ARTICLE

From Linked Open Data to Collections as Data

A Reproducible Framework Using Federated Queries

Meltem Dişli, Giulia Osti, Gustavo Candela, and Richard Zijdeman

ABSTRACT

Libraries are adopting Linked Open Data (LOD) and Collections as Data (CaD) approaches to present their collections as datasets for direct computational use. However, research focused on federated and reproducible access to these datasets is limited. This work aims to develop a federated and reproducible approach for extracting CaD from LOD repositories. In this context, data extracted from the single authors Jorge Juan y Santacilia and María de Zayas y Sotomayor, as well as from multiple authors from the Spanish Golden Age movement (1492–1659), are used as examples. Federated and reproducible queries are conducted using the Wikidata SPARQL public endpoint and three institutional LOD repositories on Jupyter Notebooks. The data are exported in a format compatible with computational tools (e.g., CSV) by focusing on works of a single author or works from a specific movement. Additionally, the work allows for the visualization of the queries. The results of this work provide a valuable framework for both digital humanities researchers working on datasets and libraries aiming to present their collections as accessible data for computational analysis.

INTRODUCTION

Libraries host rich digital materials in a wide diversity of formats and content, including video, audio, text, images, metadata, postcards, and biographies, among others. Advances in technology provide new opportunities for making these digital resources accessible and reusable. In this context, libraries are increasingly embracing both conceptual approaches, such as Collections as Data (CaD), which promote the idea of enabling data-driven scholarship, and Semantic Web technologies, such as Linked Open Data (LOD), to enable computational access to collections.

Tim Berners-Lee introduced the Semantic Web and LOD two decades ago as an extension of the traditional web to foster the use of machine-readable documentation using Resource Description Framework (RDF) and SPARQL.[1] In particular, libraries have adopted these technologies to describe and enrich their catalogs in which Wikidata has played a leading role.[2] Meanwhile, the CaD initiative has emerged as a set of guidelines, resources, and community-driven practices to support the transformation of cultural heritage (CH) collections into machine-readable, bulk-downloadable datasets. Grounded in library work and oriented toward multivocality, CaD encourages institutions to reconceptualize their collections, fostering critical engagement with the many layers of societal impact that could be generated from their reuse as data.

Leading institutions have adopted both the CaD and LOD approaches, at times combining them, making collections available as reusable datasets.[3] Most LOD repositories come with detailed

About the Authors

Dr. Meltem Dişli (dislimeltem@gmail.com; corresponding author) is Assistant Professor, Hacettepe University. **Giulia Osti** (giulia.osti@ucdconnect.ie) is PhD Candidate, University College Dublin. **Dr. Gustavo Candela** (gcandela@ua.es) is Lecturer, University of Alicante. **Dr. Richard Zijdeman** (richard.zijdeman@iisg.nl) is Head of Data & Augmentation Department, HSNDB, International Institute of Social History.

Submitted: 9 May 2025. Accepted for Publication: 1 August 2025. Published: 15 December 2025.

https://doi.org/10.5860/ital.v44i4.17432

documentation that supports reproducibility scenarios. Reproducibility—the ability to replicate an experiment to obtain similar results—is crucial for knowledge transfer and collaboration.[4] This aspect is valued not only by STEM disciplines but also by computational approaches to the humanities and the social sciences, as reproducibility can enhance the credibility and impact of research.[5] In this sense, Jupyter Notebooks has emerged as a powerful tool to provide reproducible code, including documentation in text format that can be hosted and run in cloud services.[6] However, while numerous LOD repositories are accessible through different services, there has been relatively limited work exploring how the information they hold can be reused through approaches that are both reproducible and federated—that is, querying and integrating distributed datasets across multiple SPARQL endpoints or application programming interfaces (APIs) in a coordinated and reproducible manner.[7]

The purpose of this work is to provide a reproducible and federated framework for reusing LOD in libraries, informed by the CaD principles. The principles outlined here aim to critically address the creation, circulation, and use of digitized and born-digital collections in a form that is structured for direct computational processing. Considering the Vancouver Statement on CaD, this effort aims at expanding the opportunities to engage with CaD (Principle 1), benefitting from the affordances offered by technical interoperability inherent to LOD (Principle 6).[8] It is grounded in best reproducibility practices—a foundational element of the statement, highlighting the importance of providing end users with all the necessary information to work with CaD in order to support responsible reuse. Last, it strives to disclose and critically comment on all steps taken, fostering transparency in our approach (Principle 5). In accordance with this purpose, this work aims to answer the following questions:

1. How can federated queries be used to extract and structure LOD as CaD?
2. What are the essential steps to enable a reproducible framework for the reuse of LOD as CaD?

The main contributions of this work are as follows: (1) a conceptualization of the key steps needed to enable the reuse of LOD from libraries within a CaD framework, and (2) a set of Jupyter Notebooks illustrating how CaD can be extracted via federated queries and prepared for computational reuse. The intended audiences of this work include library professionals interested in developing new workflows to support the reuse of their LOD repositories as CaD and scholars in the humanities and social sciences exploring computational methods.

The paper is structured as follows: following a brief overview of the current state of LOD and CaD in libraries, we introduce the framework employed for reusing LOD in libraries and subsequently detail our findings. Finally, the paper concludes with a summary of the adopted framework and outlines potential avenues for future research.

RELATED WORK

In recent years, significant progress has been made in enabling libraries to make their digital collections accessible and facilitate computational research. Initiatives have emerged to promote the publication and reuse of digital collections in innovative and creative ways, such as OpenGLAM, the International GLAM Labs Community, Research Data Alliance Collections as Data Interest Group, and AI4LAM.[9] European Commission has also been working toward the establishment of a common data space for CH collections.[10] Libraries have enabled the use of cutting-edge methods based on artificial intelligence and machine learning. In parallel, alongside the FAIR (Findable, Accessible, Interoperable, Reusable) and the CARE (Collective benefit,

Authority to control, Responsibility, Ethics) principles, approaches such as CaD foster the publication of digital collections, supporting computational and responsible use.[11]

The Always Already Computational: Collections as Data project, followed by Collections as Data: Part to Whole, have played a pivotal role in making library collections machine-readable and computationally processable.[12] As a result of these initiatives, numerous institutions, including the Library of Congress, the British Library, the National Library of Scotland, and the National Library of the Netherlands, have begun providing access to their collections as data. These efforts have led to the development of a workflow, checklists, a conceptual model, and implementation examples that guide libraries in transforming their collections into CaD.[13] Such initiatives are crucial to the creation of data platforms, secure environments in which data can be shared and reused following best practices and according to the policies and rules established by organizations[14]—such as the common European data space for CH collections. Another crucial initiative for enhancing the accessibility and reusability of CH collections is LOD. Libraries have explored the application of the Semantic Web to their catalogs to create rich knowledge graphs in the form of interconnected nodes representing facts (e.g., Miguel de Cervantes is an author).

Some examples include the National Library of Spain (BNE), the National Library of the Netherlands, the Library of Congress, the Biblioteca Virtual Miguel de Cervantes (BVMC), and the National Library of France (BNF). These institutions have used different data modeling approaches to describe their metadata. For instance, the Library of Congress employed BIBFRAME, the National Library of France used the IFLA Library Reference Model,[15] and the National Library of the Netherlands used Schema.org. Other approaches, such as the LUX: Yale Collections Discovery project, aggregate content from several data sources, describe the metadata using conceptual models for CH content such as CIDOC Conceptual Reference Model (CRM), and make the collections available as LOD. Additionally, dedicated APIs and SPARQL endpoints provide access to the data, enabling federated queries across multiple repositories.[16]

To publish linked data, institutions have explored external resources to enrich and interlink their repositories. Examples of commonly used external resources include GeoNames, the Virtual International Authority File (VIAF), and Wikidata, which has gained widespread adoption in the library domain as a collaborative, community-driven platform for open data curation.[17] Beyond Wikidata, platforms such as the Social Sciences and Humanities Open Marketplace can also support the collaborative enrichment of datasets and publications.[18]

In terms of enabling reuse and following best practices promoted by the International GLAM Labs Community, practices including visualization, data quality analysis, and metadata diversity analysis play a crucial role.[19] Instead of being end goals in themselves, these function as indicators to evaluate dataset reusability and help surface structural qualities that can support different types of reuse.

These efforts provide an extensive demonstration of how LOD made available by libraries can be published and reused. Our contribution is a practical implementation, underlining that LOD and CaD should be considered complementary. Through the examples presented (e.g., retrieving multiple works by a single author across different repositories, or mapping the birthplaces of Spanish Golden Age authors), we show how a federated and reproducible framework for reusing the LOD repositories as CaD can enhance the accessibility, reuse, interoperability, reproducibility, and transparency of CH data. By providing reproducible code in Jupyter Notebooks, we also illustrate the challenges of vocabulary alignment across diverse endpoints, sparking a discussion

on the socio-technical dimension of federated queries. Responding to previous work, our work not only encourages institutions to explore new uses of their data but also demonstrates potential ways in which collections can be queried to support collaborative research and curation.[20] In this sense, our framework serves as a bridge among curators, researchers, technical experts, and administrators, fostering communication and cooperation across these roles.

A REPRODUCIBLE FRAMEWORK TO EXTRACT CAD FROM LOD USING FEDERATED QUERIES

This section describes the conceptual items part of the reproducible and federated framework we propose in this piece to enable the reuse of LOD in libraries. Figure 1 illustrates the main steps described in the following sections.

Figure 1. The proposed framework to extract and enable the reuse of linked open data as collections as data.

Identification

The first step corresponds to the identification of the content that could be suitable to craft a CaD from LOD. Identification can be guided by thematic relevance—such as historical periods (e.g., World War I[21]), specific authors (e.g., the Spanish mariner Jorge Juan Santacilia [https://www.cervantesvirtual.com/portales/jorge_juan_santacilia/]), or particular events (e.g., the Spanish Civil War exile [https://www.cervantesvirtual.com/portales/exilio/])—but must take into account data accessibility, structure, and their relation to a specific media type. For instance, *Chronicling America* provides access to historical newspapers with structured metadata, containing records on events like the *Titanic* sinking or the Alaska purchase (https://guides.loc.gov/chronicling-america-topics). Portals like *History Lab* offer robust materials, such as Henry Kissinger's phone conversations and US declassified documents, or more basic data, such as access to their bibliographic metadata. In this step, it is crucial to evaluate whether a collection is structured, machine-readable, and accessible through interoperable services. In this context, beyond features such as licenses, formats, the ontologies used to describe the content, the annotations provided, or collection size, the availability of a SPARQL endpoint or an API is critical.

Extraction

When the digital collections are available as LOD, a SPARQL endpoint might be available to access the data. Even if the content is not provided in the form of LOD, there are existing tools, such as Cow, LDWizard, OpenRefine, and RML, to transform the information into LOD.[22] SPARQL enables users to define federated queries by using the command SERVICE to search across several repositories. For instance, a researcher might be interested in retrieving from a particular repository the information about a writer (e.g., works, date of birth, name, and external identifiers) and cross-comparing this information with what is held in other repositories. Wikidata has emerged as a powerful repository in the galleries, libraries, archives, and museums (GLAM) sector, providing a mature infrastructure that enables scalability and long-term sustainability.[23] Wikidata public SPARQL endpoint makes it possible to run federated queries across repositories. Figure 2 shows an example of a federated SPARQL query to retrieve data from Wikidata and the BNE. The query starts from an entity in Wikidata, the author Jorge Juan Santacilia (identified by Q2085725), and pairs it with its corresponding identifier (P950) from the BNE. The SERVICE instruction enables querying the BNE SPARQL endpoint, retrieving all the works associated with the entity, along with their labels.

Figure 2. Example of a federated SPARQL query.

```
PREFIX bne-def: <https://datos.bne.es/def/>
SELECT * WHERE {
  wd:Q2085725 wdt:P950 ?id
  BIND( uri ( concat ( "https://datos.bne.es/resource/" , ?id )) as ?bneID )
  SERVICE <http://datos.bne.es/sparql> {
    ?bneID bne-def:OP5001 ?work .
    ?work rdfs:label ?label .
  }
}
```

In addition, when the information is available through a static website (https://www.cervantesvirtual.com/portales/jorge_juan_santacilia/imagenes_jorge_juan/), it can be transformed into RDF by using software libraries such as RDFLib and Apache Jena. Structured data can be created using vocabularies such as Schema.org, FOAF, or CIDOC CRM.[24] Reusing existing vocabularies is at the heart of the LOD philosophy. Online indexes such as the Awesome Humanities Ontologies or Linked Open Vocabularies help to educate about existing vocabularies. The latter already provides reuse statistics as a recommendation feature, with more full-fledged vocabulary recommenders underway (e.g., CLARIAH's Vocabulary Recommender). On the policy level, recommendations for specific vocabularies are made, such as in the Netherlands' Digital Heritage Reference Architecture (DERA), where Schema.org is recommended as the first go-to vocabulary.[25]

Publication

Once the dataset has been validated, the next step involves making it available through appropriate platforms. However, publication should go beyond simply making data available online—it should be about publishing data with reuse in mind. Datasets can be published through platforms such as Zenodo, the Social Sciences and Humanities Open Marketplace, Wikidata, or GitHub, which support persistent identifiers, rich metadata, and various licensing options. To foster reuse, previous work has also defined guidelines and best practices to publish the datasets, including compiling a README file and making available documentation regarding data scopes

and limitations.[26] Emerging approaches, such as the various proposals concerning datasheet and data envelopes for CH datasets, recommend publishing information on provenance, intended and known uses, potential biases, and creators, among others.[27]

Enabling Reuse

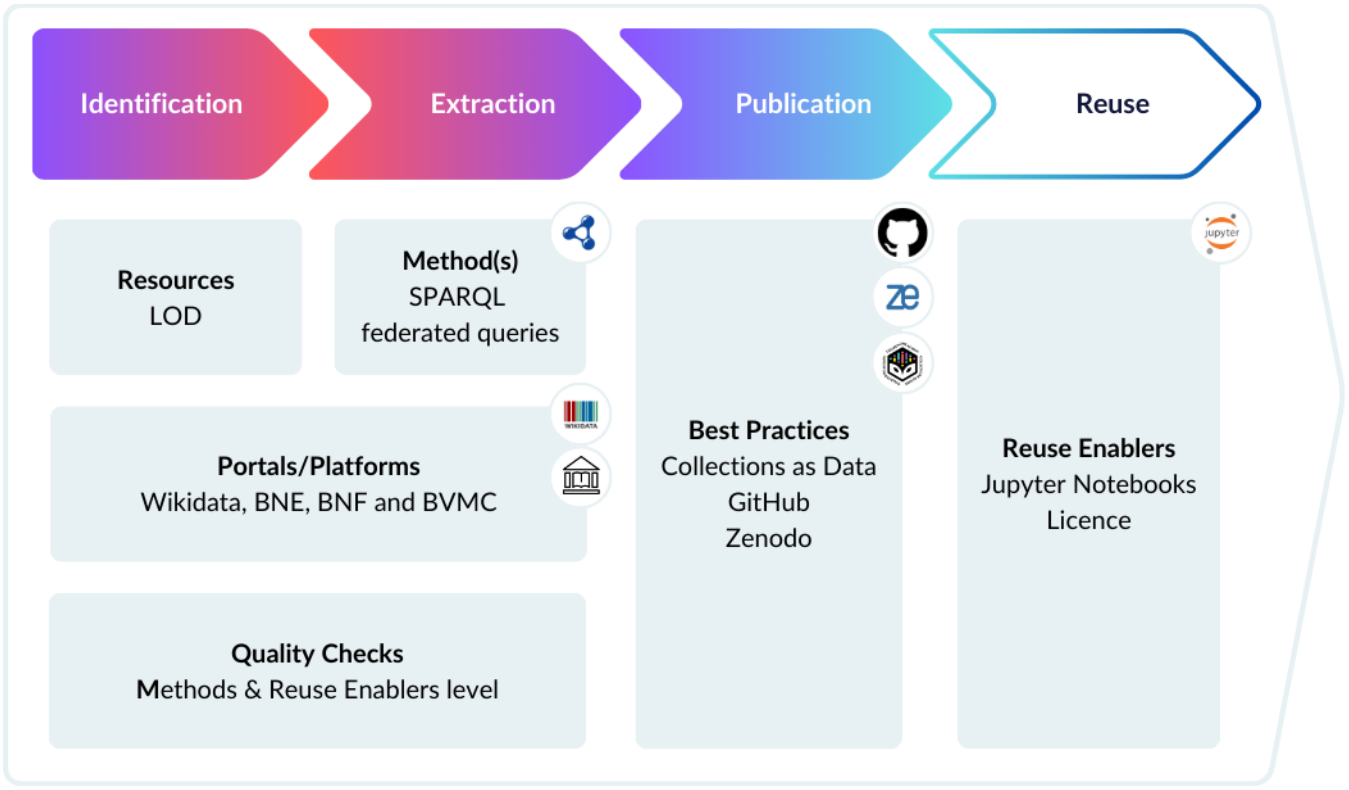
The final step in the framework focuses on enabling reuse on a deeper level, or making it actionable, thereby supporting technical workflows and interpretive engagements. Several conditions make reuse actionable, and often the chosen approach reflects different levels of complexity depending on "*who the work is done for.*"[28] One key strategy involves providing reproducible workflows through Jupyter Notebooks,[29] combining code, documentation, and examples of how data can be transformed and analyzed.

LOD is published with variable levels of consistency, completeness, and semantic alignment. Therefore, quality checks are essential. Consistent and high-quality data directly contribute to interoperability and enhance reusability, especially in cross-institutional or federated contexts, as evidenced by previous work assessing LOD repositories in different domains, including the CH sector.[30] In this context, it is important that datasets meet data quality criteria such as availability, interlinking, consistency, completeness, accuracy, interoperability, and relevancy.[31]

RESULTS

This section presents the application of the framework proposed in this work to a selection of LOD repositories made available by relevant GLAM organizations, providing details about the implementation and the scope of this approach. A conceptualization of the framework architecture is presented in Figure 3.

Figure 3. Framework application architecture. LOD is linked open data.



In this case study, we explore the use of federated queries through SPARQL endpoints to retrieve and pre-process publication data related to Jorge Juan y Santacilia (1713–1773), María de Zayas y Sotomayor (1590–1661), and four other authors part of the Spanish Golden Age movement. For this purpose, we used Wikidata SPARQL public endpoint and three institutional LOD repositories, namely the BNE, BNF, and BVMC. The queries were tested and collated in Python with SPARQLWrapper along with other packages to support queries parallelization and improved requests handling, such as async.io and aiohttp, using Jupyter Notebooks to ensure reproducibility and enable reuse scenarios, following current best practices in their redaction.[32] All the materials have been published on GitHub (available at https://github.com/semanticnoodles/federated-cad) and Zenodo (available at https://doi.org/10.5281/zenodo.15346714). The notebooks are configured to run directly in Binder, a free, cloud-based service that allows users to browse the repository and launch interactive Jupyter Notebooks in a sandbox environment.

The design of the queries follows a specific rationale, illustrating the steps to retrieve a single author's works and multiple authors' works based on a movement property (https://www.wikidata.org/wiki/Property:P135), as unpacked in the following thematic sections. The query results are put together and exported as CSV files to obtain CaD in its simplest form. In terms of quality, our approach is guided by several of the key criteria mentioned in the Enabling Reuse subsection. Availability is ensured by relying on publicly accessible SPARQL endpoints and open licensed datasets. Interoperability is primarily achieved using Wikidata identifiers, which serve as common reference points across repositories. Accuracy is supported through normalization steps, particularly in standardizing author names and incorporating language codes, although further disambiguation and data cleaning steps are required. Relevance is considered during the query design phase, ensuring the data selection is aligned with the specific context of the application. Last, interpretability is addressed by structuring the results consistently, following the aforementioned principles from CaD.

Query-building Rationale

In the context of LOD, building a SPARQL query is not simply a technical task, but a dual process that combines design logic with research intent. While collecting data about the literary works attributed to an author from a specific national repository, the query should trace through persistent identifiers and fetch structured metadata like titles, editions, places, and date of publication. Considering the query-building examples we provide in our repository for BNE, BNF and BVMC, they essentially embody what federated access is about—using SERVICE clauses to retrieve fine-grained, institution-specific bibliographic records. The use of OPTIONAL clauses is critical as entities may not detail every property uniformly; by marking edition, language, or place of publication as optional, the query is prevented from halting in the case of missing data. Such modularity allows an easy recombination of query blocks for different use cases. Furthermore, avoiding excessive nesting—setting a LIMIT value to control result size—and reducing overly broad UNION patterns can prevent server overload and reduce the chances of getting timeouts.

Although Wikidata offers a centralized and standardized entry point through authority identifiers, the current structure and its semantics may considerably vary across repositories, an aspect that emerges clearly from the notebooks accompanying this paper. Because each institution sets up ontologies, property relationships, and descriptive practices according to its needs, query formulation becomes a highly repository-specific task. Accessing data from these sources requires direct experience working with a repository endpoint, other than familiarity with the vocabularies in use in that context.

Single Author (Multiple Works) Queries

Once the SPARQL queries are crafted, matching the syntax requirements of each repository, they are ready to be integrated into a streamlined workflow that enables the retrieval and consolidation of bibliographic metadata. This is the approach we applied to retrieve the metadata for both Jorge Juan y Santacilia and María de Zayas y Sotomayor. By querying distributed SPARQL endpoints using their respective Wikidata identifiers, we collate metadata such as titles, editions, publication details, and languages into unified datasets. The process is designed to be scalable yet pragmatic: While a result limit is applied, it remains sufficiently generous (1,000 datasets) given the typical output range for a single historical author.

Figure 4 shows the distribution of languages in which the works of María de Zayas y Sotomayor and Jorge Juan y Santacilia are available, based on data from the LOD repositories included in the work. The findings indicate that 71% of María de Zayas y Sotomayor's works are published in Spanish, although there are also works in English, French, and Italian. In the case of Jorge Juan, his works appear in Spanish, French, English, German, and Dutch. The presence of their works in multiple European languages beyond Spanish suggests that both authors have readerships extending beyond Spain.

Figure 4. Publication languages of the works of María de Zayas y Sotomayor and Jorge Juan y Santacilia.

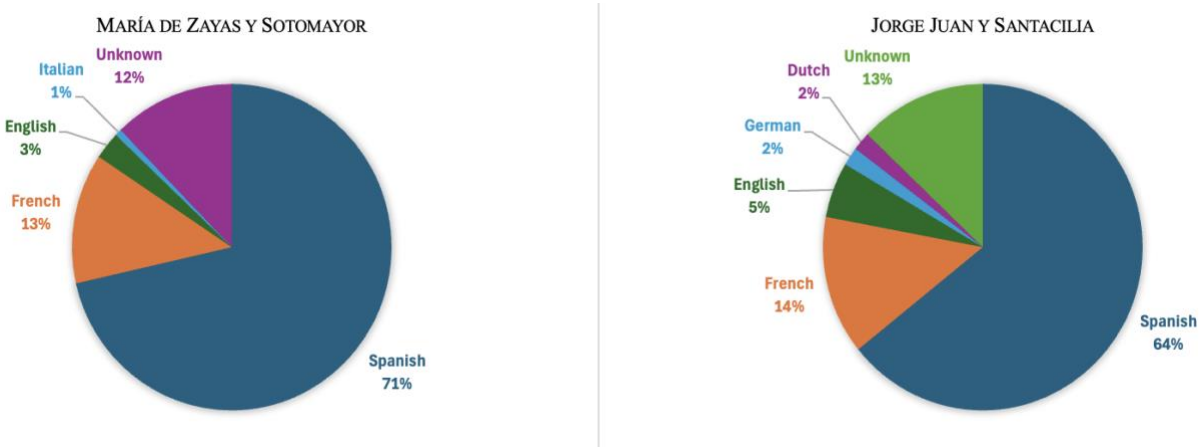
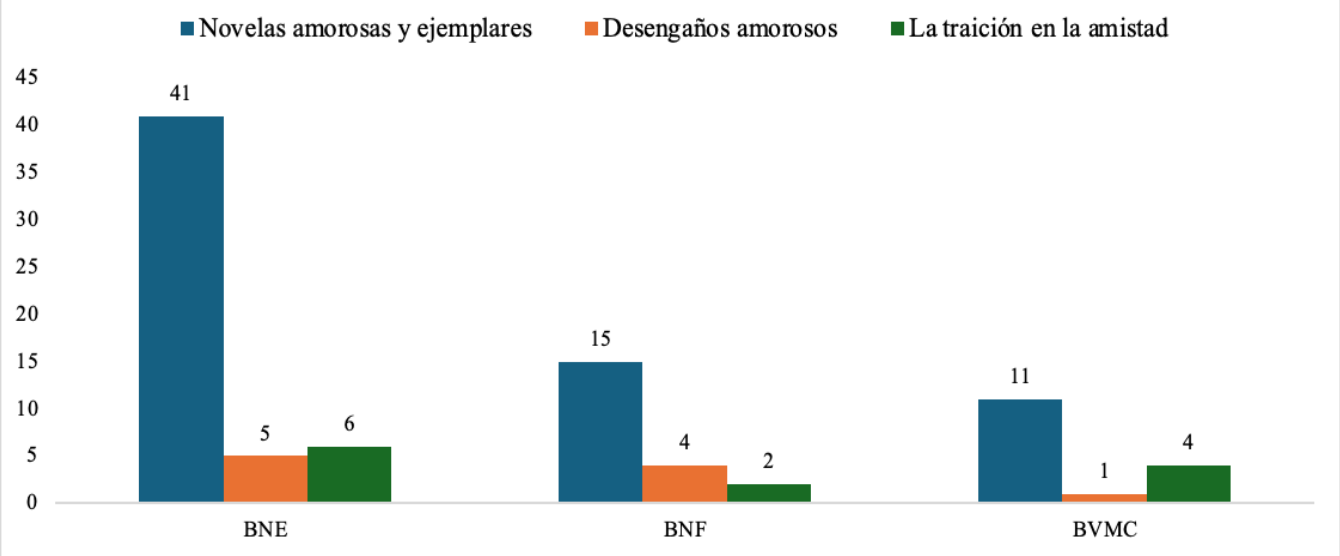


Figure 5 illustrates the distribution of the most frequently found works of María de Zayas y Sotomayor in the BNE, BNF, and BVMC LOD repositories. In these datasets, works may appear under different titles or in various forms. For instance, the work *Desengaños amorosos* is indexed as *Parte segunda del sarao y entretenimiento honesto* in some metadata. After deduplicating works listed under different titles, their frequencies were calculated. In all three LOD repositories, *Novelas amorosas y ejemplares* is the most frequently found work. The frequencies of the other two works are similar, with *La traición en la amistad* ranking second in both the BNE and BVMC and *Desengaños amorosos* ranking second in the BNF.

Figure 5. Common works of María de Zayas y Sotomayor in the selected linked open data repositories: National Library of Spain (BNE), National Library of France (BNF), and Biblioteca Virtual Miguel de Cervantes (BVMC).



Multiple Authors (Multiple Works) Discovery Query

Drawing from the workflow established for single authors, we attempted to perform a discovery-oriented task centered on a broader cultural context, the Spanish Golden Age. Rather than start with a predefined list of authors, the process begins with a SPARQL query to identify all writers associated with this movement as mapped in Wikidata; this demonstrates how federated queries can consolidate dispersed works. Once that information is obtained, a subset of five authors who are all represented in our chosen repositories (listed in Table 1) was selected for further analysis. Their identifiers are extracted and used to populate dynamic query templates, which enable the systematic interrogation of each repository for bibliographic metadata about these figures and their works.

Table 1. Authors from the Spanish Golden Age whose works are mapped across our chosen repositories: National Library of Spain (BNE), National Library of France (BNF), and Biblioteca Virtual Miguel de Cervantes (BVMC)

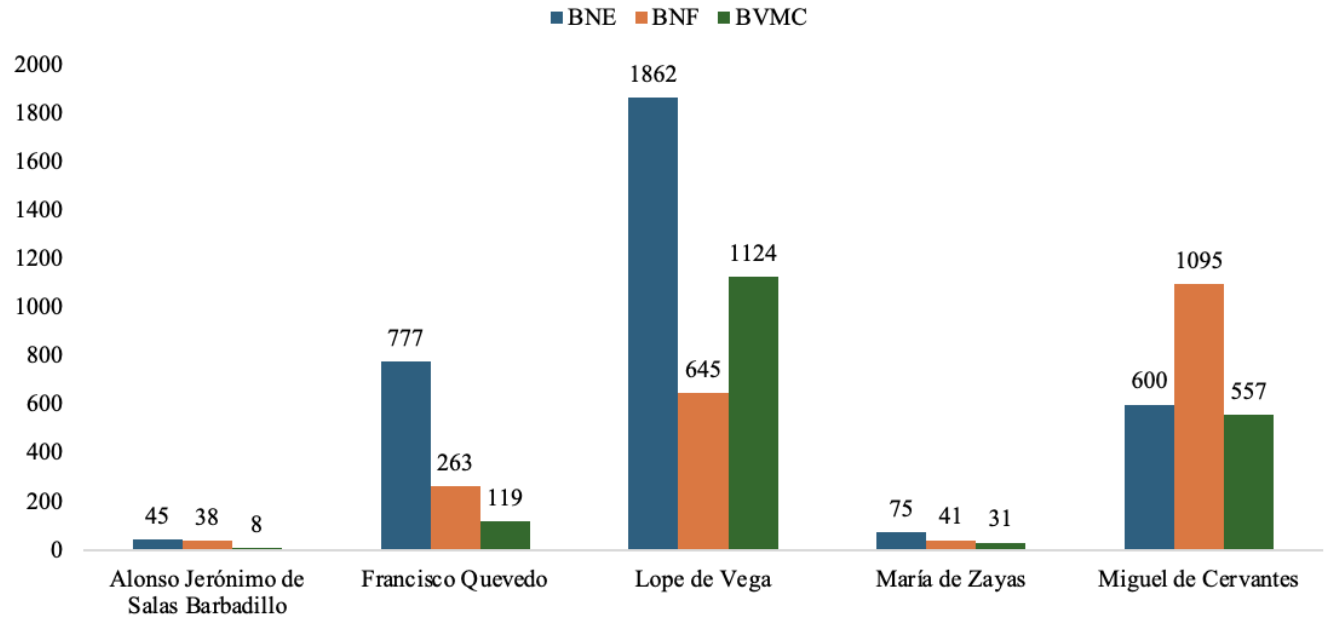
Author	Wikidata ID	BNE ID	BNF ID	BVMC ID
Alonso Jerónimo de Salas Barbadillo	Q3893270	XX1054399	12054 809	895
Francisco Quevedo	Q201315	XX1066651	118873287	6
Lope de Vega	Q165257	XX1719671	11927819k	72
María de Zayas y Sotomayor	Q2905689	XX913942	119878263	105
Miguel de Cervantes	Q5682	XX1718747	118957747	40

Author names may appear in various forms across metadata records created by different institutions, which can lead to inconsistencies in how information about the same individual is represented across different institutional datasets. Therefore, these identifiers play a critical role in enabling researchers to access accurate data and establish consistent links between various data sources.

Due to the significantly larger scope of this dataset, spanning multiple authors and potentially hundreds of records per repository, the workflow presented in the corresponding notebook introduces more sophisticated mechanisms for query execution and data handling. Result limits had to be reduced substantially to avoid overwhelming endpoints, and pagination was implemented to retrieve complete datasets in manageable chunks. To maintain efficiency, the querying process was parallelized using asynchronous functions, allowing multiple requests to run concurrently while managing retries and rate limits. These asynchronous queries are further streamlined through a flexible templating system, allowing for dynamic customization across authors and repositories.

Figure 6 illustrates the datasets for Spanish Golden Age authors according to selected LOD repositories. As seen, both BNE and BVMC have the largest number of datasets by Lope de Vega, while BNF holds more datasets by Miguel de Cervantes. It is also noticeable that the datasets of Alonso Jerónimo and María de Zayas y Sotomayor are significantly fewer across all LOD repositories compared to other authors. These findings could be useful for libraries looking to develop their collections around specific authors, as well as for researchers focused on a particular author when selecting a library for their studies.

Figure 6. Dataset composition for the Spanish Golden Age movement: Number of records per author by linked open data repositories: National Library of Spain (BNE), National Library of France (BNF), and Biblioteca Virtual Miguel de Cervantes (BVMC).



DISCUSSION

While the framework conceptualized in this work demonstrates the potential of federated SPARQL querying across LOD repositories, its current implementation is limited in scope and subject to several constraints. First, not all libraries provide access to LOD versions of their bibliographic records, and among those that do, not all are linked or represented through properties in Wikidata. This affects the discoverability and interoperability of data across collections, as seen in our multiple authors test case, where we focused on a subset of five well-represented authors. However, in other cases, alternative control systems such as VIAF identifiers could serve as a

valuable fallback for entity reconciliation and reuse. Second, the dynamic and evolving nature of LOD repositories presents challenges for reproducibility. Changes in data structure, content, or availability over time can affect both the consistency of query results and, in the long run, the reliability of this workflow. Similarly, fluctuations in the availability or stability of the SPARQL endpoints can also affect automated retrieval. Technical challenges also emerged. For example, inconsistencies in metadata quality, such as malformed publication year entries (e.g., "1806]" or "1941]"), can hinder data normalization and analysis. Moreover, the number of results retrieved via SPARQL queries may need capping to prevent timeouts in complex workflows such as the one presented for the Spanish Golden Age, which may lead to incomplete datasets in cases where more extensive records exist.

Another significant constraint is the heterogeneity of the ontologies employed across institutions. This complicates query formulation and increases the need for custom handling of repository-specific schemas to provide the metadata. Compounding this is the limited availability of documentation and practical examples for federated SPARQL queries, which constrain broader adoption and experimentation. More accessible information and documented examples would substantially lower the entry barrier for researchers and practitioners seeking to reuse LOD resources.

Finally, the test run of the proposed framework is limited in scope to the employment of LOD repositories made available by libraries, strategically using Wikidata to search across different repositories. While additional work is required to reuse the obtained CSV datasets, particularly for public-facing applications, this effort can be streamlined with tools like CollectionBuilder. Rooted in the CaD movement, CollectionBuilder enables the creation of interactive, metadata-driven websites using CSV files and minimal infrastructure. However, to make effective use of such tools, some post-processing is necessary—for example, handling non-standardized language values, reviewing missing identifiers, or normalizing field structures to match the expected input schema.

CONCLUSIONS

Building on previous research, we outlined a federated and reproducible framework to extract CaD from LOD repositories made available by libraries, enabling their reuse. We performed a test run on a selection of three LOD repositories. Our evaluation suggests that this framework can serve both as a practical tool and a conceptual model to illustrate how to leverage the use of federated queries for DH practitioners and data curators. By doing so, we have illustrated the compatibility and integrability of the LOD and CaD approaches.

Another takeaway of our work is that perhaps too often and too quickly practitioners point to the technical barriers of federated SPARQL queries. As a community, we need documentation of data models, more strict safeguarding of data quality, and social contracts on the use of shared vocabularies as a first step to move decentralized knowledge sharing forward.

Future work to be explored includes the refinement of the framework using detailed provenance and integration beyond Wikidata. It is virtually impossible to add identifiers for all libraries and archives for all kinds of entities. Convergence toward shared value lists, such as VIAF, represented as LOD at local endpoints, would allow for a far more efficient retrieval of information from a more diverse set of heritage institutes, enhancing the representation of digital heritage as collections in the form of data.

ACKNOWLEDGMENTS

We would like to thank the Collections as Data initiative, the International GLAM Labs Community, the Research Data Alliance Collections as Data Interest Group, and the GLAM Workbench. This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real), under Grant No. 18/CRT/6224.

ENDNOTES

[1] Richard Cyganiak, David Wood, and Markus Lanthaler, eds., "RDF 1.1 Concepts and Abstract Syntax," World Wide Web Consortium, February 2014, https://www.w3.org/TR/rdf11-concepts/; Tim Berners-Lee, James Hendler, and Olli Lassila, "The Semantic Web in Scientific American," *Scientific American Magazine* 284, no. 5 (May 2001): 34–43, https://doi.org/10.1038/scientificamerican0501-34.

[2] Gustavo Candela et al., "A Systematic Review of Wikidata in GLAM Institutions: A Labs Approach," in *Linking Theory and Practice of Digital Libraries—28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Ljubljana, Slovenia, September 24–27, 2024, Proceedings, Part II*, ed. Apostolos Antonacopoulos et al., 34–50 (Springer, 2024), https://doi.org/10.1007/978-3-031-72440-4_4.

[3] Chris Dijkshoorn et al., "The Rijksmuseum Collection as Linked Data," *Semantic Web* 9, no. 2 (2018): 221–30, https://doi.org/10.3233/SW-170257; Gustavo Candela, "An Automatic Data Quality Approach to Assess Semantic Data from Cultural Heritage Institutions," *Journal of the Association for Information Science and Technology* 74, no. 7 (2023): 866–78, https://doi.org/10.1002/asi.24761; Gustavo Candela, María Dolores Sáez, MPilar Escobar Esteban, and Manuel Marco-Such, "Reusing Digital Collections from GLAM Institutions," *Journal of Information Science* 48, no. 2 (2022): 251–67, https://doi.org/10.1177/0165551520950246; Sally Chambers et al., "Position Statements: Collections as Data: State of the Field and Future Directions" (Zenodo, 2023), https://doi.org/10.5281/zenodo.7897735.

[4] Chris Drummond, "Reproducible Research: A Minority Opinion," *Journal of Experimental and Theoretical Artificial Intelligence* 30, no. 1 (2018): 1–11, https://doi.org/10.1080/0952813X.2017.1413140; Melanie Feinberg et al., "The New Reality of Reproducibility: The Role of Data Work in Scientific Research," *Proceedings of the ACM on Human-Computer Interaction* 4, no. CSCW1 (May 28, 2020): 1–22, https://doi.org/10.1145/3392840.

[5] Fiona Fidler and John Wilcox, "Reproducibility of Scientific Results," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Metaphysics Research Lab, 2021), https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/; Feinberg et al., "The New Reality of Reproducibility"; Drummond, "Reproducible Research."

[6] Gustavo Candela, Sally Chambers, and Tim Sherratt, "An Approach to Assess the Quality of Jupyter Projects Published by GLAM Institutions," *Journal of the Association for Information Science and Technology* 74, no. 13 (2023): 1550–64, https://doi.org/10.1002/asi.24835; Vernon Gayle and Roxanne Connelly, "The Stark Realities of Reproducible Statistically Orientated Sociological Research: Some Newer Rules of the Sociological Method,"

Methodological Innovations 15, no. 3 (2022): 207–21, https://doi.org/10.1177/20597991221111681.

[7] Feinberg et al., "The New Reality of Reproducibility"; Guillermo Vega-Gorgojo, "LOD4Culture: Easy Exploration of Cultural Heritage Linked Open Data," *Semantic Web* 15, no. 5 (2024): 1563–92, https://doi.org/10.3233/SW-233358; Jonathan Blaney, "Introduction to the Principles of Linked Open Data," *Programming Historian* 6 (2017), https://programminghistorian.org/en/lessons/intro-to-linked-data.

[8] Thomas Padilla, Hannah Scates Kettler, Stewart Varner, and Yasmeen Shorish, "Vancouver Statement on Collections as Data" (Zenodo, 2023), https://doi.org/10.5281/zenodo.8342171.

[9] "OpenGLAM Principles," OpenGlam, https://openglam.org/principles/; Mahendra Mahey et al., *Open a GLAM Lab* (International GLAM Labs Community, 2019), https://doi.org/10.21428/16ac48ec.f54af6ae; "Collections as Data Interest Group," Research Data Alliance, https://www.rd-alliance.org/groups/collections-as-data-ig/activity/; "Artificial Intelligence for Libraries, Archives, and Museums," AI4LAM, https://sites.google.com/view/ai4lam.

[10] "Commission Recommendation of 10.11.2021 on a Common European Data Space for Cultural Heritage" (European Commission, 2021), https://digital-strategy.ec.europa.eu/en/news/commission-proposes-common-european-data-space-cultural-heritage.

[11] Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3, no. 1 (2016): article 160018, https://doi.org/10.1038/sdata.2016.18; Stephanie Russo Carroll et al., "The CARE Principles for Indigenous Data Governance," *Data Science Journal* 19 (2020): 43, https://doi.org/10.5334/dsj-2020-043; Padilla et al., "Vancouver Statement on Collections as Data."

[12] Thomas Padilla et al., "Always Already Computational: Collections as Data" (Zenodo, 2019), https://doi.org/10.5281/zenodo.3152935; "Collections as Data Futures: A Recap, A Resource, Next Steps," Collections as Data – Part to Whole, May 4, 2023, https://collectionsasdata.github.io/part2whole/recap/.

[13] Gustavo Candela, Sally Chambers, and Alba Irollo, "A Workflow to Publish Collections as Data: The Case of Cultural Heritage Data Spaces," Social Sciences & Humanities Open Marketplace, 2023, https://marketplace.sshopencloud.eu/workflow/I3JvP6; Benjamin Charles Germain Lee, "The 'Collections as ML Data' Checklist for Machine Learning and Cultural Heritage," *Journal of the Association of Information Science Technology* 76, no. 2 (2025): 375–96, https://doi.org/10.1002/ASI.24765; Meltem Dişli, "Veri Olarak Kültürel Miras Koleksiyonları [Cultural Heritage Collections as Data]" (PhD diss., Hacettepe University, 2024), https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/34990.

[14] Milena Dobreva, Krassen Stefanov, and Krassimira Ivanova, "Data Spaces for Cultural Heritage: Insights from GLAM Innovation Labs," in *From Born-Physical to Born-Virtual: Augmenting*

Intelligence in Digital Libraries, ed. Yuen-Hsien Tseng, Marie Katsurai, and Hoa N. Nguyen, 492–500 (Springer, 2022), https://doi.org/10.1007/978-3-031-21756-2_41.

[15] Consolidation Editorial Group of the IFLA FRBR Review, Pat Riva, Patrick Le Boeuf, and Maja Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information* (International Federation of Library Associations and Institutions, 2025), https://repository.ifla.org/handle/20.500.14598/40.2.

[16] "Wikidata:SPARQL Query Service/Federation Report," Wikidata, last edited October 2, 2025, https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Federation_report.

[17] Gustavo Candela, Pilar Escobar, Rafael C. Carrasco, and Manuel Marco-Such, "A Linked Open Data Framework to Enhance the Discoverability and Impact of Culture Heritage," *Journal of Information Science* 45, no. 6 (2019): 756–66, https://doi.org/10.1177/0165551518812658; Candela et al., "A Systematic Review of Wikidata in GLAM Institutions."

[18] Laure Barbot et al., "Contextualizing Research Tools & Services Through Workflows in the SSH Open Marketplace," *Journal of Open Humanities Data* 10 (2024): 22, https://doi.org/10.5334/johd.192.

[19] Gustavo Candela, "Browsing Linked Open Data in Cultural Heritage: A Shareable Visual Configuration Approach," *Journal of Computational Cultural Heritage* 18, no. 1 (2024): article 9, https://doi.org/10.1145/3707647; Candela, "An Automatic Data Quality Approach"; Rafael C. Carrasco, Gustavo Candela, and Manuel Marco-Such, "Measuring the Diversity of Data and Metadata in Digital Libraries" (arXiv, 2023), https://doi.org/10.48550/ARXIV.2301.01193.

[20] Gayle and Connelly, "The Stark Realities of Reproducible Statistically Orientated Sociological Research"; Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship."

[21] Mikko Koho et al., "WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data," *Semantic Web* 12, no. 2 (2021): 265–78, https://doi.org/10.3233/SW-200392.

[22] Gustavo Candela, "Towards a Semantic Approach in GLAM Labs," *Journal of Information Science*, advance online publication (2023), https://doi.org/10.1177/01655515231174386; Gustavo Candela et al., "An Ontological Approach for Unlocking the Colonial Archive," *Journal on Computing and Cultural Heritage* 16, no. 4 (2023): article 74, https://doi.org/10.1145/3594727.

[23] Candela et al., "A Systematic Review of Wikidata in GLAM Institutions"; "SPARQL Endpoint Interface to Python," SPARQLWrapper, http://rdflib.github.io/sparqlwrapper.

[24] Candela, "Towards a Semantic Approach in GLAM Labs"; Marilena Daquino et al., "Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data," *ACM Journal on Computing and Cultural Heritage* 10, no. 4 (2017): article 21, https://doi.org/10.1145/3051487.

[25] Bram Gaakeer et al., "Digitaal Erfgoed Referentie Architectuur (DERA) – Versie 4.0" (Zenodo, 2021), https://doi.org/10.5281/zenodo.5562062.

[26] Gustavo Candela et al., "A Checklist to Publish Collections as Data in GLAM Institutions," *Global Knowledge, Memory and Communication* 74, nos. 5–6 (2023): 1323–55, https://doi.org/10.1108/GKMC-06-2023-0195.

[27] Lee, "The 'Collections as ML Data' Checklist for Machine Learning and Cultural Heritage"; Henk Alkemade et al., "Datasheets for Digital Cultural Heritage Datasets," *Journal of Open Humanities Data* 9, no. 1 (2023): 17, https://doi.org/10.5334/johd.124.

[28] Thomas Padilla, "On a Collections as Data Imperative" (UC Santa Barbara, 2017), https://escholarship.org/uc/item/9881c8sv.

[29] Tim Sherratt, "GLAM-Workbench/Recordsearch" (Zenodo, 2023), https://doi.org/10.5281/zenodo.7553047.

[30] Gustavo Candela, Pilar Escobar, María Dolores Sáez, and Manuel Marco-Such, "A Shape Expression Approach for Assessing the Quality of Linked Open Data in Libraries," *Semantic Web* 14, no. 2 (2023): 159–79, https://doi.org/10.3233/SW-210441; Nora Abdelmageed and Lois Hutubessy, "A Systematic Approach towards Higher Quality Linked Open Data at Nieuwe Instituut," *SEMANTiCS—20th International Conference on Semantic Systems* 3795 (2024): paper 9, https://ceur-ws.org/Vol-3759/paper9.pdf.

[31] Amrapali Zaveri, et al., "Quality Assessment for Linked Data: A Survey: A Systematic Literature Review and Conceptual Framework," *Semantic Web* 7, no. 1 (2015): 63–93, https://doi.org/10.3233/SW-150175.

[32] "SPARQLWrapper"; "Asyncio: Reference Implementation of PEP 3156," Python, accessed April 7, 2025, http://www.python.org/dev/peps/pep-3156/; "Aiohttp: Async Http Client/Server Framework (Asyncio)," MacOS :: MacOS X, Microsoft :: Windows, POSIX, Python, accessed April 7, 2025, https://github.com/aio-libs/aiohttp; Candela, Chambers, and Sherratt, "An Approach to Assess the Quality of Jupyter Projects Published by GLAM Institutions."