# HERITRACE: A User-Friendly Semantic Data Editor with Change Tracking and Provenance Management for Cultural Heritage Institutions

*Arcangelo Massari[1] [1] and Silvio Peroni[2] [1]*

*[1] Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy*

## Abstract (English)

HERITRACE is a data editor designed for galleries, libraries, archives and museums, aimed at simplifying data curation while enabling non-technical domain experts to manage data intuitively without losing its semantic integrity. While the semantic nature of RDF can pose a barrier to data curation due to its complexity, HERITRACE conceals this intricacy while preserving the advantages of semantic representation. The system natively supports provenance management and change tracking, ensuring transparency and accountability throughout the curation process. Although HERITRACE functions effectively out of the box, it offers a straightforward customization interface for technical staff, enabling adaptation to the specific data model required by a given collection. Current applications include the ParaText project, and its adoption is already planned for OpenCitations. Future developments will focus on integrating the RDF Mapping Language (RML) to enhance compatibility with non-RDF data formats, further expanding its applicability in digital heritage management.

## Abstract (Italian)

HERITRACE è un editor di dati semantici progettato per gallerie, biblioteche, archivi e musei, con l'obiettivo di semplificare la curatela dei dati consentendo agli esperti non tecnici di gestirli in modo intuitivo senza compromettere la loro integrità semantica. Sebbene le tecnologie del Web Semantico possano rappresentare una barriera alla curatela dei dati a causa della loro complessità, HERITRACE offusca questa complessità preservando al contempo i vantaggi della rappresentazione semantica. Il sistema supporta nativamente la gestione della provenance e il tracciamento delle modifiche, garantendo trasparenza e responsabilità durante il processo di curatela. Pur funzionando efficacemente fin dal primo utilizzo, HERITRACE offre un'interfaccia di personalizzazione semplice per il personale tecnico, consentendo l'adattamento al modello di dati specifico richiesto da una determinata collezione. Tra le applicazioni attuali figura il progetto ParaText, e l'adozione è già prevista per OpenCitations. Gli sviluppi futuri si concentreranno sull'integrazione del RDF Mapping

---

[1] arcangelo.massari@unibo.it

[2] silvio.peroni@unibo.it

Language (RML) per migliorare la compatibilità con i formati di dati non RDF, ampliando ulteriormente la sua applicabilità nella gestione del patrimonio digitale.

# 1   Introduction

Traditionally, GLAM (galleries, libraries, archives, and museums) experts have relied on their interpretative skills and domain knowledge to curate metadata. However, the digitization of cultural heritage data has introduced new challenges, including the representation of data in various machine-readable and semantic-actionable formats and their preservation in heterogeneous databases and repositories. This scenario has created significant barriers for domain experts lacking technical expertise, particularly in Semantic Web technologies.

The widespread adoption of Semantic Web technologies in the GLAM sector underscores their growing relevance. Numerous prominent institutions, including the Library of Congress (Laurence 2013), Bibliothèque nationale de France (da Silveira et al. 2020), and Deutsche National Bibliothek (Hannemann and Kett 2010)2.1**Error! Reference source not found.** have embraced the Linked Open Data (LOD) paradigm to enhance the findability, accessibility, interoperability and reusability of their collections (Wilkinson et al. 2016). RDF (Pan 2009) has been instrumental in these efforts, enabling seamless integration of diverse data formats and fostering computational analysis. Candela highlights how RDF facilitates not only data structuring but also its reuse and enrichment, making it an essential middleware technology for GLAM collections (Candela 2023). Additional studies emphasize the pivotal role of RDF and LOD in creating interoperable datasets and fostering global collaboration across institutions (Alexiev 2018; Candela et al. 2022; Ranjgar et al. 2024).

While Semantic Web technologies have unlocked new possibilities, they have also led to a paradoxical situation. On the one hand, these technologies have made human intervention more critical due to the semantic interpretation of data that cannot be automated. On the other hand, they have limited the number of curators to those who are experts in the Semantic Web, thereby creating challenges in workforce scalability and accessibility. This duality has resulted in contrasting scenarios within the GLAM domain. Some collections have embraced Semantic Web technologies, requiring staff with advanced technical expertise for maintenance, while others have avoided these technologies to prevent curatorial complexities.

Two examples illustrate these divergent paths: the FICLIT Digital Library (ADLab - Laboratorio Analogico Digitale and /DH.arc - Digital Humanities Advanced Research Centre 2022) and OpenCitations (Silvio Peroni and Shotton 2020), both managed by the University of Bologna. The FICLIT Digital Library, relying on Omeka S, faces limitations due to its simplistic semantic tools and lack of SPARQL query capabilities, leading to challenges in change tracking and transparent provenance management. In contrast, OpenCitations fully embraces Semantic Web technologies but grapples with the issue of incorrect or missing data, a problem that requires human discernment for correction.

The central problem that arises from these scenarios is the gap between the complex digital technologies and the domain expertise of GLAM professionals. This gap hinders effective data curation and limits the potential of digital collections to represent and disseminate cultural heritage accurately and comprehensively. The solution proposed in this project is the development of a framework that facilitates domain experts without skills in Semantic Web technologies in enriching and editing such semantic data intuitively, irrespective of the underlying ontology model and the technologies adopted for storing such data.

The challenges are manifold. A critical goal is to create a system that is user-friendly for several kinds of end-users, including librarians, museologists, gallery curators, archivists, administrators and IT professionals who are tasked with setting up and maintaining the framework. Another significant challenge is provenance management. In the context of GLAM institutions, where the historical and source context of data is paramount, a data management system must accurately track and document the responsible agents and primary data sources. Change-tracking is also a fundamental requirement. The system needs to efficiently monitor and record all modifications to the data, allowing for transparency and accountability in the curation process. Customization is a further challenge that a data management system for cultural heritage must address. Recognizing that different GLAM domains have unique requirements for how resources are represented and managed, a customizable interface should be tailored to various data models, enabling the representation of diverse resource types according to specific domain needs. Finally, interfacing with pre-existing data presents a substantial challenge, as GLAM institutions often already possess vast collections, which organize their data through the adoption of different data models. This requirement is particularly important for ensuring that the transition to a new data management system is smooth and does not disrupt the ongoing operations of the institution.

To address all the aforementioned problems, we have developed a novel Web-based tool called HERITRACE (Heritage Enhanced Repository Interface for Tracing, Research, Archival Curation, and Engagement). HERITRACE is a novel RDF (Resource Description Framework) data editor which addresses the increasing complexities faced by cultural heritage institutions – including GLAMs – when curating metadata. This article extends extensively a previous publication – included in the Book of Abstracts of AIUCD 2024 (Di Silvestro and Spampinato 2024) – adding additional information about the technologies used for implementing HERITRACE, the comparison with existing tools, and the features and functions it implements.

The rest of this paper is structured as follows: Section 0 delves into related work, distinguishing between the adoption of Semantic Web technologies (Section 2.1) and the comparative analysis of semantic editors (Section 2.2), providing an in-depth comparative analysis of HERITRACE and existing systems in the GLAM domain. Section 0 describes the architecture and key functionalities of HERITRACE, with a focus on usability, provenance, and change tracking. Finally, Section 17 discusses future directions and broader implications of HERITRACE, concluding with its potential contributions to digital heritage management.

## 2   Related Work

The integration of Semantic Web technologies into GLAM institutions has become a pivotal development in the management and dissemination of cultural heritage. GLAM institutions are increasingly adopting RDF and Linked Open Data (LOD) principles to structure their metadata and enrich their collections. This section reviews two main aspects: the widespread adoption of Semantic Web technologies by GLAM institutions and the development of semantic editors designed to enable domain experts to curate RDF data intuitively.

### 2.1   Adoption of Semantic Web Technologies in GLAM Institutions

The past decade has witnessed a significant transformation in how cultural heritage institutions manage and share their collections, with many organizations embracing Semantic Web technologies and LOD principles. A survey of major national libraries reveals the extent of this transformation. Table 1 presents a comprehensive overview of national and international libraries that have embarked on this digital journey, showcasing the breadth of this adoption across Europe and beyond.

The Österreichische Nationalbibliothek (Rachinger 2008), with roots dating back to the 14th century Habsburg Empire, has converted its authority files and bibliographic records to LOD, making over 1.8 million objects accessible via the Web. These include medieval manuscripts, early printed books, and the world's largest collection of papyri. The Biblioteca Nacional de España's platform (Wulff 2024) demonstrates similar ambitions, offering RDF representations of its catalog that connect authors, works, and subjects, providing access to Spanish literary heritage, from medieval manuscripts to Golden Age literature. The Biblioteca Virtual Miguel de Cervantes (Candela et al. 2018) has focused on semantically enriching its digital collections of Hispanic literature, creating connections between literary works, their authors, and their historical context.

The Bibliothèque nationale de France (da Silveira et al. 2020) exemplifies how LOD can enhance access to French cultural heritage, aggregating information about authors, works, and subjects previously scattered across various databases. Their collection spans from illuminated medieval manuscripts to contemporary publications, representing over five centuries of French intellectual and artistic achievement. Similar efforts are visible at the Bibliothèque nationale du Luxembourg (Kremer 2021), which has prioritized the semantic representation of its multilingual collections, reflecting Luxembourg's unique position at the crossroads of European cultures. Also, the British National Bibliography (Corine et al. 2017) has focused on publishing bibliographic metadata as LOD, allowing for complex queries that were impossible with traditional catalog systems.

The Deutsche National Bibliothek's semantic services (Hannemann and Kett 2010) provide access to German-language cultural resources, including extensive collections from the Reformation period to the present day, while the Library of Congress's LOD services not only cover American cultural heritage but also provide reference points for cultural institutions worldwide through their authority data (Laurence 2013). The National Library of Finland (Hyvönen 2020) has integrated its data with the broader Finnish LOD infrastructure, making accessible collections related to Nordic culture and history, while the National Library of the Netherlands (Beek et al. 2014) has focused on connecting its historical collections through semantic relationships, providing access to holdings from the Dutch Golden Age.

However, it is the Europeana project (Isaac and Haslhofer 2013) that represents the most ambitious implementation in the field, aggregating and semantically enriching cultural heritage data from institutions across Europe. Its data model has become a reference for representing cultural heritage information using Semantic Web technologies, enabling users to explore connections between diverse cultural artifacts, from medieval manuscripts to modern photographs.

These implementations, while successful in making cultural heritage data more accessible and interoperable, highlight a critical need: tools that allow domain experts - librarians, archivists, and curators - to manage and enrich semantic data without requiring extensive technical knowledge. The complexity of current LOD management systems often creates a barrier between cultural heritage professionals and their own data, limiting their ability to maintain and enhance these valuable resources. This situation underscores the importance of developing user-friendly semantic editors that can work with various data models while remaining accessible to non-technical users.

| Institution | URL | Reference |
|---|---|---|
| Österreichische Nationalbibliothek | https://labs.onb.ac.at | (Rachinger 2008; Petz 2023) |
| Biblioteca Nacional de España | http://datos.bne.es | (Wulff 2024) |
| Biblioteca Virtual Miguel de Cervantes | https://data.cervantesvirtual.com | (Candela et al. 2018) |
| Bibliothèque nationale de France | https://data.bnf.fr | (da Silveira et al. 2020) |
| Bibliothèque nationale du Luxembourg | https://data.bnl.lu | (Kremer 2021) |
| British National Bibliography | https://bl.natbib-lod.org/ | (Corine et al. 2017) |
| Europeana | https://pro.europeana.eu | (Isaac and Haslhofer 2013) |
| Deutsche National Bibliothek | https://www.dnb.de | (Hannemann and Kett 2010) |
| Library of Congress | https://id.loc.gov | (Laurence 2013) |
| National Library of Finland | https://data.nationallibrary.fi | (Hyvönen 2020) |
| National Library of the Netherlands | https://data.bibliotheken.nl | (Beek et al. 2014) |

*Table 1 Overview of LOD repositories published by GLAM organizations, adapted from (Candela 2023).*

## 2.2   Comparative Analysis of Semantic Editors in GLAM

Building on the discussion of Semantic Web adoption, this subsection shifts the focus on semantic editors. A comparative evaluation of systems, including OmekaS (Salarelli 2016), Semantic MediaWiki (Krötzsch, Vrandečić, and Völkel 2006), Research Space (Oldman and Tanase 2018), and CLEF (Marilena Daquino et al. 2022), reveals varying strengths and limitations across criteria like user-friendliness, provenance management, change tracking, customization, and compatibility with heterogeneous data sources. These evaluation criteria are based on those used to assess the CLEF system in the related paper. This ensures that our assessment criteria are not only relevant but also consistently applied across similar platforms within the digital heritage domain, as summarized in Table 2.

| Name | User friendly (Users) | User friendly (Admin) | Provenance Mgmt. | Change-tracking | Customization | Heterogeneous data sources |
|---|---|---|---|---|---|---|
| OmekaS | ✓ | ✓ | | | ✓ | |
| Semantic MediaWiki | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Research Space | ✓ | | ✓ | | ✓ | ✓ |
| CLEF | ✓ | ✓ | ✓ | ✓ | | |

*Table 2 Comparison of Data Management System Features for the GLAM Sector.*

OmekaS, recognized for its user-friendly interface, primarily serves museums and educational institutions with its intuitive web-publishing platform. However, it exhibits certain limitations in more complex operational aspects. Notably, OmekaS does not inherently track provenance. This limitation can affect the credibility and traceability of the information presented. Additionally, OmekaS lacks inbuilt change-tracking capabilities. Data interfacing in OmekaS presents another challenge. To import pre-existing data in bulk, users must rely on the CSV Import plugin (Berthereau and Corporation for Digital Scholarship 2015). This plugin necessitates restructuring the original data to fit its specific format with mandatory field names, which can be a cumbersome and time-consuming process. This requirement for data formatting reduces the platform's flexibility in handling heterogeneous data sources.

Semantic MediaWiki significantly enhances the popular MediaWiki platform by integrating semantic capabilities. This blend of features balances user-friendliness for non-technical end-users and the more complex needs of technical administrators. One of the key strengths of Semantic MediaWiki is its customization potential, although it requires a degree of familiarity with both the MediaWiki environment and underlying semantic concepts. In terms of data provenance management, Semantic MediaWiki provides robust support. However, its capabilities for change-tracking are not native to the system but are instead supplemented through the use of external plugins. A notable example is the Semantic Watchlist plugin (WikiTeq 2022), which effectively monitors changes within the wiki. These changes are stored in a relational database rather than in RDF format, which, while practical for tracking purposes, may not align seamlessly with the semantic structure of the data. This discrepancy could potentially restrict the depth of change analysis and the ability to contextualize changes within the semantic framework of the data. Addressing the interfacing with heterogeneous data sources, Semantic MediaWiki initially focused solely on importing OWL ontologies. To broaden its RDF support, the RDFIO extension was introduced (Lampa et al. 2017). This extension enables the loading of RDF triples, but it is confined to the N-Triples format and notably lacks support for named graphs. This limitation is significant as it restricts the platform's adaptability in various environments that may require more complex semantic data structures.

Research Space, tailored for the academic and research community, excels in user-friendliness for end-users, offering diverse data visualization options such as graphs and temporal maps. However, it maintains a level of complexity for administrators, demanding a steep learning curve. The platform requires a solid understanding of HTML, handlebars and other ResearchSpace-specific components for creating templates, which may be cumbersome even for those with technical expertise. In terms of data provenance, Research Space automatically associates data with its source, ensuring traceability and credibility. However, it lacks a change-tracking system, which could limit its effectiveness in environments where monitoring data modifications over time is crucial. Regarding data interfacing, Research Space allows uploading RDF data directly, which is advantageous for projects involving such formats. However, after the data is uploaded, an administrator's intervention is required to customize the interface appropriately to display the items correctly. This aspect indicates that while Research Space can interface with heterogeneous data sources, doing so involves a significant level of programming complexity for system administrators.

CLEF is designed to manage complex digital libraries, archives, and research data, particularly in the humanities. It offers an administrator-friendly interface and focuses on user-friendliness for end-users, making it suitable for a wide range of audiences within its domain. CLEF's provenance management is robust, utilizing named graphs. Moreover, it does feature change tracking capabilities, including synchronization with GitHub, but lacks a direct system to restore previous versions. Expanding on the capabilities of CLEF, it is important to note that this system does not allow for extensive customization. Moreover, unlike some of its counterparts, CLEF is not designed to upload and manage pre-existing RDF data as-is. This limitation is significant because the software is structured to add items one by one from scratch directly through the user interface. This approach, while potentially beneficial for building new databases, limits the platform's ability to seamlessly integrate and manage existing large-scale datasets. Furthermore, even though CLEF does not impose a specific data model, it organizes data in a format akin to nanopublications for managing provenance. This structure means that if a pre-existing triplestore is connected, the system is not readily equipped to explore the data without a prior reorganization to make it compatible with CLEF's framework.

The comparative analysis highlights how these systems cater to different needs within the GLAM sector. While tools like OmekaS focus on simplicity and accessibility for non-technical users, they lack advanced features like provenance tracking and change management. Semantic MediaWiki and Research Space offer richer functionalities but come with a steeper learning curve, especially for administrators. CLEF stands out for its robust provenance management and synchronization with external platforms, yet its limited ability to handle pre-existing RDF data restricts its flexibility. These insights underline the need for a balanced approach that combines usability with advanced semantic capabilities, paving the way for tools that can bridge these gaps effectively.

# 3    HERITRACE Functionalities

The user interface (UI) of HERITRACE is designed to cater to the needs of domain experts in the GLAM sector, ensuring that they can interact with complex datasets without requiring technical expertise. The UI is web-based, allowing access from various devices and platforms, including both PCs and smartphones, and supports professionals by enabling them to work across different environments. Access to the system is restricted to pre-authorized personnel, with login credentials managed via ORCID (Haak et al. 2012), ensuring secure and streamlined authentication.

This flexibility and security are demonstrated in HERITRACE's deployment within the ParaText project, which serves as a real-world case study for the system's capabilities. Developed by the University of Bologna, ParaText focuses on managing bibliographic metadata for textual resources. While access to ParaText is restricted to authenticated and authorized personnel, all examples and screenshots illustrating HERITRACE's functionalities in this paper are drawn from its deployment in the ParaText project.

While HERITRACE is not primarily engineered as a comprehensive search tool— unlike specialized systems such as OSCAR (Heibi, Peroni, and Shotton 2019) — it provides a basic interface for visualizing the contents of a dataset. The catalog interface in HERITRACE presents a two-panel layout. On the one hand, users can find a "Categories" panel that lists all available types of objects described in the system, with each category displaying the total number of items it contains. For example, Figure 1 shows various publication types such as "Article in Book" (153 items), "Issue" (25 items), "Journal" (27 items), and so forth. On the other hand, users can view the items belonging to the selected category. The interface includes practical features such as a "Sort by" dropdown menu allowing users to order items by different properties, an "Items per page" selector for controlling the number of displayed results, and pagination controls for navigating through the list of items. Finally, each item in the list is displayed as a clickable link.

**Catalogue**

**Categories**

| Category | Count |
|---|---|
| Article in Book | 153 |
| Issue | 25 |
| Journal | 27 |
| Journal Article | 77 |
| Monograph | 66 |
| Proceedings Paper | 10 |
| Volume | 62 |

**Items in category: Article in Book**

Sort by: Title

Items per page: 50        1 / 4

Pagani, L. (2015). Al crocevia di lingua e letteratura. Il grammatico Filosseno come esegeta di Omero. In: Lemmata. Beiträge zum Gedenken an Christos Theodoridis.

Muzzolon, M. (2005). Aristarco negli scolii ad Aristofane. In: Interpretazioni antiche di Aristofane.

Perrone, S. (2005). Aristofane e la religione negli scholia vetera alle Rane. In: Interpretazioni antiche di Aristofane.

Montanari, F. (2011). Correcting a copy, editing a text. Alexandrian ekdosis and papyri. In: From scholars to scholia. Chapters in the history of ancient Greek scholarship.

Pagani, L. (2011). Pioneers of grammar. Hellenistic scholarship and the study of language. In: From scholars to scholia. Chapters in the history of ancient Greek scholarship.

Ascheri, P. (2011). The Greek origins of the Romans and the Roman origins of Homer in the Homeric scholia and in POxy. 3710. In: From scholars to scholia. Chapters in the history of ancient Greek scholarship.

*Figure 1 HERITRACE catalog interface.*

When a user clicks on an item in the catalog, they are presented with a detailed view of the resource. This interface displays the resource's metadata fields, such as type, identifier, and title as shown in Figure 2, depending on the structure of the specific dataset. Each field is labeled and paired with its current value, allowing users to view or modify the information through inline editing. Intuitive controls, like text inputs and dropdown menus, facilitate straightforward updates, with real-time validation to ensure data consistency. Action buttons at the top of the interface enable users to perform key operations. These include canceling edits without saving changes, deleting the resource, or accessing the "Time Machine" to view and restore previous versions. Users can also expand the metadata by adding new fields, such as additional identifiers or descriptions.

*Figure 2 Detailed view of a resource in HERITRACE, showing metadata fields and editing options.*

The "Time Machine" feature in HERITRACE, illustrated in Figure 3, provides users with a timeline for managing the evolution of each entity in the database. Every time an entity is created, modified, or deleted the system generates a new snapshot. These snapshots capture key provenance metadata, including the timestamp of generation, the responsible agent, the primary data source, a description of the entity, and a detailed list of the modifications made.

The timeline interface allows users to view previous versions of a resource, facilitating comparisons and enabling the restoration of earlier states if the current version is no longer suitable. When a user decides to restore an entity to a previous snapshot, HERITRACE automatically updates the resource and ensures that all linked resources are also adjusted accordingly. For example, if a bibliographic resource is restored to a previous version where its associated identifiers or descriptive metadata differ from the current state, HERITRACE ensures that the linked records are reverted to previous snapshots as well. This integrated approach minimizes the risk of

inconsistencies and guarantees that the dataset remains coherent and accurate throughout its evolution. The "Time Machine" thus is a critical tool for maintaining transparency, accountability, and reliability in metadata curation and management.

Deleted resources are not permanently removed from the system but remain accessible through "Time Vault". This interface, which closely resembles the catalog, provides a dedicated view for exploring the history of deleted resources. Users can review essential information about deleted resources, such as the deletion timestamp, the responsible agent, and the associated modifications. Additionally, the Time Vault enables quick access to the most recent version of a deleted resource, allowing users to restore it if necessary. This approach ensures that no data is irretrievably lost while maintaining a complete and auditable history of all actions performed within the system.



Figure 3 The HERITRACE timeline interface.

In addition to managing existing resources, HERITRACE provides an interface for creating new records, ensuring that curators can enrich their datasets while maintaining the same standards of accuracy and consistency. This process begins with the "New Record" section, where users can easily define new entities and their associated metadata. The system dynamically adjusts the available metadata fields based on the selected type, ensuring compatibility with the dataset's underlying structure. For example, as shown in Figure 4, if "Article" is selected as the entity type, the interface presents fields such as "Identifier," "Title," "Description," "Abstract," "Keyword," "Author," "Editor," "Collaborator," and "Publisher." Users can expand and fill these fields as needed. Each field includes clear labels and helpful tooltips to guide users in entering accurate and relevant metadata.

To ensure the integrity and reliability of the data, HERITRACE incorporates a robust validation system that applies both to the creation of new records and to the modification of existing entities. As users populate or update metadata fields, the system performs real-time checks to validate the entered information against predefined rules and constraints. These constraints, defined through SHACL (Pareti and Konstantinidis 2022), may include conditions such as mandatory fields, value types (e.g., text, numbers, or URIs), or the allowable number of values for a specific property. Validation feedback is provided immediately, highlighting any issues and guiding users to correct errors or omissions before saving changes. This ensures that all new and modified records adhere to the dataset's structural and semantic requirements, minimizing errors and promoting consistency across the collection.

*Figure 4 HERITRACE interface for creating a new resource.*

## 3.1 Configurations and customization

HERITRACE is designed to function seamlessly with any RDF-based data set, allowing institutions to deploy the system without the need for extensive initial configuration. However, for those who wish to tailor the system to their specific needs, HERITRACE offers extensive configuration options to enhance its functionality and user experience. This subsection delves into a more technical aspect of HERITRACE, focusing on the role of the technical staff or configurator rather than the domain expert. The system is designed to simplify the configurator's task by allowing customization through well-known languages, particularly SHACL (Pareti and Konstantinidis 2022) and YAML (Ben-Kiki, Evans, and Ingerson 2009).

SHACL (Shapes Constraint Language) is a W3C standard used to validate RDF data against a set of conditions. It enables the definition of constraints on data models, ensuring that the data adheres to specific structural and semantic rules. In HERITRACE, SHACL is employed to define the data model, specifying classes, properties, and constraints for each entity type. This ensures data consistency and integrity across the dataset. In HERITRACE, SHACL determines not only which properties belong to a particular class but also how these properties should be managed and validated within the system.

For instance, if the SHACL shape related to the class `fabio:JournalArticle` states that the `prism:publicationDate` property can have a data type of `xsd:date`, `xsd:gYearMonth`, or `xsd:gYear`, the system will present a dropdown menu allowing users to select one of these date types. Depending on the selection, the appropriate input method will be displayed: a calendar for a full date, a month-year selector for `xsd:gYearMonth`, and a simple number input for `xsd:gYear`.

As said, SHACL also defines cardinality constraints using `sh:minCount` and `sh:maxCount`, which dictate how many times a property can be added. For example, a property like `dcterms:title` might have a max count of 1, indicating it can only appear once, whereas properties like `datacite:hasIdentifier` might have no upper limit, allowing multiple entries.

Beyond basic property definitions, SHACL supports sophisticated validation mechanisms. Regular expression patterns can be specified to validate specific formats, such as a DOI (Digital Object Identifier). By embedding a pattern within the SHACL definition, the system can ensure that the DOI adheres to its expected format. Additionally, SHACL allows for custom error messages to be specified, providing clear feedback if the input does not match the required pattern.

Complementing SHACL, HERITRACE uses YAML configuration files to control the visual presentation of data. YAML allows administrators to define user-friendly labels for properties, decide which fields to display in the interface, and specify how certain properties should be represented. For example, an administrator might configure a property to be displayed with values retrieved via a SPARQL query. For instance, resource names visible in the figures are backed by unique URIs, as these resources are RDF entities. However, ParaText represents these URIs as complete bibliographic references. This transformation is achieved dynamically through SPARQL queries, ensuring that domain experts can interact with intuitive bibliographic descriptions rather than abstract URIs. This level of customization ensures that the interface aligns with institutional priorities and enhances usability for end-users. The SHACL and YAML configuration files used for the ParaText project are available on Zenodo (Massari 2025).

## 3.2    Provenance management and change-tracking

In the context of HERITRACE, provenance refers to the metadata associated with modifications to a dataset, documenting essential metadata such as when the modification occurred, who was responsible for it, what specific changes were made, and the primary source of the data involved (Gil et al. 2010). Provenance management ensures a transparent and auditable history of curatorial actions, which is crucial for maintaining accountability and reliability in metadata workflows. HERITRACE employs the OpenCitations Data Model (OCDM) (M. Daquino et al. 2020) to implement

its provenance management system, extending the PROV Ontology (PROV-O) (Lebo, Sahoo, and McGuinness 2013) with additional mechanisms to record changes in the data.

Provenance in the OCDM revolves around the concept of snapshots, each representing a particular state of an entity, such as a book or an artifact, at a specific moment in time. Snapshots are stored as instances of `prov:Entity` and are linked to prior states of the entity through the `prov:wasDerivedFrom` property. This linkage allows the reconstruction of an entity's history, providing a comprehensive view of how it evolved. HERITRACE records a set of core metadata for each snapshot to ensure complete traceability. This metadata includes the timestamp of creation (`prov:generatedAtTime`), which identifies the precise moment a snapshot was created, and, when relevant, the timestamp of invalidation (`prov:invalidatedAtTime`), which marks when a snapshot was superseded by subsequent modifications. Additionally, the metadata includes the responsible agent (`prov:wasAttributedTo`), detailing who performed the modification—whether an individual, an organization, or an automated process—and the primary data source (`prov:hasPrimarySource`), which links the modification to the external or internal source that informed the change. Together, these elements form a robust provenance framework, ensuring that all curatorial actions are fully documented and that the dataset retains a high degree of integrity and accountability.

While provenance management focuses on contextual metadata, change-tracking addresses the content of modifications. The OCDM employs a document-based approach to change tracking applied to data, where only the differences, or deltas, between successive snapshots of an entity are stored. This method avoids the inefficiency of retaining complete backups of each version of the entity. Instead, HERITRACE records SPARQL update queries to track entity changes. Specifically, additions to an entity are captured as `INSERT DATA` queries, while removals are stored as `DELETE DATA` queries. These deltas are linked to the entity through the `oco:hasUpdateQuery` property, introduced by the OCDM (S. Peroni, Shotton, and Vitali 2016). This approach reduces the storage requirements significantly and simplifies the process of reconstructing previous versions of the entities. To return to the previous state, HERITRACE reverses the recorded SPARQL update queries. For example, if a triple was added to an entity, the system removes it to revert to the earlier state, and if a triple was removed, it reinstates it. This process is facilitated by a dedicated tool called the Time Agnostic Library (Massari 2024b), which automates the application and inversion of update queries, ensuring an efficient and precise restoration process.

The integration of provenance management and change tracking in HERITRACE is exemplified in Figure 5, which illustrates the lifecycle of an entity through successive modifications. Initially, the creation of an entity is captured as a snapshot, with associated provenance metadata detailing the timestamp, the responsible agent, and the primary source. As modifications occur, HERITRACE generates deltas that document the specific triples added or removed. For instance, if a new author is added to a bibliographic record, the corresponding delta is stored as an `INSERT DATA` query, while the removal of outdated information is captured as a `DELETE DATA` query. This process ensures that both the context and content of changes are meticulously documented and easily accessible.
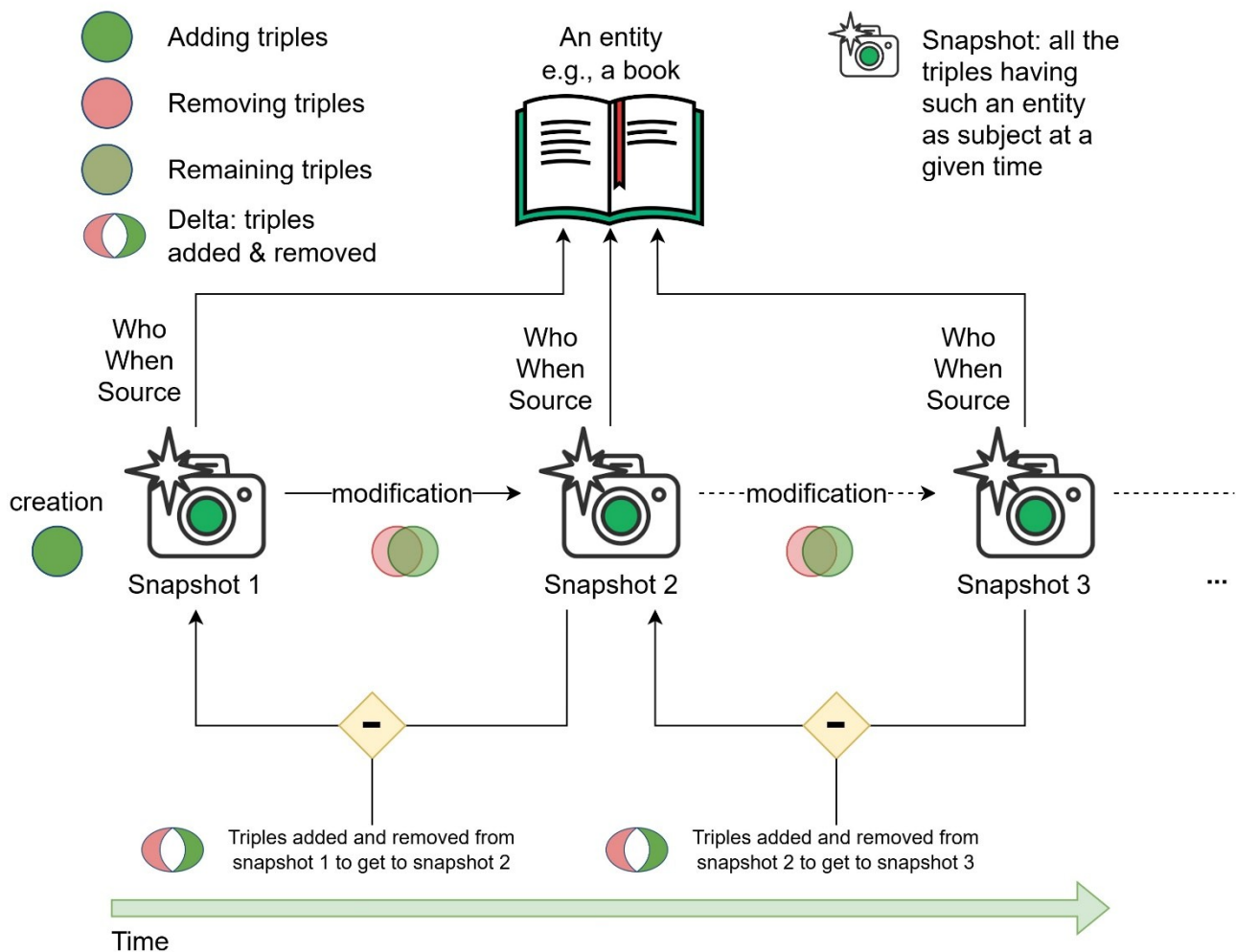
*Figure 5 OCDM change-tracking mechanism, showing the evolution of an entity through successive snapshots.*

For those interested in exploring HERITRACE further, the system along with its documentation are available on GitHub and Software Heritage (Massari 2024a), providing essential resources for implementation and use.

# 4   Discussion and future directions

As highlighted in Section 2.2, existing semantic data management systems such as OmekaS, Semantic MediaWiki, Research Space, and CLEF excel in specific areas but also exhibit limitations. OmekaS is renowned for its user-friendly interface, which lowers the entry barrier for non-technical users. However, its lack of built-in provenance and change-tracking mechanisms reduces its effectiveness in workflows requiring accountability and traceability. Furthermore, users are required to organize their datasets in CSV files with predefined columns to import them. This necessity to adapt and restructure the original data format before loading it into the system can be

cumbersome and time-consuming, limiting flexibility when interfacing with heterogeneous data sources.

Semantic MediaWiki, by contrast, provides robust provenance management through extensions like Semantic Watchlist and offers significant customization options. However, its reliance on relational databases instead of RDF as the primary storage format for change-tacking data creates a disjunction between its semantic capabilities and its data infrastructure. Furthermore, Semantic MediaWiki can only import triples via a plugin that does not support quadruples or named graphs, limiting its ability to capture contextual information that is often critical in Linked Data environments.

Research Space demonstrates strong provenance capabilities, automatically associating data with its source, and is particularly valued for its data visualization tools. However, its lack of integrated change-tracking mechanisms limits its ability to monitor and restore modifications over time. Moreover , its configuration requires significant technical expertise, particularly for tailoring the interface to project-specific needs. CLEF offers robust provenance management through named graphs and a change tracking system via an external versioning system, GitHub. However, its limited flexibility in uploading and managing pre-existing RDF datasets constrains its applicability in projects involving large-scale data.

On the other hand, HERITRACE offers a unified solution that integrates usability, provenance, and change-tracking into a single framework. By employing the OpenCitations Data Model (OCDM), HERITRACE ensures detailed provenance, recording when and by whom modifications were made, and linking changes to their primary sources. Its delta-based change-tracking approach, which utilizes SPARQL update queries, provides efficient storage and facilitates the reconstruction of previous dataset states without the need for full backups. Unlike Semantic MediaWiki, HERITRACE fully supports the use of RDF and is not restricted to triples, allowing for richer contextualization and representation of data.

Moreover, the importance of customization in the GLAM domain cannot be overstated. Institutions manage a wide array of objects—from books and artworks to digital-born resources—each requiring distinct metadata models. While Research Space offers a high degree of adaptability, its reliance on a proprietary templating language introduces significant challenges. Configurators must learn a custom syntax developed specifically for the platform, which is neither widely known nor intuitive, making customization a daunting task, even for technical staff. In contrast, HERITRACE employs SHACL, a W3C standard widely recognized for its role in defining and validating RDF data structures. SHACL enables administrators to define data models and entity constraints using a familiar, standardized syntax. Additionally, HERITRACE simplifies graphical customization by using a YAML configuration file. This file allows configurators to adjust interface elements by modifying predefined parameters. This dual approach ensures that HERITRACE can be tailored to meet the diverse needs of GLAM institutions, accommodating their heterogeneous data models while remaining accessible and efficient to configure. Furthermore, HERITRACE functions effectively out of the box, providing an immediate solution for GLAM institutions without the need for extensive initial configuration.

This streamlined usability is particularly critical in facilitating the adoption of Semantic Web technologies by cultural heritage institutions. As highlighted in Section 0 when mentioning the FICLIT Digital Library, many collections are wary of the curatorial challenges associated with these

technologies' complexity. This reluctance is especially pronounced in Italy, where, as shown in Table 1, none of the national libraries have yet implemented these technologies. In contrast to this stagnation, European institutions leveraging RDF benefit from adherence to FAIR principles and enhanced international collaboration. By offering a user-friendly tool, HERITRACE seeks to bridge this gap, enabling non-technical domain experts to reap the benefits of RDF-based systems without the associated challenges. Through this approach, HERITRACE aims to encourage the adoption of Semantic Web technologies within the Italian GLAM sector, fostering more effective and interoperable cultural heritage management.

Nevertheless, HERITRACE focus on RDF is a limitation. Cultural heritage data often exists in other formats, especially tabular data (e.g., CSV files) and relational databases. This diversity creates challenges for institutions aiming to integrate legacy systems or non-RDF datasets into a unified semantic framework. To address this gap, a key future direction for HERITRACE is the integration of the RDF Mapping Language (RML) (Dimou et al. 2014). RML is a powerful extension of the W3C-standard R2RML (Relational Database to RDF Mapping Language) and enables the transformation of heterogeneous data formats into RDF. By supporting mappings from tabular data, JSON, XML, and relational databases to RDF, RML would significantly enhance the system's flexibility. Its declarative approach allows users to define mapping rules in a straightforward and reusable manner, making it an ideal solution for bridging diverse data sources. For GLAM institutions, this integration would mean that even data not natively represented in RDF could be seamlessly incorporated into HERITRACE, further increasing the system's utility.

In addition to technical enhancements, usability testing will play a pivotal role in HERITRACE's future development. While the system has been designed with user accessibility in mind, real-world testing with GLAM professionals is essential to identify potential barriers and optimize the interface. Testing will focus on evaluating the intuitiveness of the user interface, the efficiency of key features like the Time Machine and customization tools, and the overall user experience for both non-technical and technical stakeholders. Insights from usability studies will guide iterative improvements, ensuring that HERITRACE continues to meet the evolving needs of its target audience.

Finally, a broader objective for HERITRACE is to expand its adoption and collaboration within the GLAM community. Current deployments, such as the ParaText project at the University of Bologna, demonstrate the system's potential to address real-world curatorial challenges. Moreover, HERITRACE is planned for adoption by OpenCitations (Silvio Peroni and Shotton 2020), an open infrastructure dedicated to the free dissemination of bibliographic and citation data. Both authors of this paper are involved in the OpenCitations project, which facilitates the alignment of the system with the organization's specific requirements. The adoption of HERITRACE will focus on managing metadata and citation datasets, tracking changes, and documenting data provenance. These applications are already providing an opportunity to test HERITRACE in real-world contexts characterized by large-scale, dynamic datasets.

# 5   Conclusion

In summary, HERITRACE presents itself as a practical solution in the field of semantic data management, with a particular focus on the needs of the GLAM sector. The system provides a user-friendly interface that caters to both nontechnical and technical users, alongside features such as provenance management, change tracking, and the ability to customize according to specific needs. Its capability to integrate with existing datasets enhances its practicality. Through these efforts, HERITRACE seeks to support efficient and transparent data curation processes, enabling institutions to manage their digital collections with accuracy and flexibility.

## Acknowledgements

## References

ADLab - Laboratorio Analogico Digitale and /DH.arc - Digital Humanities Advanced Research Centre. 2022. "FICLIT Digital Library." Bologna, Italy: University of Bologna.

Alexiev, Vladimir. 2018. "Museum Linked Open Data: Ontologies, Datasets, Projects." *Digital Presentation and Preservation of Cultural and Scientific Heritage* 8 (September):19–50. https://doi.org/10.55630/dipp.2018.8.1.

Beek, W. G. J., R. J. Hoekstra, F. Maas, A. Merono, and I. B. Leemans. 2014. "Linking the STCN and Performing Big Data Queries in the Humanities." In .

Ben-Kiki, Oren, Clark Evans, and Brian Ingerson. 2009. "Yaml Ain't Markup Language (Yaml™) Version 1.1." *Working Draft 2008* 5:11.

Berthereau, Daniel and Corporation for Digital Scholarship. 2015. "CSV Import." https://omeka.org/s/modules/CSVImport/.

Candela, Gustavo. 2023. "Towards a Semantic Approach in GLAM Labs: The Case of the Data Foundry at the National Library of Scotland." *Journal of Information Science*.

Candela, Gustavo, Pilar Escobar, Rafael C. Carrasco, and Manuel Marco-Such. 2018. "Migration of a Library Catalogue into RDA Linked Open Data." Edited by Christoph Schlieder. *Semantic Web* 9 (4): 481–91. https://doi.org/10.3233/SW-170274.

Candela, Gustavo, María Dolores Sáez, MPilar Escobar Esteban, and Manuel Marco-Such. 2022. "Reusing Digital Collections from GLAM Institutions." *Journal of Information Science* 48 (2): 251–67. https://doi.org/10.1177/0165551520950246.

Corine, Deliot, Wilson Neil, Costabello Luca, and Vandenbussche@ Pierre-Yves. 2017. "The British National Bibliography: Who Uses Our Linked  Data?" In *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative. https://doi.org/10.23106/DCMI.952137546.

Daquino, M., S. Peroni, D. Shotton, G. Colavizza, B. Ghavimi, A. Lauscher, M. Romanello, and P. Zumstein. 2020. "The OpenCitations Data Model." In *The Semantic Web – ISWC 2020. 19th International Semantic Web Conference. Proceedings, Part II*, 447–63. Athens, Greece: Springer. https://doi.org/10.6084/M9.FIGSHARE.3443876.V7.

Daquino, Marilena, Mari Wigham, Enrico Daga, Lucia Giagnolini, and Francesca Tomasi. 2022. "CLEF. A Linked Open Data Native System for Crowdsourcing." https://doi.org/10.48550/ARXIV.2206.08259.

Di Silvestro, Antonio, and Daria Spampinato. 2024. *Me.Te. Digitali. Mediterraneo in Rete Tra Testi e Contesti, Proceedings Del XIII Convegno Annuale AIUCD2024*. AIUCD. https://doi.org/10.6092/UNIBO/AMSACTA/7927.

Dimou, Anastasia, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 2014. "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data." In *Proceedings of the 7th Workshop on Linked Data on the Web*, edited by Christian Bizer, Tom Heath, Sören Auer, and Tim Berners-Lee. Vol. 1184. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

Gil, Y., J. Cheney, P. Groth, O. Hartig, S. Miles, L. Moreau, and P. Silva. 2010. "Provenance XG Final Report." *W3C*. http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/.

Haak, Laurel L., Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. 2012. "ORCID: A System to Uniquely Identify Researchers." *Learned Publishing* 25 (4): 259–64. https://doi.org/10.1087/20120404.

Hannemann, Jan, and Jürgen Kett. 2010. "Linked Data for Libraries." In *Proc of the World Library and Information Congress of the Int'l Federation of Library Associations and Institutions (IFLA)*.

Heibi, Ivan, Silvio Peroni, and David Shotton. 2019. "Enabling Text Search on SPARQL Endpoints through OSCAR." Edited by Alejandra Gonzalez-Beltran, Alejandra Gonzalez-Beltran, Francesco Osborne, Silvio Peroni, and Sahar Vahdati. *Data Science* 2 (1–2): 205–27. https://doi.org/10.3233/DS-190016.

Hyvönen, Eero. 2020. "Linked Open Data Infrastructure for Digital Humanities in Finland." *Digital Humanities in the Nordic and Baltic Countries Publications* 3 (1): 254–59. https://doi.org/10.5617/dhnbpub.11195.

Isaac, Antoine, and Bernhard Haslhofer. 2013. "Europeana Linked Open Data – Data.Europeana.Eu." *Semantic Web* 4 (3): 291–97. https://doi.org/10.3233/SW-120092.

Kremer, Christine. 2021. "Bibliothèque nationale du Luxembourg : multiple et ouverte à tous." *Arabesques*, no. 102 (July), 26–27. https://doi.org/10.35562/arabesques.2657.

Krötzsch, Markus, Denny Vrandečić, and Max Völkel. 2006. "Semantic MediaWiki." In *The Semantic Web - ISWC 2006*, edited by Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora M. Aroyo, 4273:935–42. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11926078_68.

Lampa, Samuel, Egon Willighagen, Pekka Kohonen, Ali King, Denny Vrandečić, Roland Grafström, and Ola Spjuth. 2017. "RDFIO: Extending Semantic MediaWiki for Interoperable Biomedical Data Management." *Journal of Biomedical Semantics* 8 (1): 35. https://doi.org/10.1186/s13326-017-0136-y.

Laurence, Corinne M. 2013. "Linked Data and the Library of Congress." *Library Philosophy and Practice*, 2–24.

Lebo, T., S. Sahoo, and D. McGuinness. 2013. "PROV-O: The PROV Ontology." W3C. PROV-O. April 30, 2013. http://www.w3.org/TR/2013/REC-prov-o-20130430/.

Massari, Arcangelo. 2024a. "HERITRACE (Heritage Enhanced Repository Interface for Tracing, Research, Archival Curation, and Engagement)." https://archive.softwareheritage.org/swh:1:snp:4d1d83b7043649a21900fcbf6465f0879672228e;origin=https://github.com/opencitations/heritrace.

———. 2024b. "Opencitations/Time-Agnostic-Library: 3.6.13." Software Heritage Archive. https://archive.softwareheritage.org/swh:1:snp:9d4ed12adc4b2bdfe96f9629cb99b1ab3c559261;origin=https://github.com/opencitations/time-agnostic-library.

———. 2025. "Configuration Files for ParaText: SHACL and Display Rules for Semantic Data Management Using HERITRACE." Zenodo. https://doi.org/10.5281/ZENODO.14741864.

Oldman, Dominic, and Diana Tanase. 2018. "Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace." In *The Semantic Web – ISWC 2018*, edited by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, 11137:325–40. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_20.

Pan, Jeff Z. 2009. "Resource Description Framework." In *Handbook on Ontologies*, 71–90. Springer.

Pareti, Paolo, and George Konstantinidis. 2022. "A Review of SHACL: From Data Validation to Schema Reasoning for RDF Graphs." In *Reasoning Web. Declarative Artificial Intelligence*,

edited by Mantas Šimkus and Ivan Varzinczak, 13100:115–44. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-95481-9_6.

Peroni, S., D. Shotton, and F. Vitali. 2016. "A Document-Inspired Way for Tracking Changes of RDF Data." In *Detection, Representation and Management of Concept Drift in Linked Open Data*, edited by L. Hollink, S. Darányi, A.M. Peñuela, and E. Kontopoulos, 26–33. Bologna: CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1799/Drift-a-LOD2016_paper_4.pdf.

Peroni, Silvio, and David Shotton. 2020. "OpenCitations, an Infrastructure Organization for Open Scholarship." *Quantitative Science Studies* 1 (1): 428–44. https://doi.org/10.1162/qss_a_00023.

Petz, Georg. 2023. "Linked Open Data. Zukunftsweisende Strategien." *Bibliothek Forschung Und Praxis* 47 (2): 213–22. https://doi.org/10.1515/bfp-2023-0006.

Rachinger, Johanna. 2008. "The Austrian National Library: A New Orientation in the Shadows of a Long History." *Alexandria: The Journal of National and International Library and Information Issues* 20 (1): 151–60. https://doi.org/10.1177/095574900802000105.

Ranjgar, Babak, Abolghasem Sadeghi-Niaraki, Maryam Shakeri, Fatema Rahimi, and Soo-Mi Choi. 2024. "Cultural Heritage Information Retrieval: Past, Present, and Future Trends." *IEEE Access* 12:42992–26. https://doi.org/10.1109/ACCESS.2024.3374769.

Salarelli, Alberto. 2016. "Gestire piccole collezioni digitali con Omeka: l'esperienza di MoRE (A Museum of REfused and unrealised art projects)." *Bibliothecae.it* Vol 5 (November):177-200 Paginazione. https://doi.org/10.6092/ISSN.2283-9364/6393.

Silveira, Lúcia da, Fabiano Couto Corrêa da Silva, Sara Caselani Zilio, and Larissa Silva Cordeiro. 2020. "Convergência de Práticas Linked Open Data Na Bibliothèque Nationale de France (BNF DATA)." *Revista ACB: Biblioteconomia Em Santa Catarina* 25 (1): 21–40.

WikiTeq. 2022. "Semantic Watchlist." https://www.mediawiki.org/wiki/Extension:Semantic_Watchlist.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Wulff, Enrique. 2024. "Digital Humanities: The Case Study of the National Library in Spain." In *Advances in Information Quality and Management*, edited by Mehdi Khosrow-Pour, D.B.A., 1–19. IGI Global. https://doi.org/10.4018/978-1-6684-7366-5.ch052.