

Article

A Reproducible Framework to Extract Multimodal Collections as Data in Museums: A Wikidata Approach

Gustavo Candela ^{1,2,*} , Christel Annemieke Romein ^{3,4,5}  and Julie M. Birkholz ^{6,7} 

¹ Departamento de Lenguajes y Sistemas Informáticos, University of Alicante, 03690 San Vicente del Raspeig, Spain

² Wikimedia España, 47008 Valladolid, Spain

³ HAICu-Project, KiTes-BMS, University of Twente, 7522 NB Enschede, The Netherlands; c.a.romein@utwente.nl

⁴ Walter Benjamin Kolleg, University of Bern, 3012 Bern, Switzerland

⁵ READ-COOP SCE, 6020 Innsbruck, Austria

⁶ Department of History and GhentCDH, Ghent University, 9000 Ghent, Belgium; julie.birkholz@ugent.be

⁷ KBR-Royal Library of Belgium, 1000 Brussels, Belgium

* Correspondence: gcandela@ua.es

Abstract

GLAM (Gallery, Library, Archive, and Museum) institutions have been making digital collections available for decades. New initiatives to publish, preserve, and reuse data have emerged, with collaborative projects such as Wikidata playing a significant role. This work provides a framework for extracting multimodal collections as data, using Wikidata as the primary source, along with a selection of research scenarios for reusing this data in line with emergent trends in data publication and reuse. Results showed that current trends in data preservation can be achieved through open-source code, cloud services, and collaborative platforms. Future work to be explored includes adopting best practices for provenance documentation and refining reuse scenarios.

Keywords: collections as data; Wikidata; culture heritage

1. Introduction

Preserving digital cultural heritage ensures that cultural and historical information remains accessible, authentic, and usable for future generations. Without active preservation, these materials can become inaccessible over time due to technological change, file corruption, or obsolete software [1]. Simultaneously, there are also increasingly new technologies to ensure computational access and responsible use of collections. Thus, Galleries, Libraries, Archives, and Museums (GLAMs), as a main provider of cultural heritage data, are presented with challenges of both ensuring long-term preservation with shifting technological opportunities and policy instruments for making their digital collections available.

Open Science practices have been defined and implemented to promote open access to scientific publications, research data, models, algorithm collects, software, protocols, notebooks, or workflows [2]. In parallel, emergent data sharing initiatives focused on federation, security, and sovereignty have emerged during the last years such as Solid and data spaces [3,4]. Community efforts such as GLAM Labs¹; AI4LAM²; and international initiatives such as Collections as Data³, FAIR [5], and CARE [6] provide guides for best practices for the publication of machine-actionable collections that support computational access and ethical uses of these data [7,8].



Academic Editor: George Pavlidis

Received: 23 June 2026

Revised: 17 July 2026

Accepted: 23 July 2026

Published: 27 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Increasingly, institutions have turned to the Semantic Web to describe their digital collections and catalogues [9], with Linked Open Data used as the gold standard for enriching and maximising the reuse of cultural heritage data. For example, the National Library of Scotland, the Smithsonian American Art Museum or the Rijksmuseum (Amsterdam), published their data using Application Programming Interfaces (APIs) or as downloadable files [10]. They identified existing sources covering different domains with which to enrich their digital collections. Some examples include GeoNames, Virtual International Authority File (VIAF), and Wikidata, a collaborative edition platform owned by Wikimedia Foundation that has played a leading role in the GLAM sector [11–13]. Thus, we see Linked Data ontologies and knowledge graph infrastructures in GLAMs are used not only to enrich the information on their collections but also in making this information publicly accessible as linked open data in a way to efficiently integrate and implement an unprecedented amount of often unstructured, siloed data at lightning speed.

Despite these affordances, the implementation of linked data technologies in GLAMs remain rare. This is due to a number of issues, largely that the technology and approaches are not ones that traditionally exist in these institutions nor are taught to collection curators and managers. Despite using catalogue systems with standard ontologies which aid to organise, store, and manage collections, the integration of these data with the Semantic Web demands a shift in conceptual data models and data formats. A number of models have been developed for bibliographic records—the IFLA-Library Reference Model (IFLA-LRM), the Resource Description and Access (RDA), and Bibliographic Framework (BIBFRAME)—as well as for cultural heritage objects—the Europeana Data Model (EDM). These models are distinct. Consequently, the question remains: how can these knowledge bases that use Semantic Web principles to collect structured data be employed for preservation purposes?

To investigate this, we take Wikidata as an example and show that this can be extracted and reused in a reproducible manner. We also show the potential (re)uses of the extracted data. The two main contributions of this work are (i) a reproducible framework for extracting multimodal collections as data using Wikidata as a primary source and (ii) a selection of research scenarios to reuse this data according to emergent trends in data publication and reuse. With that, this work is relevant to GLAM practitioners, curators, and researchers who wish to reuse and enrich digital collections made available by GLAM institutions.

This article is organised as follows: after a description of the state of the art, we present the framework to extract multimodal collections as data from Wikidata and its application to a selection of museums. The article concludes with the main conclusions and final remarks.

2. Related Work: The Semantic Web and GLAM

In 2006, Tim Berners-Lee, the inventor of the Web, wrote a memo on the Semantic Web, which outlined “a common framework that allows data to be shared and reused across application, enterprise, and community boundaries”, of which LOD served as a technique to describe knowledge [14]. Thus, in contrast to referencing an unstructured description of a place, person, or object, for example, a textual description in book, linked data through standards such as the Resource Description Framework (RDF) provides a standardised structure to organise, store, and link information on these entities. For example, a historical statement referencing the Belgian author of the well-known comic book series Tintin could be stated as such “Georges Prosper Remi wrote The Adventures of Tintin” and expressed as a triplet consisting of a subject (“:Georges Prosper Remi”), a predicate (“:wrote”), and an object (“:The Adventures of Tintin”). Each of these items are represented with unique identifiers (Uniform Resource Identifiers—URIs) that machines are able to read and retrieve.

Crucially, such a model permits linkage to external repositories, thereby enriching metadata beyond the holding institution's own records. The product of applying these principles to a digital collection is a knowledge graph, the construction of which hinges on a number of interdependent concerns—data modelling, enrichment, and quality foremost among them.

GLAM institutions hold collections of considerable thematic and material breadth, integrating text, images, maps, metadata, postcards, and manuscripts [15].⁴ To render this heterogeneity usable, they have adopted a variety of publication strategies, ranging from international standards for serving images as the Library of Congress⁵ and Europeana.⁶ A growing body of work has, in turn, taken stock of the state of {collections as data} across several national contexts, formulating best practices and guidelines for the publication of machine-actionable collections [16].

Museums, in particular, have been conspicuously active in publishing, enriching and aggregating their catalogues. At one end stands large national databases such as the French Joconde, which documents over a million works held in French art collections [17,18]. Elsewhere, the emphasis falls on machine-actionable release: the Rijksmuseum in Amsterdam offers its catalogue as a data dump expressed in established vocabularies, notably CIDOC-CRM,⁷ and comparable metadata datasets have been derived from the Prado [19]. A further strand publishes catalogue metadata as Linked Open Data (LOD), as the Smithsonian American Art Museum and the Getty Provenance Index⁸ have done an approach whose value is far from merely technical, since metadata covering dealer stock books and sales catalogues can support the identification of art looted during the Second World War. Underpinning much of this activity are shared ontologies: CIDOC-CRM,⁹ widely adopted for describing the physical nature of cultural heritage, and community initiatives such as Linked Art, which promote agreed data models for the domain [20].¹⁰ Running alongside these efforts, experiments with multimodal datasets—combining text, images, and metadata—have sought to advance knowledge discovery and model development, though not without raising challenges in storage and data modelling [20,21].

Once published, such data lends itself to reuse along several axes. One line of enquiry has applied automated methods to assess the quality of the LOD repositories released by cultural heritage (CH) institutions [22]. More recently, the coupling of large language models with external information retrieval—in the form of Retrieval-Augmented Generation (RAG) systems—has been deployed to widen access to archival materials [23]. None of this proceeds without adequate description, and detailed documentation has accordingly become indispensable, both to encourage reuse and to render the structure and content of CH datasets intelligible [24].

Computational notebooks, and Jupyter Notebooks in particular, have meanwhile acquired a notable place in the GLAM toolkit, furnishing a varied repertoire of reproducible and inventive examples [25,26]. Their uptake, however, remains uneven. Adoption within (Western) GLAM and across data and research infrastructures remains emergent, marked by considerable disparities in technical capacity, workflow design, and integration with cloud services [27].

Wikidata

Wikimedia projects serve a range of functions in the GLAM context, from linking and enriching catalogue records to publishing multimedia content [11]. Many institutions have created properties in Wikidata to connect their records to external sources [28]. Listing 1 shows a SPARQL query for retrieving Wikidata properties used to link to these institutions. Through its connection to Wikimedia Commons, Wikidata can also surface the multimedia content held there, such as photographs of artworks in museum collections.¹¹ Wikidata

properties such as wdt:P18 and wdt:P6802, accordingly, serve to record the links to these images.¹² Underpinning these capabilities is Wikibase, the software on which Wikidata runs and which institutions increasingly deploy in their own right to produce machine-readable, FAIR data through cloud services [5,29].

Listing 1. SPARQL query to retrieve Wikidata properties to link museums.

```
SELECT DISTINCT ?museum ?property
WHERE {
  ?museum wdt:P31/wdt:P279* wd:Q33506 .
  ?museum wdt:P1687 ?property
}
LIMIT 500
```

Wikidata's adoption in the GLAM sector has been examined from several angles, and the recurring verdict is favourable. The infrastructure sustained by the Wikimedia Foundation and maintained by an active community has emerged as a viable option for cultural heritage institutions [30,31], not least because it addresses needs that are otherwise costly to meet: within the Digital Humanities, it serves both to recover missing identifiers across all categories of named entity and as a source for controlled-vocabulary integration [32], while affording a transparent, collaborative environment in which cultural heritage data can be described and linked [33]. The balance is not wholly one-sided—the benefits and the attendant challenges of working in the sector have both been documented [34]—yet the barrier to entry is falling, as recent cloud services make Wikibase more accessible to newcomers in the DH and GLAM communities [29].

The body of work surveyed above is, taken together, substantial, but it leaves one issue unaddressed. Existing efforts have, for the most part, reused established datasets: whether to generate visualisations or to evaluate (new) AI-based services—and where frameworks for data extraction have been proposed, complete with reproducible code, they have not taken the collection as their principal object. What is missing, in short, is an approach that brings these three concerns together at once: the GLAM institution as primary content, a governing purpose, and using reproducible code as the means.

3. Methodology

The proposed methodology is based on a framework that works in four steps:

1. Identification of sources;
2. Data extraction;
3. Publication;
4. Reuse.

Figure 1 shows the architecture employed for a tool we named PLOW—Preservation through Linked Open data and Wikidata. The individual steps are described in detail below.

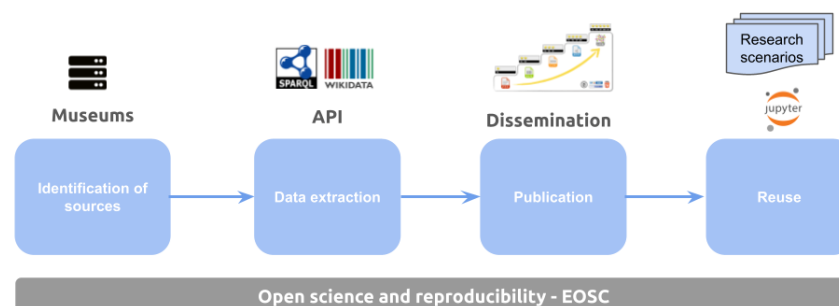


Figure 1. Architecture defined for PLOW—Preservation through Linked Open data and Wikidata.

3.1. Identification of Sources

The first step is the identification of a data source, for which several technical options exist. GLAM datasets are commonly published through GLAM Labs—the Data Foundry at the National Library of Scotland being a representative case—and through the dedicated data sections of major libraries, such as the API et jeux de données of the National Library of France or the equivalent position at the Rijksmuseum in Amsterdam.¹³ These rely chiefly on data dumps that may be downloaded and reused with little friction; others, by contrast, expose their holdings through APIs—SPARQL, OAI-PMH and IIIF among them. Cross-domain hubs occupy a third position: Wikidata, in particular, has become a significant identifier hub in the GLAM sector, enabling the integration of metadata with multimedia content.

Identifying a source, however, is not the same as selecting one, and several criteria bear on that choice. Foremost among them is licensing: while many datasets are released under open licences such as Creative Commons, a substantial number remain copyrighted, or carry no clear statement of reuse at all [16]. Considerations of a different order also intrude: among them, the protection of cultural heritage placed at risk by war, as in Ukraine, Gaza, and Afghanistan.¹⁴ To these may be added the familiar measures of fitness for purpose: data quality, completeness, timeliness, and availability, together with thematic relevance, the degree, to which a source speaks to the subject or research question at hand. Such relevance is best made concrete: a study of artists in exile during the Spanish Civil War, of the paintings of a particular movement, or of works looted by the Nazis during the Second World War will each privilege very different collections [35,36].

3.2. Data Extraction

This step involves extracting the data from the source identified. The output may take a range of serialisation formats, JSON, XML, or RDF, among the most common, and frequently extends beyond textual metadata to multimedia content such as images, maps, or OCR-derived text. Extraction need not rely solely on the originating institution alone: aggregators such as Europeana or collaboratively edited platforms such as Wikidata can serve equally well as points of access.

In some cases, this step can be related to a particular topic, literary movement or event (e.g., World War II). In other cases, the data is extracted based on membership relations such as being part of a particular catalogue (e.g., Rijksmuseum Amsterdam or National Library of France). In this context, Wikidata has created many properties for GLAM institutions in order to link their catalogues [37]. For example, Listing 2 shows a SPARQL query to retrieve artworks from Wikidata linked to the Prado Museum by means of the dedicated property `wdt:P8905`. Other examples rely on using the `CONSTRUCT` clause in SPARQL to build a knowledge graph from the triples retrieved by a query.

Listing 2. SPARQL query to retrieve Art paintings from Wikidata linked to the Prado Museum.

```
SELECT DISTINCT ?s ?sLabel ?author
          ?authorLabel (SAMPLE(?date) AS ?year)
WHERE { ?s wdt:P8905 ?prado .
        ?s wdt:P18 ?image .
        ?s wdt:P180 ?depicts .
        ?s wdt:P571 ?date .
        ?s wdt:P170 ?author
        SERVICE wikibase:label { bd:serviceParam wikibase:language
                                "[AUTO_LANGUAGE],mul,en". }
}
```

GROUP BY ?s ?sLabel ?author ?authorLabel

Once extracted, metadata, and images alike may also surface in aggregators such as Europeana, where they can be integrated and browsed alongside further digital collections.¹⁵

3.3. Publication

The final step concerns the publication of the dataset and its subsequent reuse. Established guidance exists for the former: best-practice frameworks set out the principal steps for publishing machine-actionable collections, within Europeana and beyond [16].

A useful organising principle here is Berners-Lee's five-star deployment scheme (see Figure 2), which ranks open data by degree of accessibility: from material merely placed online under an open licence, through structured and non-proprietary formats, to the use of URIs and, at the highest level, linking to other data to provide context. The scheme is valuable precisely because it reframes publication not as a binary of open versus closed but as a gradient of reusability, against which an institution may locate its own holdings and chart a path of improvement [38].

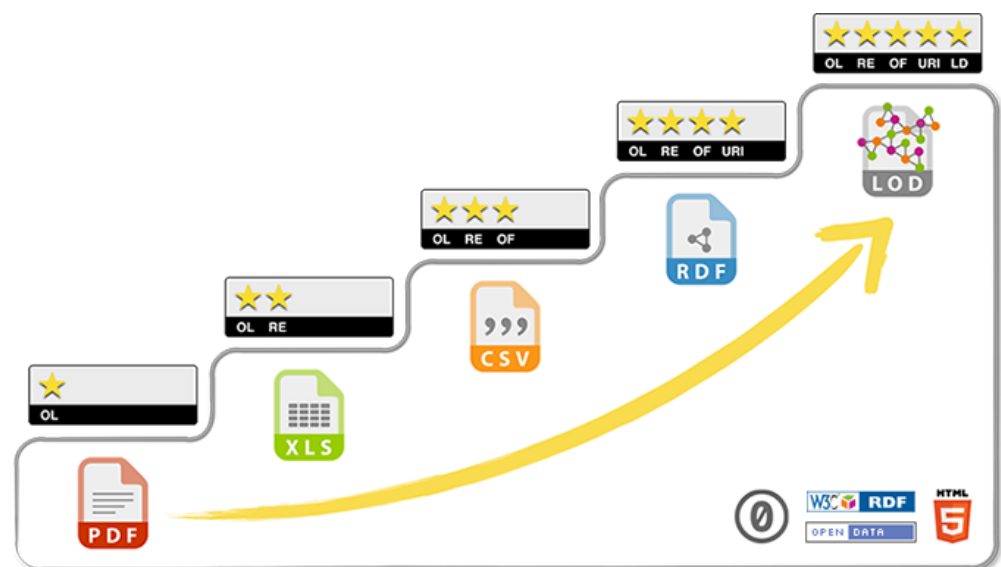


Figure 2. Five-star Open Data, in which each star describes a higher level of openness and usability for data. <https://5stardata.info/en/> (accessed on 25 May 2026).

3.4. Reuse

Reuse, for its part, has been approached from several directions, each building on the content once published. Among the practical instruments, the Jupyter Notebook has gained particular traction, allowing institutions to furnish prototype examples that combine code, charts, documentation, and interactive widgets in a single reproducible artefact [27]. Institutions have begun to employ Jupyter Notebooks as a powerful tool to provide examples of use through prototypes. They combine code, charts, documentation, and interactive widgets [26,39]. Further examples and prospective research scenarios drawn from GLAM data are set out in Section 2.

Of particular interest is the integration of published collections with data and research infrastructures. The Common European Data Space for Cultural Heritage and the European Collaborative Cloud for Cultural Heritage (ECCCH) both operate at this level, as does the European Open Science Cloud (EOSC)—whose Jupyter Notebook service, notably, permits reproducible code to be run on dedicated cloud servers. Visibility and reuse are, in turn, served by platforms such as the Social Sciences and Humanities Open Marketplace, Zenodo, and OpenAIRE.

One further consideration deserves emphasis: versioning. Institutions frequently wish to publish their catalogues as a series of updates rather than a single release, and platforms

such as Zenodo accommodate this directly, issuing a distinct DOI and a suggested citation for each version [40].

4. Results

In this section, we describe the application of the framework in Section 3 to several GLAM institutions. We selected Wikidata as the main source since it has been extensively used in GLAM institutions and specifically for museums as there is a large share of information on paintings (compared to other sources of books, periodicals, or archival material) providing a large sample for experimenting with the workflow and quality testing. The (re)use of these data are also clear given that all data on Wikidata are available under a CC0 licence. Furthermore, Wikidata provides an API to extract these data, providing a transparent and technical workflow for extracting the relevant data. The museums were selected according to the following criteria:

- 1. Availability of a dedicated Wikidata property to link artworks;
- 2. Historical importance, cultural influence, as well as digital transformation;
- 3. Provision of documentation of the catalogue.

This resulted in the following cases: Musée d’Orsay, National Gallery, National Gallery of Ireland, Prado Museum, and the Royal Museum of Fine Arts Antwerp.

Wikidata was used as the primary repository for extracting multimodal content, including metadata, and images. To extract the metadata, the SPARQL query included in Appendix A was employed. The data was retrieved in May 2026. Note that images are stored on Wikimedia Commons, where Wikidata only preserves a link. Table 1 show a summary of the results following the application of the framework. The data were retrieved in July 2026. Duplicates were identified by checking for records containing more than one external identifier associated with the museum under analysis.¹⁶

A GitHub repository was created including the extracted data, a README file with additional documentation and a collection of Jupyter Notebooks illustrating how to extract and reuse the data.¹⁷ Several Python packages, such as SPARQLWrapper (version 2.0.0), were employed to query the Wikidata SPARQL endpoint.¹⁸ The data was extracted as a JSON file. The Jupyter notebooks were hosted and executed using the services provided by EOSC. To do so, a dedicated virtual machine was deployed to run the code provided. This work intends to promote its use amongst the community. Best practices were used to develop and publish the collection of Jupyter Notebooks [25,39].

Table 1. Summary of the data retrieved from Wikidata. The Wikidata property column includes the property used to extract the data. Note that all the records provide at least one image.

Institution	Wikidata Property	No. Records	No. Duplicates	Years
Musée d’Orsay	wdt:P4659	1923	3	1730–2000
National Gallery	wdt:P13325	2433	1	1249–2000
National Gallery of Ireland	wdt:P8906	1046	0	1310–2004
Prado Museum	wdt:P8905	3916	3	100–2000
Royal Museum of Fine Arts Antwerp	wdt:P4905	2045	6	1300–2000

4.1. Scenarios of Reuse

This section outlines a selection of scenarios in which the extracted data might be reused, each grounded in prior work spanning topics ranging from the integration of data repositories to the application of various AI-based techniques [27,30]. The scenarios are addressed to a mixed readership of practitioners, curators, researchers, and students, and their purpose is illustrative rather than prescriptive: they show how data may be extracted and preserved in keeping with the best practices promoted by international initiatives,

without pretending to have a full specification or a worked implementation. What they offer, in short, is a narrative account of added value rather than a manual.

Federated queries. Institutions tend to publish digital collections as isolated collections, effectively making a data silo and thus hindering collaboration and their reuse. However, advances in technology have created a new context in which methods such as Named Entity Recognition (NER) and Entity Linking are helping enrich digital collections. When linking records to external repositories, new opportunities arise to enrich metadata, improve interoperability, and enable broader access to cultural and research collections. Wikidata federates several SPARQL repositories, enabling access to different repositories from relevant institutions through Wikidata [41].

For example, Listing 3 shows a federated SPARQL query to retrieve bibliographic records from Biblioteca Virtual Miguel de Cervantes (BVMC) in which the painter Diego Velázquez is represented as a subject. Note that the SERVICE command is employed to access the BVMC SPARQL endpoint from Wikidata. The BIND command is employed to filter the bibliographic records using the Picasso's identifier in BVMC. Other examples could be based on the integration of painting from catalogues such as the Prado Museum and the National Gallery, focused, for instance, on a particular painter such as Diego Velázquez.

Listing 3. SPARQL query to retrieve bibliographic records from Biblioteca Virtual Miguel de Cervantes in which the painter Diego Velázquez is represented as a subject.

```
SELECT DISTINCT ?work ?workLabel
WHERE {
    ?s wdt:P18 ?image .
    ?s wdt:P170 wd:Q5593 .
    BIND(<https://data.cervantesvirtual.com/person/7331>
    as ?bvmcID)
    SERVICE
    <http://data.cervantesvirtual.com/openrdf-sesame/repositories/data>
    {
        ?work dc:subject ?bvmcID .
        ?work rdfs:label ?workLabel .}}}
```

According to these observations, museum data could be enriched with data from libraries and archives, such as the National Library of France, to create integrated research scenarios that combine different types of records. This could help address challenges related to data quality, interoperability, and reuse.

AI-based computer vision models. Art works in Wikidata are described using many properties (e.g., title and inception). However, given Wikidata's collaborative editing approach, properties are not consistently used across all entities. The property wdt:P180 is used to indicate what an item depicts, in particular, to include entities visually depicted in an image (e.g., Spanish Civil War).¹⁹ Computer vision models can be employed to enhance the value of this property by leveraging the multimedia content generated by the data extraction process, thereby enabling the refinement and enhancement of image descriptions. Metadata can be automatically generated to describe images provided by GLAM institutions in greater detail, such as paintings, maps, or postcards [27].

See, for instance, Figure 3 representing *Las Meninas* from Diego Velázquez. The Wikidata identifier wd:Q208758 provides metadata about the painting, including the property wdt:P180 with several values such as people included in the painting, the painter himself or a dog.



Figure 3. Las Meninas. Source: Wikimedia Commons.

Federated data storage systems. Solid is an emerging initiative for storing federated data while ensuring security and sovereignty [3]. Previous research has analysed the use of Solid in the GLAM context, combined with Wikidata as a federated storage system.²⁰ Software libraries and tutorials in different programming languages, such as Java and Python, have been developed to facilitate their adoption.²¹ These new technologies facilitate data portability across applications while enhancing transparency and privacy in data access.

RAG-based search engine. Wikidata may be interrogated by several means—keyword matching or SPARQL, among them—yet each has its limits: a keyword set may fail to capture the intent behind a query, and SPARQL’s steep learning curve places it beyond many users. Wikimedia has accordingly begun exploring semantic search techniques, including vector embeddings, as a more forgiving mode of access.²² Coupled with large language models and natural language processing, such methods can materially improve retrieval [23], and earlier work has already applied Retrieval-Augmented Generation to web-archive content and digitised collections, thereby opening those holdings to conversational access.²³ Hence, the artwork descriptions held in Wikidata records might themselves underwrite a RAG search engine—one that affords new routes into the content, curbs hallucination, and injects domain-specific knowledge into the underlying model.

LLM or SLM-based chatbot. A natural specialisation of the retrieval-augmented scenario sketched above is a conversational agent addressed not to the researcher but to the museum visitor: a chatbot, in effect, with which one might interrogate a single art object one is standing before. The premise is straightforward: the multimodal records extracted by the framework—title, creator (wdt:P170), date of inception (wdt:P571), the entities recorded under wdt:P180, and the associated image wdt:P18—furnish a bounded, machine-readable context against which a visitor’s questions may be answered. Rather than draw on its own parametric memory, with all the attendant risk of confabulation, the model retrieves the relevant record and renders it as natural-language prose, and

where the visitor asks what a painting depicts, the very property whose enrichment was discussed above supplies the answer. The choice between a large language model (LLM) and a small language model (SLM) is, in this setting, less a technical detail than a design decision with institutional consequences. An LLM offers fluency, breadth, and a capacity to field unanticipated questions, but it is costly to serve continuously to the public and, when cloud-hosted, returns us to the difficulty noted above of data leaving European jurisdiction. An SLM, a model of a few billion parameters capable of running on modest, or even on-premise hardware, inverts these trade-offs. Its narrower internal knowledge is, paradoxically, an advantage rather than a constraint: confined to answering from the retrieved record alone, it is more readily held to the institution's own authoritative data, and less inclined to volunteer plausible invention. Hence, for a bounded domain such as the holdings of a single museum, pairing a modest SLM with strong retrieval grounding may prove not merely the cheaper option but also the more trustworthy one [42].

Wikidata records carry labels in many languages; the mechanism is already exploited by the SERVICE clause in Listing 2—a visitor might both pose questions and receive answers in their own tongue, without additional translation infrastructure. In order to foster transparency, each response may be accompanied by a reference to the Wikidata item from which it was drawn, so that the curious or the sceptical can verify the claim at its source. It should be stressed, however, that such a system inherits the limitations of its underlying data. Where Wikidata's coverage of the Prado is partial or where `wdt:P180` has been applied unevenly, the agent will be correspondingly silent or incomplete, and no degree of retrieval grounding wholly forecloses error. A visitor-facing use would therefore demand careful evaluation against curatorial ground truth before it could responsibly stand between the public and the collection.

These research scenarios provide a wide range of approaches for reusing the content extracted by the framework proposed in this work.

4.2. Discussion

This work complements existing approaches to the extraction, publication, and reuse of digital collections, but it occupies a position none of them does: it takes digital preservation as its governing concern, a collaborative platform—Wikidata—as its means, and the museum collection as its principal content. Its contribution, then, is to enrich the current landscape of worked examples, supplying the community with additional cases that follow established best practices. In this work, we showed how such an approach can be applied to works from museums; however, it can be extended to other domains and contexts by updating the data extraction step, and in particular for this approach, the specific SPARQL query.

In some cases, particularly when querying a large number of records, the Wikidata API may return a timeout error. However, Wikidata provides SPARQL mechanisms, such as limiting the number of returned results, to help avoid these issues. Another possible approach is to install a local RDF triple store to execute the queries locally. An additional topic to consider is data quality, as Wikidata is a collaboratively edited platform. Although errors such as duplicate records or incomplete metadata may occasionally occur, Wikidata is continuously monitored and curated by its community. In general, the quality of its data is high, particularly for well-established entities, and the platform incorporates numerous quality control mechanisms to detect and correct inconsistencies over time.

This framework contributes to museum practice by integrating key museum processes such as documentation, collection management, research, interoperability, and long-term preservation. These topics are essential to promote the reuse of the digital collections as well as addressing issues related to transparency and provenance [24].

Several limitations bound the scope of what has been attempted here. Wikidata does not hold a complete representation of the Prado's catalogue; a substantial selection of records has nonetheless been linked to it, thanks to the collaborative efforts of the community. It would be valuable, even so, for the institution to offer access by additional means—OAI-PMH, IIIF, or SPARQL, each weighing richness, intelligibility, and complexity differently—and, looking further ahead, through sovereignty-preserving approaches such as data spaces. Using Wikidata entails several challenges that must be addressed. For example, data uploaded to Wikidata is highly reused by AI companies to create commercial products, while GLAM institutions receive no compensation. In exchange, the visibility of the digital collections increases, impacting society. In some cases, the data modelling provided by Wikidata might not be suitable to represent indigenous knowledge. A more inclusive approach is required in these cases in order to facilitate an equitable and non-discriminatory environment.

The EOSC cloud services are constrained in memory and disk storage; for the examples presented here, those resources sufficed, but more demanding AI-based work would require provision with greater GPU and memory capacity. Alternatives such as Google Colab supply additional resources, though at the cost of depositing data outside Europe—a consideration of some weight for European institutions and researchers.

The reuse scenarios are illustrative and their implementation and formal evaluation constitute future work. They have been presented as textual sketches: prototypes in narrative form rather than working systems, and their implementation remains a task for further work; refinement may in any case prove necessary, since several of the data and research infrastructures invoked are themselves still under development.

This work can be extended in several directions. The SPARQL queries might be refined to draw in additional classes of institutions, libraries, and archives, and data research infrastructures, such as those discussed above, could be pursued in the service of preservation and reuse alike.

5. Conclusions

GLAM institutions have, in recent years, been exploring new ways of making their digital collections available, and the preservation of digital heritage has become a matter of real public consequence—a means of safeguarding cultural heritage for the future. A range of initiatives have emerged to promote the publication and responsible reuse of the data these institutions hold. In this research, we showed how Wikidata can be used in these tasks.

We proposed a reproducible framework for extracting multimodal collections as data using Wikidata as a primary source. In keeping with Open Science principles, this work has described a reproducible approach for extracting multimodal collections as data from museums via Wikidata. The framework was applied to a selection of museums, with Wikidata serving as the principal repository, and was accompanied by narrative research scenarios demonstrating how the extracted datasets might be reused. Taken together, this demonstrates how data preservation and extraction can be met through open-source code, cloud, services and collaborative platforms—without recourse to bespoke or proprietary infrastructure.

Two directions invite further work. The first is the adoption of best practices to record provenance in a machine-readable form [40]. The second is the refinement of the reuse scenarios themselves—supplying the structured, machine-actionable detail that would allow others to take them up in earnest. Of the two, the latter is likely to do the most to encourage adoption.

Author Contributions: Conceptualisation: G.C.; Formal Analysis: all; Investigation: all; Methodology: G.C.; Visualisations: G.C.; Resources: all; Software: G.C.; Writing original draft: G.C.; Writing—review and editing: all. All authors have read and agreed to the published version of the manuscript.

Funding: C. Annemieke Romein’s research within the HAICu-project (digital Humanities Artificial Intelligence and Cultural Heritage) is funded by the NWO NWA 1518.22.10 grant. The contributions of Julie M. Birkholz are supported by the FWO-funded Large-Scale Infrastructure project CLARIAH-VL+ (I001525N) and BELSPO FED-tWIN project (Prf-2019-040).

Data Availability Statement: The data supporting the findings of this study are publicly available at Zenodo (<https://doi.org/10.5281/zenodo.21552395>).

Acknowledgments: The authors thank the International GLAM Labs Community, the Impact Centre of Competence, Wikimedia Spain, and DARIAH-EU.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
BVMC	Biblioteca Virtual Miguel de Cervantes
CARE	Collective Benefit, Authority to Control, Responsibility, Ethics
CH	Cultural Heritage
DOI	Digital Object Identifier
ECCCH	European Cultural Heritage Cloud
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
GLAM	Galleries, Libraries, Archives, and Museums
GPU	Graphics Processing Unit
IIIF	International Image Interoperability Framework
LLM	Large Language Model
LOD	Linked Open Data
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
RAG	Retrieval-Augmented Generation
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
VIAF	Virtual International Authority File

Appendix A

Listing A1 was employed to extract the metadata from the Wikidata public SPARQL endpoint for the Prado Museum.

Listing A1. SPARQL query to retrieve Art paintings from Wikidata linked to the Prado museum.

```
SELECT DISTINCT ?s ?sLabel
(SAMPLE(? date) AS ?year)
(GROUP_CONCAT(DISTINCT ?author; SEPARATOR="; ") AS ?authorAll)
(GROUP_CONCAT(DISTINCT ?depicts; SEPARATOR="; ") AS ?depictsAll)
(GROUP_CONCAT(DISTINCT ?depictsTxt; SEPARATOR="; ")
AS ?depictLabels)
WHERE {?s wdt:P8905 ?prado .
?s wdt:P18 ?image .
OPTIONAL {?s wdt:P180 ?depicts .
?depicts rdfs:label ?depictsTxt .
FILTER (LANG(?depictsTxt) = "en")}}
```

```

?s wdt:P571 ?date .
?s wdt:P170 ?author
SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],mul,en". }
}
GROUP BY ?s ?sLabel ?author ?authorLabel ?year

```

Notes

- ¹ <https://glamlabs.io/> (accessed on 27 May 2026).
- ² <https://ai4lam.org/> (accessed on 27 May 2026).
- ³ <https://collectionsasdata.github.io/> (accessed on 27 May 2026).
- ⁴ See, for example, <https://www.glamlabs.io/projects> (accessed on 27 May 2026).
- ⁵ <https://guides.loc.gov/chronicling-america> (accessed on 21 May 2026).
- ⁶ <https://pro.europeana.eu/page/datasets> (accessed on 17 May 2026).
- ⁷ <https://data.rijksmuseum.nl/docs/data-dumps> (accessed on 27 May 2026).
- ⁸ <https://www.getty.edu/research/provenance/> (accessed on 27 May 2026).
- ⁹ <https://cidoc-crm.org/> (accessed on 27 May 2026).
- ¹⁰ <https://linked.art/> (accessed on 27 May 2026).
- ¹¹ See, for example, https://commons.wikimedia.org/wiki/Museo_del_Prado (accessed on 27 May 2026).
- ¹² See, for example, Picasso's painting Guernica in Wikidata available at <https://www.wikidata.org/wiki/Q175036> (accessed on 27 May 2026).
- ¹³ See, for example, <https://api.bnf.fr/> (accessed on 27 May 2026).
- ¹⁴ See, for example, the AISTER project available at <https://aister.uni.lu/> (accessed on 27 May 2026).
- ¹⁵ <https://www.europeana.eu/en/collections/organisation/2658-the-prado-museum> (accessed on 13 May 2026).
- ¹⁶ See, for instance, <https://www.wikidata.org/wiki/Q5834840> (accessed on 27 May 2026).
- ¹⁷ <https://github.com/hibernator11/preserving-digital-heritage> (accessed on 27 May 2026).
- ¹⁸ <https://sparqlwrapper.readthedocs.io> (accessed on 27 May 2026).
- ¹⁹ See, for example, the Guernica entity in Wikidata available at <https://www.wikidata.org/wiki/Q175036> (accessed on 27 May 2026).
- ²⁰ <https://github.com/hibernator11/solid-glam> (accessed on 27 May 2026).
- ²¹ https://solidproject.org/for_developers (accessed on 27 May 2026).
- ²² https://www.wikidata.org/wiki/Wikidata:Embedding_Project (accessed on 27 May 2026).
- ²³ <https://github.com/iipc/WACAD/tree/main/notebooks> (accessed on 27 May 2026).

References

1. Terras, M. Opening access to collections: The making and using of open digitised cultural content. *Online Inf. Rev.* **2015**, *39*, 733–752. <https://doi.org/10.1108/OIR-06-2015-0193>.
2. European Commission. Open Science. 2026. Available online: https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science_en (accessed on 15 May 2026).
3. Mansour, E.; Sambra, A.V.; Hawke, S.; Zereba, M.; Capadisli, S.; Ghanem, A.; Aboulnaga, A.; Berners-Lee, T. A Demonstration of the Solid Platform for Social Web Applications. In *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion, Montreal, QC, Canada, 11–15 April 2016*; CHE: Republic and Canton of Geneva, Switzerland; pp. 223–226. <https://doi.org/10.1145/2872518.2890529>.
4. Dobрева, M.; Stefanov, K.; Ivanova, K. Data Spaces for Cultural Heritage: Insights from GLAM Innovation Labs. In *Proceedings of the From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries—24th International Conference on Asian Digital Libraries, ICADL 2022, Hanoi, Vietnam, 30 November–2 December 2022*; Tseng, Y., Katsurai, M., Nguyen, H.N., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13636, pp. 492–500. https://doi.org/10.1007/978-3-031-21756-2_41.
5. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. <https://doi.org/10.1038/sdata.2016.18>.

6. Carroll, S.R.; Garba, I.; Figueroa-Rodríguez, O.L.; Holbrook, J.; Lovett, R.; Materechera, S.; Parsons, M.A.; Raseroka, K.; Rodriguez-Lonebear, D.; Rowe, R.; et al. The CARE Principles for Indigenous Data Governance. *Data Sci. J.* **2020**, *19*, 43. <https://doi.org/10.5334/DSJ-2020-043>.
7. Mahey, M.; Al-Abdulla, A.; Ames, S.; Bray, P.; Candela, G.; Derven, C.; Dobрева-McPherson, M.; Gasser, K.; Chambers, S.; Karner, S.; et al. *Open a GLAM Lab*; International GLAM Labs Community, Book Sprint: Doha, Qatar, 2019; p. 164. <https://doi.org/10.21428/16ac48ec.f54af6ae>.
8. Padilla, T.; Allen, L.; Frost, H.; Potvin, S.; Russey Roke, E.; Varner, S. *Final Report—Always Already Computational: Collections as Data*, version 1; Zenodo: Geneva, Switzerland, 2019. <https://doi.org/10.5281/zenodo.3152935>.
9. Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web in Scientific American. *Sci. Am. Mag.* **2001**, *284*. Available online: <https://www.scientificamerican.com/article/the-semantic-web/> (accessed on 17 July 2026).
10. Dijkshoorn, C.; Jongma, L.; Aroyo, L.; van Ossenbruggen, J.; Schreiber, G.; ter Weele, W.; Wielemaker, J. The Rijksmuseum collection as Linked Data. *Semant. Web* **2018**, *9*, 221–230. <https://doi.org/10.3233/SW-170257>.
11. Zhao, F. A systematic review of Wikidata in Digital Humanities projects. *Digit. Scholarsh. Humanit.* **2022**, *38*, 852–874. <https://doi.org/10.1093/llc/fqac083>.
12. Freire, N.; Isaac, A. Wikidata's Linked Data for Cultural Heritage Digital Resources: An Evaluation Based on the Europeana Data Model. *Int. Conf. Dublin Core Metadata Appl.* **2020**, *2020*, 59–68. <https://doi.org/10.23106/dcmi.952141718>.
13. Candela, G.; Cupér, M.; Holownia, O.; Gabriëls, N.; Dobрева, M.; Mahey, M. A Systematic Review of Wikidata in GLAM Institutions: A Labs Approach. In *Proceedings of the Linking Theory and Practice of Digital Libraries—28th International Conference on Theory and Practice of Digital Libraries, TPD L 2024, Proceedings, Part II, Ljubljana, Slovenia, 24–27 September 2024*; Antonacopoulos, A., Hinze, A., Piwowarski, B., Coustaty, M., Nunzio, G.M.D., Gelati, F., Vanderschantz, N., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; Volume 15178, pp. 34–50. https://doi.org/10.1007/978-3-031-72440-4_4.
14. Berners-Lee, T. Linked Data. 2006. Available online: <http://www.w3.org/DesignIssues> (accessed on 20 May 2026).
15. Shang, C. Quantitative assessment of the value of intangible cultural heritage art supported by multimodal machine learning. *Discov. Artif. Intell.* **2026**, *6*, 7. <https://doi.org/10.1007/S44163-025-00697-9>.
16. Candela, G.; Gabriëls, N.; Chambers, S.; Dobрева, M.; Ames, S.; Ferriter, M.; Fitzgerald, N.; Harbo, V.; Hofmann, K.; Holownia, O.; et al. A checklist to publish collections as data in GLAM institutions. *Glob. Knowl. Mem. Commun.* **2023**, *74*, 1323–1355. <https://doi.org/10.1108/GKMC-06-2023-0195>.
17. Moissinac, J.C.; Rouzé, F.; Wadhwa, P.; Germain, B. Toward a Semantic Representation of the Joconde Database. In *Proceedings of the Semapro, IARIA, Nice, France, 25–29 October 2020*.
18. Ministère de la Culture. *Collections des Musées de France*; Base Joconde: Paris, France, 2026.
19. Gonzalez Cifuentes, T.; Torrejon Vazquez, D. *Dataset de obras del Museo Nacional del Prado*; Zenodo: Geneva, Switzerland, 2026. <https://doi.org/10.5281/zenodo.19261880>.
20. Sanderson, R. Implementing Linked Art in a Multi-Modal Database for Cross-Collection Discovery. *Open Lib. Humanit.* **2024**, *10*. <https://doi.org/10.16995/olh.15407>.
21. Nguyen, T.; Storås, A.M.; Thambawita, V.; Hicks, S.A.; Halvorsen, P.; Riegler, M.A. Multimedia Datasets: Challenges and Future Possibilities. In *Proceedings of the MultiMedia Modeling, Bergen, Norway, 9–12 January 2023*; Dang-Nguyen, D.T., Gurrin, C., Larson, M., Smeaton, A.F., Rudinac, S., Dao, M.S., Trattner, C., Chen, P., Eds.; Springer: Cham, Switzerland, 2023; pp. 711–717.
22. Presutti, V. Data Quality and Ethics in Cultural Heritage Knowledge Graphs. In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2025, Volume 1: KDIR, Marbella, Spain, 22–24 October 2025*; Coenen, F., Nolle, L., Aveiro, D., Fernández-Breis, J.T., Masciari, E., Gruenwald, L., Bernardino, J., Torres, R., Eds.; SCITEPRESS: Setúbal, Portugal, 2025; p. 9.
23. Kelly, P.; Schild, J.; Jafari, A. FolkRAG: A retrieval-augmented generation system for cultural heritage materials. *Neural Comput. Appl.* **2025**, *37*, 20281–20297. <https://doi.org/10.1007/s00521-025-11455-4>.
24. Alkemade, H.; Claeysens, S.; Colavizza, G.; Freire, N.; Lehmann, J.; Neudecker, C.; Osti, G.; van Strien, D. Datasheets for Digital Cultural Heritage Datasets. *J. Open Humanit. Data* **2023**, *9*, 17. <https://doi.org/10.5334/johd.124>.
25. Pimentel, J.F.; Murta, L.; Braganholo, V.; Freire, J. Understanding and improving the quality and reproducibility of Jupyter notebooks. *Empir. Softw. Eng.* **2021**, *26*, 65. <https://doi.org/10.1007/S10664-021-09961-9>.
26. Sherratt, T. *GLAM Workbench*, version v2.0.0; Zenodo: Geneva, Switzerland, 2025. <https://doi.org/10.5281/zenodo.15597489>.
27. Candela, G.; Dobрева, M.; Alkemade, H.; Holownia, O.; Mahey, M.; Ames, S.; Renaud, K.; Vodopivec, I.; Lee, B.C.G.; Padilla, T.; et al. A Use Case Lens on Digital Cultural Heritage. *J. Comput. Cult. Herit.* **2026**, *accepted*. <https://doi.org/10.1145/3828548>.
28. Dişli, M.; Candela, G.; Gutiérrez, S.; Fontenelle, G. Open Data Practices of Art Museums in Wikidata: A Compliance Assessment. *J. Open Humanit. Data* **2025**, *11*, 71. <https://doi.org/10.5334/johd.438>.
29. Lindemann, D.; Candela, G.; Marchetti, A.; Pellizzari di San Girolamo, C.C.; Olea, I.; Varvantakis, C.; Assis, T.; Moitinho de Almeida, V.; Obregon Sierra, A.; Salgado, A.; et al. *The Wikibase Ecosystem in DH and GLAM*, version v2; Zenodo: Geneva, Switzerland, 2026. <https://doi.org/10.5281/zenodo.21110147>.

30. Thornton, K. *Wikidata for Digital Preservationists*; Technical Report; Digital Preservation Coalition: Glasgow, Scotland, 2021. <https://doi.org/10.7207/twgn21-19>.
31. Thornton, K.; Seals-Nutt, K. In-Person Long Paper. Metadata. 2022. Available online: <https://youtu.be/Xx6-Z4EDxEk> (accessed on 19 May 2026).
32. Trognitz, M.; Mandell, R.A.; Štuhec, S.; Palacz, J. Wikidata as a Reconciliation Anchor: Curating Data for Long-Term Preservation. *J. Open Humanit. Data* **2026**, *12*, 27. <https://doi.org/10.5334/johd.450>.
33. Sichani, A.M.; Kono, K.; Winters, J. Building Cultural Heritage Data Infrastructures with Wikidata: The Case of the Congruence Engine Data Register. *J. Open Humanit. Data* **2026**, *12*, 35. <https://doi.org/10.5334/johd.448>.
34. Smith-Yoshimura, K. Experimentations with Wikidata/Wikibase. Hanging Together Blog. 2020. Available online: <https://hangingtogether.org/experimentations-with-wikidata-wikibase/> (accessed on 19 May 2026).
35. Salinas, C.G. *Las Artistas del Exilio Republicano Español: El Refugio latinoamericano*; Arte Grandes Temes, Ediciones Cátedra: Madrid, Spain, 2019.
36. J. Paul Getty Trust. Getty Provenance Index. Available online: <https://www.getty.edu/research/provenance/> (accessed on 19 May 2026).
37. Romein, C.A.; Wagner, A.; van Zundert, J.J. Building and Deploying a Classification Schema Using Open Standards and Technology. *J. Digit. Leg. Hist.* **2023**, *2*. <https://doi.org/10.21825/dlh.85751>.
38. Romein, C.A.; Kemman, M.; Birkholz, J.M.; Baker, J.; Dd Gruijter, M.; Meroño-Peñuela, A.; Ries, T.; Ros, R.; Scagliola, S. State of the Field: Digital History. *History* **2020**, *105*, 291–312. <https://doi.org/10.1111/1468-229X.12969>.
39. Candela, G.; Chambers, S.; Sherratt, T. An approach to assess the quality of Jupyter projects published by GLAM institutions. *J. Assoc. Inf. Sci. Technol.* **2023**, *74*, 1550–1564. <https://doi.org/10.1002/ASI.24835>.
40. Romein, C.A.; Hodel, T.; Gordijn, F.; van Zundert, J.J.; Chagué, A.; van Lange, M.; Jensen, H.S.; Stauder, A.; Purcell, J.; Terras, M.M.; et al. Exploring Data Provenance in Handwritten Text Recognition Infrastructure. *J. Data Min. Digit. Humanit.* **2024**. <https://doi.org/10.46298/jdmdh.10403>.
41. Dişli, M.; Osti, G.; Candela, G.; Zijdemán, R. From Linked Open Data to Collections as Data: A Reproducible Framework Using Federated Queries. *Inf. Technol. Libr.* **2025**, *44*, 1–15. <https://doi.org/10.5860/ital.v44i4.17432>.
42. Mahadeshwar, R.; Cranenburgh, A.v.; Caselli, T.; Nissim, M. Evaluating the Impact of Source Diversity for RAG in Historical Research. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), Palma, Mallorca, Spain, 11–16 May 2026*; Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., Toral, A., Eds.; ELRA Language Resources Association: Paris, France, 2026; pp. 716–734. <https://doi.org/10.63317/4yz9z3uvzasd>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.