unified access to distributed services, datasets and tools[8] [9].

In such settings, collections and datasets must be exposed through interoperable interfaces and shared conventions to be discoverable, composable and reusable alongside resources from other communities. However, current practices in data publication for the SSH and Cultural Heritage domains still tend to prioritise repository-specific or project-specific arrangements over federation-ready, standards-aligned solutions. As documented in recent analyses of EOSC interoperability, gaps persist at all levels (technical, semantic, organisational) that limit the seamless integration of domain-specific repositories into larger federated ecosystems[9] [10].

[3] European Research Infrastructure Consortium

[4] http://dariah.cnr.it/

[5] H2IOSC is a project funded by the European Union Next Generation EU and the Italian Ministry of University and Research as part of the National Recovery and Resilience Plan (NRRP). The project enhances and federates CLARIN-IT, DARIAH-IT, E-RIHS.it, and OPERAS-IT, the four Italian nodes of the European Research Infrastructures in the Social Sciences and Humanities domain, providing an open and integrated environment for advanced research on digital data and cultural heritage. https://www.h2iosc.cnr.it/

[6] https://eosc.eu/

[7] https://sshopencloud.eu/

[8] https://sshopencloud.eu/about-sshoc

[9] https://eosc.eu/task-force-deliverab/a-landscape-overview-of-the-eosc-interoperability-framework-capabilities-and-gaps

Within this context, the key challenge is how to design algorithms and system architectures that can ingest SSH datasets originating from diverse institutions and projects, validate and normalise their internal structure, and transform them into interoperable, queryable digital library collections. Addressing this challenge requires, on the one hand, formal models and validation procedures capable of dealing with heterogeneous data and, on the other, workflow-oriented infrastructures that support the end-to-end FAIRification, semantic enrichment and publication of datasets within federated research and open science environments.

In this paper, we present the approach to managing heterogeneous data originating from multiple domains within the digital infrastructure developed for DARIAH-IT as part of the H2IOSC project: the MetaFAIR Ecosystem (MFE). Conceived as a comprehensive environment for the cataloguing, management, and publication of digital resources in the GLAM sector, MFE builds on the technical foundations of metaFAD, an open-source system originally developed by GruppoMeta (now Dedagroup) for the management of cultural-heritage digital assets. This foundation provides a modular architecture, consolidated document-management workflows, and core components such as a digital asset management system, multi-domain governance mechanisms, and an indexing engine.

Beyond this technical layer, MFE reflects the strategic vision and coordinated design effort led by the Italian Node of the Digital Research Infrastructure for the Arts and the Humanities (DARIAH-IT). Through extensive engagement with diverse communities and continuous monitoring of the evolving digital landscape across European, national, and regional initiatives, DARIAH-IT shaped the system's conceptual framework and functional priorities, ensuring that it responds to real and emerging needs within the social science and humanities research domain.

Developed within the H2IOSC project, MFE is designed to support the full data lifecycle in alignment with FAIR principles. By offering a unified environment for the cataloguing, preservation, and dissemination of digital resources, it provides dedicated tools, documentation, and specialised services that enhance data stewardship practices. Its modular and scalable architecture ensures adaptability to a wide range of technical and disciplinary requirements, making the system suitable for diverse research and cultural-heritage contexts.

1.3. Objectives

Considering the challenges outlined above, this work addresses the following overarching question: which workflow steps and associated validation and mapping algorithms are required to support end-to-end FAIRification and semantic publication of SSH datasets? Concretely, we are interested in the design of reusable, programmable workflows that (i) accept heterogeneous, locally defined schemas and packaging conventions; (ii) validate their structural coherence and completeness according to explicit, machine-checkable rules; (iii) map them onto shared semantic models and ontologies; and (iv) expose the resulting knowledge graphs through standardised interfaces suitable for integration into federated digital library and open science environments.

To answer this question, the paper proposes and analyses the MFE as a multilayer system architecture that functions as a digital library back-end for SSH data. Building on previous work on FAIR workflows and semantic interoperability in the humanities and cultural heritage domains, MFE combines a formal dataset packaging model with a workflow-oriented infrastructure for ingestion, validation, semantic mapping, and publication. As such, it contributes (i) an architectural blueprint for a Digital Library (DL) oriented FAIRification pipeline tailored to the needs of SSH and GLAM collections, and (ii) a concrete, implemented environment in which heterogeneous datasets can be transformed into interoperable, queryable collections that have the potential to be reused in other Open and Data Science platforms.

1.4. Structure of the paper

In §1 we introduced the main issue we want to address in this contribution; the following Sections of the paper are organised as follows. §2 describes the requirements and design goals that a digital-library back-end must satisfy to support end-to-end FAIRification in the SSH domain, distinguishing between

functional requirements (*e.g.*, schema support, dataset completeness, provenance and licensing) and non-functional requirements (*e.g.*, scalability and interoperability). §3 presents the workflow algorithms that operationalise FAIRification and semantic publication in the MFE, describing the main phases from ingestion and validation to semantic mapping, RDF generation, and publication through triplestore-based query interfaces. §4 then introduces the multilayer MFE architecture that implements these workflow phases at scale, detailing the interaction among data management, semantic processing, digital asset management, editorial workflows, and external integrations within the broader DARIAH-IT and H2IOSC environment. Finally, §5 summarises the main outcomes and outlines directions for future developments, with particular attention to extending schema/ontology coverage and strengthening cross-domain interoperability in federated research settings.

2. Requirements and design goals

2.1. Functional requirements

From a functional perspective, the MetaFAIR Ecosystem must first and foremost accommodate the structural and semantic diversity that characterises datasets in the SSH and in the GLAM sector. In practice, this means supporting both widely adopted community standards (Dublin Core, EAD, EAC-CPF) and the DataCite Metadata Schema for research data, as well as project-specific or institution-specific schemas. Rather than enforcing a single descriptive model, MFE must provide mechanisms to register, reference and process multiple schemas, so that heterogeneous collections can be ingested without prior "flattening" into a lowest common denominator.

Secondly, each dataset ingested into the MFE must be fully described through structured metadata, in line with the FAIR principles that emphasise rich, machine-actionable descriptions as a prerequisite for discovery and reuse. At a minimum, this includes persistent identifiers, titles, creators, temporal and spatial coverage, subject information, and explicit links to related resources, following established best practices for cultural heritage and research data. These metadata records must be systematically associated with the corresponding digital assets (files such as images, textual transcriptions, audio-visual material or tabular data) and with the containers that group them into coherent digital objects (for example, all media and documentation about a single archival record or bibliographic object). This explicit linkage between descriptions, containers and bitstreams is essential to support digital library operations such as object-level retrieval, navigation and export. Digital records stored in the MFE can be exported in XML format using the same schema selected at the time of data ingestion. This ensures a consistent and traceable lifecycle, from the initial import of raw data to its final export in a mapped and coherent schema.[10]

Thirdly, every dataset must be accompanied by paradata, providing a transparent account of the processes, decisions, transformations, and methodological assumptions that shaped the data throughout its entire lifecycle, *e.g.*: provenance documentation that explains how the data were produced, curated and transformed, including information on sources, digitisation workflows, data cleaning and enrichment steps, and known limitations or biases. Such documentation is increasingly recognised as a key component of reusable cultural heritage and research datasets, for instance, in emerging "datasheet" practices.[11] In the context of MFE, this material needs to be captured in a structured, machine-readable form so that it can be associated with the corresponding dataset and exposed through the same interfaces as the core metadata.

Finally, each dataset must carry an explicit licence and access profile that governs its reusability. Rights and licence information are a mandatory element in contemporary data citation schemas such

[10] See MFE's presentation images: https://github.com/fpinnahub/DataForPapers/raw/refs/heads/main/MFE_images.zip (retrieved February 5, 2026).

[11] The term "datasheet" here refers to the *Datasheets for Datasets* proposal by [11], which advocates attaching a structured, human- and machine-readable document to each dataset, analogous to hardware datasheets in electronics. Such datasheets systematically record the dataset's motivation, composition, collection and processing methods, intended uses, limitations, and maintenance plans, thereby improving transparency, accountability, and informed reuse.

as DataCite, and they are directly connected to the "Accessible" and "Reusable" dimensions of the FAIR guidelines. Functionally, MFE must ensure that (i) every dataset declares a licence or rights statement; and (ii) the declared access conditions are enforced consistently across the access channels provided by the system, including human-oriented interfaces, application programming interfaces (APIs) and bulk download mechanisms. This enables digital library front-ends and external services to consume MFE-hosted datasets with clear information about how they may be accessed, cited and reused.

2.2. Non-functional requirements

Beyond these functional aspects, the MetaFAIR Ecosystem (MFE) must satisfy a set of non-functional requirements that reflect its intended role as a DL back-end within national and European research infrastructures.

A first requirement concerns scalability. MFE is expected to handle an increasing number of datasets and collections, each potentially comprising thousands or millions of digital objects and associated assets. Consequently, the overall architecture must support incremental ingestion and re-ingestion workflows without incurring prohibitive overheads; validation and transformation procedures must remain tractable as dataset size grows. This calls for stateless or loosely coupled services that can be replicated. In addition, the system must be able to cope with evolving schemas and mappings: as new metadata standards are introduced or existing ones are refined, the underlying storage and workflow components should accommodate schema evolution without requiring disruptive migrations or re-engineering of the whole platform. From a DL perspective, this scalability requirement applies not only to storage and computing but also to query performance, ensuring that search and exploration over semantically enriched collections remain responsive under realistic workloads.

A second non-functional requirement is interoperability with existing and emerging digital components of the research ecosystem. On the data layer, MFE must expose its semantically enriched output through standard triplestore technologies and query languages, most notably Resource Description Framework (RDF) and SPARQL, so that knowledge graphs produced by the FAIRification workflows can be queried and integrated using off-the-shelf semantic web tools. Triplestores such as OpenLink Virtuoso are widely used both in cultural heritage contexts and in EOSC-related projects as scalable back-ends for Linked Open Data publication, making compatibility with these technologies a key design goal for any DL-oriented infrastructure. On the application layer, MFE must offer interfaces that can be consumed by heterogeneous digital library front-ends and discovery services, including REST-based APIs, SPARQL endpoints, and, where appropriate, protocols compatible with harvesting and aggregation workflows.

Finally, MFE must integrate smoothly with the broader landscape of European initiatives and national implementations, such as the H2IOSC project. In this context, interoperability entails adherence to EOSC-level guidelines and profiles for service description, authentication and authorisation, and metadata exchange, so that MFE-based services can be registered, composed and monitored alongside other research services. At the same time, the system must remain compatible with the specific requirements and conventions of humanities-oriented research infrastructures, ensuring that collections FAIRified through MFE can circulate across disciplinary and national boundaries without loss of information or functionality.

3. Workflow algorithms for FAIRification and semantic publication

The workflow for FAIRification and semantic publication can be described as a sequence of sequential steps applied to heterogeneous data. The process begins with the selection of XML-based community schemas (*e.g.*, bibliographic, archival, or GLAM standards). These schemas serve as predefined structural constraints that guide the initial parsing of incoming metadata. Ontologies such as CIDOC-CRM provide the conceptual framework used to interpret these structures, while crosswalks define the transformation rules necessary for partial schema-to-schema interoperability. As a prerequisite for any downstream semantic alignment, ingestion in MFE supposes that the incoming dataset complies with a consistent

packaging model: standardised directory hierarchies and explicit mapping files must link metadata and descriptive records to media and vocabularies, so that ingestion is accurate and repeatable.

This phase corresponds to "ingestion and validation" in the MFE workflow: once a reference template is selected, users upload digital collections to the FAIR repository; the system populates predefined fields and requires, alongside item-level metadata, the completion of collection-level metadata documenting provenance, responsibility, and reuse conditions.

To harmonize heterogeneous inputs, the workflow introduces an intermediate XML schema designed as a normalised representation of the metadata. This schema constitutes a reduced and data-centric element set. Its purpose is to act as a canonical form from which downstream transformations can be computed. A set of XSD templates defines the base schema used for converting metadata between the intermediate schema and the community-specific XML schemas, enabling bidirectional and deterministic transformations.

Metadata tracks are next instantiated as parameterized templates derived from established community standards (*e.g.*, EAD, EAC, ICCD, EDM, etc.). Ingestion of a collection into a schema triggers an automated distribution algorithm: metadata fields are assigned to structural components of the schema through rule-based matching.

Provenance and legacy information are documented in the related dataset description facet, a dedicated metadata schema designed to register all collection-level descriptors. The dataset facet is compiled by the data curator and includes structured fields that capture the origin and historical context of the collection, reuse conditions, and responsibility statements. This schema ensures that each dataset is associated with a coherent and complete provenance profile, distinct from the technical metadata extracted for individual digital assets (such as images, videos, or other media files). By providing a structured framework for dataset-level metadata, the workflow supports consistent documentation, traceability, and the integration of collection-level information throughout the FAIRification and semantic publication pipeline.

Semantic alignment is achieved through mapping functions that associate schema-level elements with ontology classes and properties. Each mapping defines a set of correspondence rules, effectively functioning as a transformation dictionary between XML elements and ontology constructs. A track may be associated with zero or more mapping functions, while each mapping requires both an existing track and an ontology module. Once defined, these mappings are applied programmatically to all collections instantiated under the corresponding track, guaranteeing uniform semantic enrichment.

The mapped metadata are then converted into RDF triples through a rule-engine that interprets the mapping as a set of Subject-Predicate-Object (SPO) generation instructions. For each metadata element, the system computes one or more semantic assertions of the form (subject, predicate, object). This procedure generates a graph-based representation suitable for integration into Linked Data environments.

The final step consists of storing the resulting RDF triples in a triplestore such as OpenLink Virtuoso. The triplestore supports ingestion, SPARQL-based querying, update operations, and exports in multiple RDF serialization formats. Its graph-based storage model allows the execution of complex queries across semantically enriched datasets, enabling algorithmic retrieval, inference, and publication of Linked Open Data endpoints.

4. MetaFAIR Ecosystem architecture

Building upon the sequential workflow for FAIRification and semantic publication, the MetaFAIR Ecosystem (MFE) provides the architectural framework that operationalises these transformations at scale. Each phase of the workflow (schema selection, metadata normalization, semantic alignment, and RDF generation) is supported by a coordinated set of technical layers designed to ensure methodological consistency and full compliance with FAIR principles. MFE is conceived as a modular, service-oriented infrastructure in which several functional layers collaborate to manage the entire lifecycle of cultural heritage data.

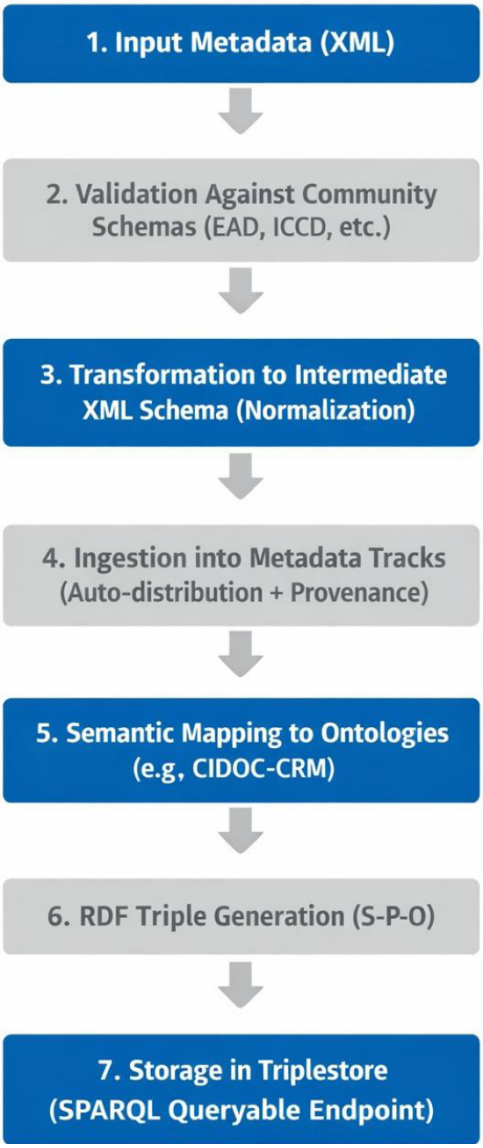

Figure 1: Workflow for FAIRification and publication of SSH metadata.

At the architectural level, the ecosystem integrates:

- a data management environment capable of handling heterogeneous metadata schemas across archival, historical-artistic, bibliographic, and other SSH domains;
- a semantic processing layer that manages ontology mappings and produces RDF triples;
- an authorities and thesauri management level which is connected to the data management layer and enriches;
- a digital asset management layer that ingests, enriches, and visualizes media objects (*e.g.*, images, videos, photos, audio, etc.);
- an editorial workflow engine ensuring controlled transitions from draft to publication;
- a dataset description layer that captures provenance, context, licensing, and responsibility at the collection level.

Together, these layers ensure that each step of the FAIRification pipeline is not merely conceptual but technically enforceable, allowing highly heterogeneous data to be processed within a unified and interoperable environment.

The internal architecture reflects this collaborative logic. The Core MFE Data Management Layer orchestrates schema instantiation, metadata editing, validation, and field distribution across domain-

specific tracks, supporting advanced browsing, querying, and versioning. The Dataset Module plays a central role by providing the structured facet in which provenance, historical context, reuse conditions, custodianship, and administrative descriptors are recorded. By separating dataset-level documentation from item-level metadata, it ensures coherence and traceability within the overall FAIRification process. The Digital Asset Management (DAM) component manages digital objects, extracting technical metadata, generating thumbnails, and enabling visualisation through dedicated viewers such as EVT (Edition Visualization Technology)[12] a modular, front-end component for the visualization and interaction with TEI-encoded scholarly editions, consuming MFE-exposed services and APIs to enable interoperable access and reuse across heterogeneous digital library environments [12]. The Portal Entity Builder (PEB) contributes the conceptual modelling environment, allowing curators and domain experts to define ontologies, classes, and semantic relationships that underpin mapping operations.

The Semantic Mapping and RDF Conversion Engine interprets these mappings as transformation rules, converting XML structures into RDF triples using a deterministic and reusable approach. These triples are indexed and stored within the Virtuoso triplestore, which exposes SPARQL endpoints supporting graph-based querying, inference, and Linked Open Data publication. Overarching all these components, the editorial workflow layer regulates the validation and publication cycle through controlled states—draft, review, approval, and publication—ensuring data quality and role coordination.

The MetaFAIR Ecosystem is designed as a connected node within a broader research infrastructure landscape. It integrates with other services within the DARIAH-IT infrastructures:

- Digital Heritage and Memory Hub (DHMH), which receives synchronized and indexed metadata via Solr for public access and cross-domain search, providing a unified discovery interface for heterogeneous cultural heritage resources.
- the Digital Philology Hub (DPH), which reuses MFE's curated and semantically enriched manuscripts descriptions within its philological research workflows. Through dedicated REST endpoints, DPH can query MFE's manuscripts repository, retrieve matching records, and incorporate selected items directly into project workspaces. This interoperability enables researchers to identify a manuscript in MFE, import it into DPH (assigning a project-specific code *i.e.*: a sigla), and continue their work with authoritative metadata seamlessly integrated into their editorial environment.

Beyond these integrations, MFE also connects with a wider ecosystem of external services and infrastructures, including:

- external data producers and consumers via REST APIs supporting record creation, update, and authority retrieval, allowing MFE to interact programmatically with platforms that contribute or reuse cultural heritage data;
- authentication and authorization services through federated identity management (Keycloak), enabling Single Sign-On across the infrastructure and ensuring that users maintain consistent roles and permissions throughout interconnected services;
- semantic web infrastructures, through the publication of RDF data via Virtuoso endpoints and support for multiple serialization formats, which extend the reach of MFE data into Linked Open Data ecosystems and enable graph-based integration, querying, and reasoning.

These integrations position MFE as both a producer and consumer of FAIR data within distributed digital ecosystems. Its connections to portals for public dissemination (DHMH), research environments for data reuse (DPH), and semantic infrastructures for linked data publication extend its role far beyond that of a traditional cataloguing environment. MFE operates instead as a foundational interoperability layer, enabling research to be built upon high-quality, provenance-aware, reusable data.

[12] EVT is an open-source tool developed at the University of Pisa under the supervision of Prof. Roberto Rosselli del Turco specifically designed to create digital editions from XML-encoded texts. See: http://evt.labcd.unipi.it/

5. Conclusions and future works

In this paper, we discussed key data management challenges in the SSH domain and illustrated how these issues could be addressed through a dedicated infrastructural component specifically designed for this purpose (MFE). Data curation and management represent a critical concern both for individual researchers and for institutions, whose interpretative frameworks and scholarly narratives are increasingly built upon digital data. Ensuring the availability of high-quality, well-curated datasets is a prerequisite for sound research practices, while the adoption of documented and shared frameworks is essential to support data reuse and interoperability across disciplinary ecosystems.

In the case of MFE a key area of application lies in the reuse of curated datasets across research environments. The integration with services such as the Digital Philology Hub demonstrates how FAIR-aligned metadata can be directly embedded into scholarly workflows, enabling researchers to discover, contextualise, and reuse authoritative descriptions without duplicating curation efforts. Similar integration patterns can be extended to other domain-specific tools, supporting collaborative annotation, critical editions, and computational analysis.

MFE was presented as an accessible service through the H2IOSC Transnational and National Access (TNA) Call,[13] allowing the collection of user feedback on its functionality and user experience.[14] The service was also presented at the event "H2IOSC meets public and private sectors" (Naples, 8–9 July 2025). In January 2026, further testing was enabled through the "H2IOSC Early Adopters" call, supporting hands-on access to MFE and other H2IOSC services until the end of the project.

Future developments will focus on expanding the range of supported schemas and ontologies, improving automated validation and mapping strategies, and strengthening cross-domain interoperability. Further work will also address scalability and performance in large, multi-institutional scenarios, as well as tighter alignment with EOSC services and emerging interoperability frameworks. In this perspective, the proposed solution provides a flexible foundation for experimentation and growth, supporting the transition from collection-centric digital libraries to data-centric, machine-actionable research ecosystems.

Acknowledgments

The H2IOSC – Humanities and Cultural Heritage Italian Open Science Cloud – project is funded, as part of the European Union Next Generation programme, by Italy's National Recovery and Resilience Plan (NRRP) under Mission 4 "Education and Research", Component 2 "From research to business", Investment 3.1 "Fund for the realisation of an integrated system of research and innovation infrastructures", Action 3.1.1 "Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe". (Project code IR0000029 – CUP B63C22000730005. Implementing Entity: CNR.)

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT 5.2 in order to: grammar and spelling check, and improving writing style. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

[13] For more information about the TNA/NA Calls, please visit: https://www.h2iosc.cnr.it/tna-na-calls/ (retrieved February 5, 2026).

[14] See MFE presentation for the TNA event: https://github.com/fpinnahub/DataForPapers/blob/main/MFE_DARIAH-IT.pdf (retrieved February 5, 2026).

References

[1] V. Bartalesi, N. Pratelli, E. Lenzi, Linking different scientific digital libraries in digital humanities: The imago case study, International Journal on Digital Libraries 23 (2022) 1–15. doi:10.1007/s00799-022-00331-4.

[2] S. Peroni, F. Tomasi, F. Vitali, The aggregation of heterogeneous metadata in web-based cultural heritage collections, International Journal of Web Engineering and Technology 8 (2013) 412–432. doi:10.1504/IJWET.2013.059107.

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[4] G. Candela, N. Gabriëls, S. Chambers, M. Dobreva, S. Ames, M. Ferriter, N. Fitzgerald, V. Harbo, K. Hofmann, O. Holownia, A. Irollo, M. Mahey, E. Manchester, T.-A. Pham, A. Potter, E. Van Keer, A checklist to publish collections as data in GLAM institutions, Global Knowledge, Memory and Communication 74 (2023) 1323–1355. doi:10.1108/GKMC-06-2023-0195.

[5] DARIAH-EU, DARIAH-EU annual report 2023, 2024. URL: 10.5281/zenodo.14007767.

[6] C. Wijesundara, S. Sugimoto, Metadata model for organizing digital archives of tangible and intangible cultural heritage, and linking cultural heritage information in digital space, LIBRES 28 (2018) 58–80. doi:10.32655/LIBRES.2018.2.2.

[7] C. Kakali, I. Lourdi, T. Stasinopoulou, L. Bountouri, C. Papatheodorou, M. Doerr, M. Gergatsoulis, Integrating dublin core metadata for cultural heritage collections using ontologies, in: Proceedings of the International Conference on Dublin Core and Metadata Applications: Application Profiles: Theory and Practice, 2007, pp. 128–139. doi:https://dl.acm.org/doi/10.5555/1344591.1344612.

[8] L. D. Miranda, K. Kutt, E. Sroka, G. J. Nalepa, Rich interoperable metadata for cultural heritage projects at jagiellonian university, 2024.

[9] European Open Science Cloud (EOSC) Secretariat, EOSC interoperability framework (v1.0), 2020. URL: https://eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf.

[10] European Commission, Directorate-General for Research and Innovation, EOSC Executive Board, EOSC interoperability framework: Report from the EOSC executive board working groups FAIR and architecture, 2021. doi:10.2777/620649.

[11] T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. M. Wallach, H. Daumé III, K. Crawford, Datasheets for datasets, Communications of the ACM 64 (2018) 86–92. doi:10.1145/3458723.

[12] R. Rosselli Del Turco, G. Buomprisco, C. Di Pietro, J. Kenny, R. Masotti, J. Pugliese, Edition visualization technology: A simple tool to visualize TEI-based digital editions, Journal of the Text Encoding Initiative (2015). doi:10.4000/jtei.1077.