DIGITAL HERITAGE (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

Mediating Art History Data Models for Native Linked Data Construction using ResearchSpace

Alessandro Adamou[1] and Polina Voronova[1]

[1]Bibliotheca Hertziana - Max Planck Institute for Art History

Abstract

Key challenges in digital art history concern translating common domain data models to the ontological framework of CIDOC-CRM and the like. This paper reports on the practical and methodological application in a domain, i.e. ritual spaces in medieval Europe, that is predicated on a very specific local model, and on how to operationalize it using the ResearchSpace platform. We illustrate the key challenges with directly building linked data in this domain for integration with other knowledge graphs in the humanities, yet by translating the complexity of CIDOC-CRM to the nuances of the domain model at runtime. Along with templates and knowledge patterns–the semantic tools made available by ResearchSpace–we further extend the platform's core functions, working around their limitations, to integrate external sources like OpenStreetMap and Zotero, in the Mapping Sacred Spaces project. The resulting workflow supports the generation of interoperable Linked Data from the outset, offering reusable modeling patterns and methodological insights applicable to other projects working with structured art history data.

CCS Concepts

• ***Applied computing*** → ***Fine arts;*** *Architecture (buildings);* • ***Information systems*** → *Crowdsourcing; Resource Description Framework (RDF); Federated databases; Mediators and data integration;*

1. Introduction

In recent years, the adoption of linked data and knowledge graphs has significantly transformed digital practices in the humanities, enabling more interconnected, nuanced, and machine-readable representations of historical and cultural knowledge. Unlike traditional databases, which often impose rigid schemas, semantic frameworks allow for flexible modeling of complex relationships, such as the functions, attributions, and provenances of cultural artifacts.

Much of the implementation effort in data-intensive projects in the humanities, such as digital archives and catalogs, is spent transforming content in databases of various formats, such as relational or XML-based, to linked data. Owing to their predefined, rigorous schemas, those databases have likely been populated using interfaces, like forms and tables, that reflect the research questions originally posed, and to which domain experts have grown accustomed. When the goal, however, is to have domain experts generate linked data natively, rather than as an afterthought, several design issues must be taken into account, such as:

1. possible misalignment between the generality of the standard ontologies employed in the knowledge graph, such as CIDOC-CRM and LRM, and the specificity of the domain requirements;
2. dealing with incompleteness, uncertainty, and approximation;
3. the need to preserve interaction paradigms for entering and displaying information, that are customary to domain experts but must incorporate data reuse, which in linked data is key.

Before the rise of virtual research environments targeting digital humanities, these use cases required custom implementations on top of general-purpose Web platforms. Dedicated systems now offer semantically structured environments for managing and visualizing complex cultural heritage data. This paper offers methodological insights into the curatorial practices for the native generation of linked data in an art historical context.

The case study at hand is *Mapping Sacred Spaces*, an ongoing project aimed at building a census of ritual architectural sites–e.g. churches, chapels or abbeys–in medieval Southern Italy, and of the liturgical furnishings therein. Southern Italy is historically marked by dynamic cultural exchange and as such it has witnessed changes in the locations and functions of many such furnishings, therefore organizing this knowledge semantically is crucial. As the resulting dataset must integrate with an institutional knowledge graph in CIDOC-CRM, which transcends the expertise of traditional art historians, the complexity of the "global" CRM model is masked by the "local" model that conforms to their research questions.

An example, which incorporates the aforementioned three orders of problems, wuld be to model an 11th-Century marble slab that came from one church and is now being used as pulpit fall in Santa Restituta (Naples). Scholars must enter these data using form fields that write CRM predicates, but guaranteeing the reuse of the associated materials, functions and churches from external sources.

The implementation sits upon ResearchSpace (RS), a platform

DOI: 10.2312/dh.20253276

for managing the full lifecycle of semantic data with built-in support for CRM [OT18]. To perform model translation at runtime, we use the RS feature of knowledge patterns (KPs), a collection of parametric queries that govern individual fields. The capabilities for data integration offered by RS are leveraged to guarantee real-time reuse of third party data, from sources like OpenStreetMap, Zotero, the Getty vocabularies, Wikidata, and IconClass, and have been expanded upon to overcome the limitations of the platform.

2. Related Work

Before dedicated platforms were established as the ways to manage the lifecycle of data, early attempts at natively producing linked data for the humanities extended general-purpose content management systems. The *Listening Experience Database*, for instance, extended Drupal with a triplestore and used named graphs to implement multi-user data curation, reusing external datasets by maintaining Solr search indices of them; however, the application has the project data model hardwired in its code and its back-end logic was never generalized to support arbitrary ontologies [ABB*19]. More recent work has gone in that direction, leveraging ontology patterns as a generalization for building art historical data [Car24].

Also built upon Drupal and complementing it with a triplestore is *WissKI*, which encourages semantic annotation as a data input practice: a paradigm that was deemed unsuitable for our use case, which relies on structured rather than textual content. *Omeka-S* allows direct editing of semantic entities using ontologies loaded into it, rather than managing mappings: it has a custom API, whereas SPARQL language support was only recently retrofitted into it. *Arches* is a geospatially-enabled platform focused on documenting immovable heritage, which covers only part of our use cases. Our choice fell upon *ResearchSpace*, which allows page templating based on entity types, exposes a SPARQL API, and has a knowledge pattern system able to map between data models at runtime, though at the sacrifice of fine-grained multi-user data curation.

For a systematic review of the systems above, we refer to a recent survey [NCM19], to which it should be added that, recently, we have also witnessed renewed usage of the Semantic MediaWiki platform for cultural heritage [Kra23], which, like WissKI, follows the paradigm of content annotation as a way of managing entities.

3. Approach

To support diverse research questions, the data entered by the scholars must be shaped so that multi-faceted filters for search and browsing can be applied together. For example, viewing all the mosaic slabs from a certain period of time requires controlled vocabularies for selecting resource types–including slabs– and techniques, and that production years or year ranges can be placed in a total ordering. Therefore, a balance between maximizing expressivity and guaranteeing rigor in data structures must be stricken.

In what follows, key issues will be highlighted with a model of liturgical furnishings as designed by art historians who cannot directly consume CRM. The approach taken is to employ semantic forms and support them with KPs, as predicated by the RS system. Here, a *semantic form* is intended as an input method whose fields can hold literal values or references to entities, which can either be built from that field (or a set thereof), or reused from the project dataset or third-party datasets. A *KP* is a set of parametric SPARQL queries or updates, all assigned to a property in the local model, that indicate how database operations–read, insert, delete, field autocompletion etc.–translate to the global model of CRM.

3.1. Local data model issues

Mapping Sacred Spaces organizes its material on a multi-layered structure that combines part-whole and location-related aspects. The middle level is the one of *objects*, which are installations with an inherently ritual function—such as altars, chancel screens, or baptismal fonts. These can be originally created for, or be preserved in *sites*, which are higher-level artifacts of architectural nature, and are themselves mereologically organized: for example, a monumental complex may contain a church, a cloister, and a museum. Finally, parts of an objects, like slabs, that should be considered atomic from an analytical perspective are bottom-level *components*.

Underneath this multi-level structure a variety of relations are at play, which call for different CRM property paths, beyond those contemplated in e.g. Linked.Art (`https://linked.art/model/`). To support a historical perspective, many topological relationships need to be modeled, which depend on the status of the artifact, as it may since have been decontextualized, stripped of its original function, preserved in uncharted storage spaces, or lost. These relations are summarized in Figure 1. When recording the last/known location of a movable artifact, it is critical to determine whether this location coincides with the object's original site [LS20]. For example, a group of sculpted panels now dispersed across museums in Sorrento, Rome, Berlin, and Washington D.C. may have once belonged to a single liturgical furnishing, likely a chancel screen, originally created for the Cathedral of Sorrento.

The research focuses primarily on mid-level objects, which are classified by their original function. Because objects may have lost that function, due to having been removed from their liturgical context, disassembled, or refunctionalized, this *resource typology* feature of the local data model may require different CRM mappings.

3.2. Data entry workflow and challenges

As the primary goal of *Mapping Sacred Spaces* is to catalog art historical items, the data pool is rationalized by making *artifacts* emerge from it: the high-level object model is artifact-centric, as is the data population workflow. This affects the structure of the input paradigms these scholars are accustomed to, namely tables and forms: we will concentrate on the latter, assuming that scholars, by filling semantic forms that create or amend artifacts (e.g. built complexes, churches, oratoria, altars or slabs), also provide peripheral data, such as places or people. This is aided by post-processing.

The project spatially organizes the artifacts in its catalog, both on the geographical, macroscopic level, and on the topographical one of monuments and complexes. This implies an interplay between spatial models on different representational levels, however, the geographical one offers more opportunities for interlinking as, to our knowledge, there are no public datasets detailing the interior topologies of Southern Italian religious buildings.

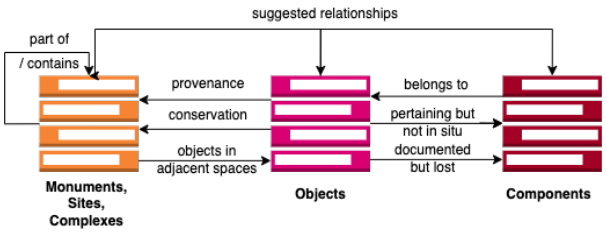

Figure 1: *Relations between types of artifacts according to the local model of Mapping Sacred Spaces.*

The curatorial process makes the relationship between complexes and monuments explicit. For that reason, scholars directly edit OpenStreetMap (OSM) and Wikidata, both being open datasets that allow refinement by the public. It was a design decision that polygons delimiting monuments and complexes should be edited directly on OSM, bearing in mind that OSM allows these to be drawn only if the perimeters are delimited by physical walls. This choice has the benefit of servicing the public with domain expertise, while data can still be imported into the project dataset using a *georeconciliation* function. Through this function, curators link a complex or monument to its OSM counterpart, which is achieved by proxying queries to Nominatim, an OSM Web API that allows search and lookup by toponym. Only one semantic link is created during this phase: at maintenance time, automated scripts take the burden of importing all the other relevant data, namely administrative information (e.g. cities, regions, countries) and coordinates (both points and polygons), and converting them to CRM. A combination of Nominatim and the OSM Overpass API is used to gather these data and update them on our database. Because OSM identifiers of buildings are volatile, and could change completely if another user were to redraw a polygon from scratch, we link to the corresponding Wikidata entity, which is also a property in OSM.

OSM identifiers of type *node*, used to identify cities and other administrative locations, are on the other hand more stable. Listing 1 shows a string search (on “cava”, which has matches in the Campania region), performed using a SPARQL wrapper of the Nominatim search service. This query can be parametrized and used as the autocomplete component of a KP that implements georeconciliation.

```
SELECT ?value ?label ?rtype WHERE {
  SERVICE osm:MSSNominatimSearchService {
    ?subject osm:q "cava" ;
    osm:accept-language "it,en" ;
    osm:osm_type ?typ; osm:type "administrative" ;
    osm:addresstype ?rtype ;
    osm:osm_id ?id ;
    osm:display_name ?label
  }
  BIND(IRI(CONCAT("http://www.openstreetmap.org/",
    ?typ, "/", ?id) ) AS ?value)
}
```

Listing 1: *Searching for administrative locations containing "cava" using a SPARQL wrapper for the OSM Nominatim API.*

After this query has built the OSM link, binding it to `?value`, a query to the Nominatim Lookup API can be used to integrate relevant data. This is performed as a *federated* SPARQL query (Listing 2), which writes to our knowledge graph, but needs to query third-party repositories; in this case, the Nominatim Lookup wrapper.

```
INSERT {
  <{value}> a crm:E53_Place
    ; crm:P168_place_is_defined_by ?wkt
    ; owl:sameAs ?wikidata
    ; crm:P89_falls_within ?pvc .
  ?pvc a crm:E53_Place
    ; crm:P2_has_type mss_type:province
    ; skos:prefLabel ?province
    .
} WHERE {
  SERVICE osm:MSSNominatimLookupService {
    ?subject osm:osm_ids "{osmid}" ;
      osm:osm_type ?typ;
      osm:osm_id ?id;
      osm:lat ?lat;
      osm:lon ?lon;
      osm:province ?pcode;
      osm:province_name ?province;
      osm:extratags 1;
      osm:wikidata ?wd .
  }}
  BIND(IRI(CONCAT("https://sacred_space.com/
␣␣␣␣resource/hertziana/place/", ?pcode)) AS ?pvc )
  BIND(STRDT(CONCAT("POINT(",?lon,"␣",?lat,")"),
    geo:wktLiteral) AS ?wkt)
  BIND(IRI(CONCAT("http://www.wikidata.org/entity/"
    ,?wd)) AS ?wikidata)
}
```

Listing 2: *Retrieving province, coordinates and Wikidata link.*

Chronological information needs to take into account values with precision up to the year of century, and also accommodate approximation or uncertainty, plus ranges. Rather than employing multiple fields to accommodate these cases, our approach proposes to use a single plaintext field and validate it against a subset of the Extended Date/Time Format specification (EDTF) of the Library of Congress [Lib22]. Listing 3 shows how a simple parser for years or year ranges – hyphen- or slash-separated – can be implemented in SPARQL, using regular expressions, within a KP for data insertion.

```
BIND("^(\\d+)\\s*(?:[/-]\\s*(\\d+))?$" AS ?re)
BIND(xsd:integer(REPLACE(?value, ?re, "$1"))
  AS ?begin)
BIND(xsd:integer(REPLACE(?value, ?re, "$2"))
  AS ?end)
BIND(BOUND(?begin) && BOUND(?end)
  && ?end > ?begin AS ?isRange)
```

Listing 3: *Parsing single years and year ranges in a SPARQL KP.*

These syntactic patterns can also be implemented in the SHACL constraint system, but having them in the KP allows the values to be used in the creation of CRM statements about the production of the artifact. Alternative approaches, such as autocompletion with values from the PeriodO gazetteer [GS16], are also possible.

Bibliographical information is always attached to the artifact entries. The approach for generating it as Linked Data is the same as with OSM. First, the team curates an external data source, namely a Zotero collection [Coh08]. Two SPARQL service wrappers, one for search and one for lookup, are configured for the Zotero REST API with access to that collection. The search wrapper is accessed in the bibliography-related KPs, for experts to link to sources in that collection. Finally, a post-processing script uses the lookup wrapper to retrieve essential publication information (e.g. publisher, editors, year) and store it into the knowledge graph to ease future reuse.

Resource types (which denote functions), materials and technique are pre-compiled on expert input, with terms extracted from the Getty AAT thesaurus, falling back to Wikidata or to custom terms if these are not in AAT. Because they amount to less than a hundred predefined terms, there is no need for direct access to AAT.

4. Implementation

Mapping Sacred Spaces was implemented on ResearchSpace version 2. This version ships with a SAIL service wrapper [RUK*13] that allows the Nominatim Search API to be queried in SPARQL: it was slightly modified to tailor it to the needs of the project and complemented with one for the Nominatim Lookup API to retrieve data for a known location. Other SAIL wrappers were added for the project's Zotero group. At the time of writing, with the data entry campaign underway, the dataset amounts to over 100k RDF triples.

ResearchSpace has a custom framework, called Ephedra, to integrate multiple repositories: this was configured to access SAIL services and the institutional knowledge graph, in order to provide recommendations to scholars. It would have been ideal for semantic forms to access Ephedra at data insertion time, so that entity labels and other essentials could be fetched as data are written to triplestore, but this is disallowed by a limitation in ResearchSpace. As a workaround, a post-processing pipeline was implemented in Python in order to integrate e.g. labels, geocoding and Zotero publishing information via Ephedra at maintenance time. To comply with rate limits, post-processing requests are throttled to one per second, ensuring responsible use as semantic data are enhanced.

Another ResearchSpace limitation is that it does not track the data entry activity of individual users, nor does it parametrize user IDs in knowledge patterns, which would have allowed us to implement a graph-based provenance system like in LED [ABB*19]: due to the small size of the research team, it was deemed acceptable for scholars to sign and timestamp their own contributions instead.

Web templates and SAIL service wrappers are packaged as a single ResearchSpace app, which is slated for release as open source alongside the knowledge patterns and, eventually, the project data.

5. Conclusion and Future Work

We have illustrated an approach to the curation of art history data, which leverages current semantic technologies in order to make the resulting dataset integrable with a larger knowledge graph, while mapping to the art historians' intended model at runtime. The implementation is packaged as an application and set of knowledge patterns for ResearchSpace and is intended as a blueprint for projects in other art historical domains to achieve interoperability.

Further work in the direction of tightening interoperability in *Mapping Sacred Spaces* includes the integration of more resources specific to art historians. The iconographic classification system Iconclass has been integrated as the way to categorize figurative decoration in the artifacts and its usage is picking up speed. Other sources that are likely to catalog part of the artifacts, such as the Italian Central Institute for Cataloguing and Documentation (ICCD), are being considered for linkage. Validation of the local data model against its own constraints, as well as those of CIDOC-CRM, will be implemented using the SHACL standard. The dataset is expected to be released as open data, with the ResearchSpace platform going public, once a critical mass of data on completely surveyed sites has been reached, targeting a late 2026 release date.

References

[ABB*19] ADAMOU A., BROWN S., BARLOW H., ALLOCCA C., D'AQUIN M.: Crowdsourcing linked data on listening experiences through reuse and enhancement of library data. *Int. J. Digit. Libr. 20*, 1 (2019), 61–79. doi:10.1007/S00799-018-0235-0. 2, 4

[Car24] CARBONI N.: Ontological patterns for modeling the validity of spatiotemporal statements. In *Proceedings of the fourth edition of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage, Tours, France, October 30-31, 2024* (2024), Bikakis A., Ferrario R., Jean S., Markhoff B., Mosca A., Asmundo M. N., (Eds.), vol. 3809 of *CEUR Workshop Proceedings*, CEUR-WS.org. URL: https://ceur-ws.org/Vol-3809/paper5.pdf. 2

[Coh08] COHEN D. J.: Creating scholarly tools and resources for the digital ecosystem: Building connections in the Zotero project. *First Monday 13*, 8 (2008). URL: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2233/2017. 4

[GS16] GOLDEN P., SHAW R.: Nanopublication beyond the sciences: the periodo period gazetteer. *PeerJ Comput. Sci. 2* (2016), e44. doi:10.7717/PEERJ-CS.44. 3

[Kra23] KRABINA B.: Building a knowledge graph for the history of Vienna with Semantic MediaWiki. *J. Web Semant. 76* (2023), 100771. doi:10.1016/j.websem.2022.100771. 2

[Lib22] LIBRARY OF CONGRESS: Extended date/time format (EDTF) specification, 2022. Accessed: 2025-04-16. URL: https://www.loc.gov/standards/datetime/. 3

[LS20] LONGO R., SCIROCCO E.: Sakralen Raum kartieren. Forschungsbericht der Max-Planck-Gesellschaft, 2020. https://www.mpg.de/16151367/biblhertz_jb_2020?c=152805. 2

[NCM19] NISHANBAEV I., CHAMPION E., MCMEEKIN D. A.: A survey of geospatial semantic web for cultural heritage. *Heritage 2*, 2 (2019), 1471–1498. URL: https://www.mdpi.com/2571-9408/2/2/93, doi:10.3390/heritage2020093. 2

[OT18] OLDMAN D., TANASE D.: Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part II* (2018), Vrandecic D., Bontcheva K., Suárez-Figueroa M. C., Presutti V., Celino I., Sabou M., Kaffee L., Simperl E., (Eds.), vol. 11137 of *Lecture Notes in Computer Science*, Springer, pp. 325–340. doi:10.1007/978-3-030-00668-6_20. 2

[RUK*13] RAKHMAWATI N. A., UMBRICH J., KARNSTEDT M., HASNAIN A., HAUSENBLAS M.: A comparison of federation over SPARQL endpoints frameworks. In *Knowledge Engineering and the Semantic Web - 4th International Conference, KESW 2013, St. Petersburg, Russia, October 7-9, 2013. Proceedings* (2013), Klinov P., Mouromtsev D., (Eds.), vol. 394 of *Communications in Computer and Information Science*, Springer, pp. 132–146. doi:10.1007/978-3-642-41360-5_11. 4