



The Semantic Reference Data Modelling Method: Creating Understandable, Reusable and Sustainable Semantic Data Models

DISCUSSION PAPER

GEORGE BRUSEKER 

NICOLA CARBONI 

MATTHEW FIELDING 

DENITSA NENOVA 

THOMAS HÄNSLI 

*Author affiliations can be found in the back matter of this article

 ubiquity press

ABSTRACT

As ontologies grow in complexity and breadth, domain experts often struggle to understand their intricacies without the assistance of semantic experts. As a result, the ontologies themselves can become significant obstacles in constructing reusable and consistent knowledge graphs of their data following these standards. To address these challenges, this contribution presents the Semantic Reference Data Model (SRDM) modelling method, a protocol that offers a practical approach to (i) simplifying ontological complexity, (ii) standardising semantic patterns and (iii) facilitating the creation of new knowledge graphs. SRDM provides a solution for overcoming conceptual challenges by offering a catalogue of entity-based templates, whereby recognisable entities within a domain are documented in single templates composed of ready-made and distinct ontological patterns covering the entity's most documented attributes. The limited number of elements available, the use of domain-specific language, and the correspondence between documented objects and real-world items make the SRDMs particularly easy to grasp, helping bridge the gap between semantic and domain experts. A case study is used to illustrate the framework's application in the cultural sector, specifically highlighting the advantages obtained in documentation, data consistency, and external data ingestion. The case study demonstrates how SRDMs can vastly simplify the creation and management of knowledge graphs, helping Digital Humanities and Cultural Heritage communities easily share essential and useful datasets.

CORRESPONDING

AUTHOR:

Nicola Carboni

School of Information Sciences
University of Illinois Urbana-
Champaign, USA

nicola.carboni@unige.ch

KEYWORDS:

linked data; cultural
heritage; CIDOC-CRM; digital
humanities; semantic web;
data curation

TO CITE THIS ARTICLE:

Bruseker, G., Carboni, N.,
Fielding, M., Nenova, D.,
& Hänsli, T. (2025). The
Semantic Reference Data
Modelling Method: Creating
Understandable, Reusable
and Sustainable Semantic
Data Models. *Journal of Open
Humanities Data*, 11: 22,
pp. 1–14. DOI: [https://doi.
org/10.5334/johd.282](https://doi.org/10.5334/johd.282)

1 CONTEXT AND MOTIVATION

The Semantic Reference Data Model (SRDM) modelling method is a proposal for creating well-documented, reusable semantic data modelling projects via composable data patterns. It aims to address a gap in the semantic data management workflow by tackling the challenges domain experts face in understanding and adopting semantic frameworks (Lozano-Tello & Gomez-Perez, 2004) and the difficulties semantic modellers encounter in communicating the ontologies that structure them. In this article we will a) explore the motivations for such a proposal, b) explain the core of this proposal regarding the suggested documentation protocol and c) illustrate the proposal through a use case, the Swiss Art Research Infrastructure.

Formal ontologies are intended, through the removal of extraneous ambiguity (Zeng, 2008) and the creation of a shared information space (Guarino et al., 2009), to serve as a lingua franca (Crofts et al., 2003) for domain experts working with diverse but overlapping datasets (Hitzler, 2021). In principle, an ontology should provide the user with the conceptual framework necessary to begin the elaboration of a semantic data model that will result in a given dataset's re-expression according to a communally agreed-upon, standardised schema (Trojahn et al., 2022). In practice, however, a disconnect is often encountered here (Ristoski & Paulheim, 2016). Ontological commitments built into the ontology can move the resulting semantic model towards a level of abstraction that begins to disassociate it from the source data, rendering one user's semantic expression of their data foreign to another's (Westerinen & Tauber, 2017). This ends up recreating the very gaps and ambiguities that had threatened the interoperability of the data in the first place. The possibility to use divergent expressions within the framework of a common ontology is a design feature that ensures conceptual manoeuvrability to knowledge engineers. Nonetheless, communicating across communities of experts in a language that makes sense to them remains a challenge which must be addressed if any given alignment is to be successfully shared with the broader community. This problem is particularly acute with foundational ontologies (Borgo et al., 2022). A notable example is CIDOC-CRM, the de facto standard ontology used for sharing humanities data (Bruseker et al., 2017). CIDOC-CRM excels at integrating datasets from across the diverse domains that comprise the humanities. However, its flexibility also allows the same input dataset to be modelled using different structures, potentially compromising interoperability and reusability. A typical result is that the adoption and application of the ontology to create a semantic data model does not necessarily lead to a seamless, reusable representation of the original dataset that the domain expert expects to realise.

To take a very basic example, the expression of a commonly-employed field like "title", using CIDOC-CRM, should be straightforward. However, CRM offers a number of classes and relations for the semantic representation of the notion that some object has a name. The basic property available is 'P1 is identified by', which has the root class E1 CRM Entity as domain and E41 Appellation as range. E41 Appellation has, in turn, a series of sub-classes, including E35 Title, which may be used for expressing titular names. Because all classes in CIDOC-CRM can be typified, there are different ways one could express the formal proposition implied in the term "title". For example, it can be expressed by modelling the fact that an entity is identified by a specific type of appellation: a title. Using an easy-to-grasp notation it can be represented as a series of edges and nodes:

```
E1 CRM Entity → P1 is identified by → E41 Appellation → P2 has type →
E55 Type "Title"
```

However, using the same type of notation, we can express the fact that an entity is identified by a title, a conceptual specification, and thus a subclass, of an appellation:

```
E1 CRM Entity → P1 is identified by → E35 Title
```

This is only a very simple example of the heterogeneity of possible expressions in CIDOC-CRM that arise because of the ontological commitments of its development, but such cases create a significant challenge to domain experts who wish to begin engaging in modelling tasks. There are, in fact, many additional semantic patterns that can be used to express the title of an entity. Each of the patterns in Listing 1 represents legitimate modelling choices, employed and used by different modellers. One may choose to favour one version or the other. At scale, this will make the data fairly incompatible, specifically at a query level, where each dataset may require a different SPARQL query to retrieve basic information containing the same formal propositional content.

```

ex:artwork rdfs:label rdfs:Literal.
ex:artwork crm:P1_is_identified_by crm:E41_Appellation .
    crm:E41_Appellation rdfs:label rdfs:Literal .
ex:artwork crm:P1_is_identified_by crm:E33_E41_Linguistic_Appellation .
    crm:E33_E41_Linguistic_Appellation crm:P190_has_symbolic_content rdfs:Literal .
ex:artwork crm:P1_is_identified_by crm:E35_Title .
    crm:E35_Title skos:prefLabel rdfs:Literal.
ex:artwork crm:P102_has_title crm:E35_Title .
    crm:E35_Title crm:P190_has_symbolic_content rdfs:Literal.
ex:artwork rdfs:label xsd:string.

```

Listing 1 Codeblock representing different ways to assign a name to an artwork using CIDOC-CRM. Prefixes used are listed on [prefix.cc](https://www.prefix.cc/).

From this simple example drawn from CIDOC-CRM, we can see that an ontology pays for the expressivity required to adequately cover a given domain at the cost of preserving, and even to some extent encouraging, a basic heterogeneity of expression that puts the reusability of the semantic data model at risk. The knowledge gap that semantics was originally intended to overcome thus risks being reintroduced, with the model itself reiterating the problems it was supposed to fix. The motivation to address this issue at a general level arose from a specific project aimed at developing a set of semantic models to support the Swiss Art Research Infrastructure (SARI).¹ SARI serves as a knowledge-hub for cultural heritage organisations, both nationally and internationally. SARI's position as a broker of sustainable data creates a strong need to process and deliver data streams from different authoritative sources through a single, consistent data pipeline. The question arose of which ontology to use for this representation, how to create reasonably representative semantic models for this data space and community, how to create stable modelling patterns for basic entities of the ontology, how to communicate effectively the ontological models created and how to apply them in order to guarantee the maximum level of compatibility across datasets.

2 DATASET DESCRIPTION

The dataset contains all the SRDM models developed for the Swiss Art Research Infrastructure, and it covers the following entities: Person, Artwork, Group, Builtwork, Place, Digital Object, Event, Bibliographic Entity, Image, Physical Information Carrier, Archival Unit.

REPOSITORY LOCATION

doi.org/10.5281/zenodo.14619668

REPOSITORY NAME

Zenodo

OBJECT NAME

Swiss Art Research Infrastructure – Semantic Reference Data Models

FORMAT NAMES AND VERSIONS

Four CSV files representing the models (SRDM_SARI_Models.csv), fields (SRDM_SARI_Fields.csv), collection (SRDM_SARI_Collections.csv) and categories (SRDM_SARI_Categories.csv). The file SRDM_Data_Protocol_Definitions.csv provides a description for all the column headers used across the files. See section 3.1 for more information on models, fields, collections and categories.

CREATION DATES

Initially created in 2020–2021 and updated over time. The current version has been exported the 2025-01-08.

DATASET CREATORS

George Bruseker, Nicola Carboni, Denitsa Nenova

¹ <https://swissartresearch.net/>.

LICENSE

Creative Commons Attribution 4.0 International

PUBLICATION DATE

2025-01-09

3 METHOD

In order to provide a functional solution to the issues in section 1, we introduce the Semantic Reference Data Model (SRDM) modelling, a method for documenting and reusing semantic data patterns that aims, inter alia, to help domain users adopt consistent semantics without necessarily having to face complex ontological, epistemological, or even technical questions from the outset. This method consists in a protocol for the construction and documentation of semantic data structures, based on a format that helps remove some of the barriers to uptake while providing a means to track the provenance of those structures throughout their life-cycle for future consultation. The result is a collection of ‘recipes’ for putting semantic data to use; patterns (Gangemi & Presutti, 2009) that can be used by domain experts to begin implementing their data on their own accord (Hammar et al., 2016) and thus achieve a move towards semantic interoperability without having to start from the ground up in knowledge engineering.

The execution of an SRDM project requires the development of two basic elements: (i) a documentation structure and (ii) a method for populating these structures.

3.1 THE DOCUMENTATION STRUCTURE

The SRDM modelling method proposes a mid level documentation structure that allows the description of core entities and their properties in a composable and reusable manner. Key to this approach is the notion of a **reference entity**. We conceive of a reference entity as a well defined and easily recognised object of documentation within a project’s domain of discourse. A reference entity is a locus of documentation which together with other reference entities make up a constellation of objects that constitute the standard reference points of the documentation that will be subject to semantic formalisation. Put another way, reference entities are the real-world things about which a domain expert gathers information as a unit and which they wish to connect together to other units of information to make a meaningful representation of the world of study. By speaking about reference entities we shift the territory of the modelling discussion back to a familiar ground, talking about the entities that domain experts want to describe, reintroducing in the semantic realm the more familiar approach to organising data that domain experts are accustomed (e.g., tabular forms and fields).

The total set of all reference entities to be modelled as such, as well as the properties to be documented for each of them, depends on the needs of each project. Some may have reason to choose as a reference entity a generic ‘visual object’, while others may focus on ‘paintings’ or postcards. The choice only depends in the first place on the datasets at hand and the overall scope/ambition of the modelling activity. There is not and can never be one single schema that will uniquely decide what kinds of entities must be used to structure a semantic expression of a given dataset or domain; hence, the challenge of semantic heterogeneity, mentioned above. However, what can maintain a level of interoperability is the use of sets of semantic patterns for shared entities and properties. What the SRDM method thus offers is a protocol for creating a documented ready-made set of semantic patterns that can be reused across projects and institutions, and which are explained in terms that connect end users to the semantics.

According to the SRDM method, each selected reference entity is documented using three levels which work together compositionally (See Figure 1) to create a reusable and understandable meta documentation of a domain or project. In particular we define:

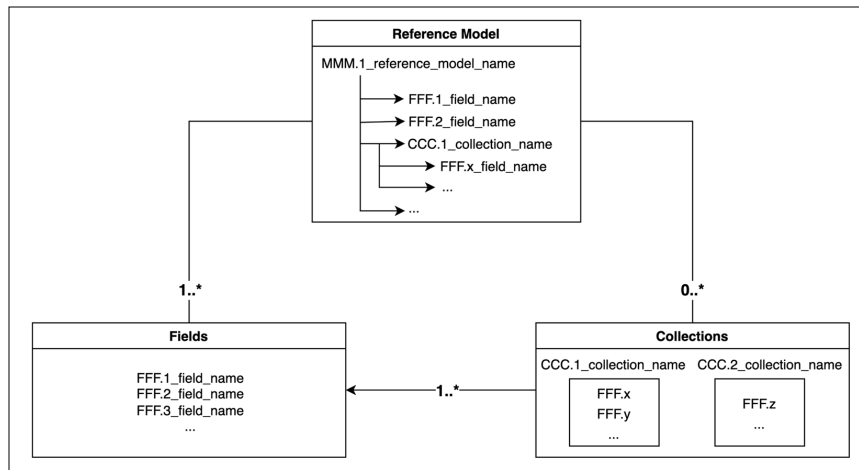


Figure 1 Overview of basic documentation structure.

Model: The main unit of documentation of a reference entity. Each model records and defines a set of fields that are generally used for the documentation of a given reference entity in a domain of discourse or project. A model is matched to a class in a target ontology (or multiple classes if necessary for the purpose of multi-instantiation) as its domain of semantic expression.

Field: A unit of documentation used to express a fact about a reference entity, expressed in terms of a property that can potentially (but not necessarily) hold for that entity within a model. Fields correspond to traditional input categories in data tables, spreadsheets, etc. (Note, however, that several fields may be required to capture the semantics of one data entry field in a traditional spreadsheet.) In an SRDM project, a field consists – at minimum – of a label, a description of the field’s intended use and a semantic translation that shows how the field is to be implemented in a semantic database. Fields are constructed to be reusable across models whenever possible.

Collection: A collection is a stable and functional bundle of fields, which represent units of documentation across models. Fields in a collection are commonly linked to one another as part of a related cluster of information that is not normally independently documented.

The use of these core constructs aims to be intuitive to domain specialists and semantic modellers alike. For example, the description of ‘photographs’, in a colloquial sense, may entail the description of one or more reference entities. What exactly is meant by ‘photograph’ in the domain or project must be elicited. What are the unique documented entities which it wishes to make independent statements about? The photographic content, the material support, the digital reproduction of the photo, all are potentially different reference entities that reflect distinct aspects of the object photograph. Then again, ‘photograph’ could mean only the material support. Therefore, the description of ‘photograph’ may entail multiple reference entities and related models or only a single reference entity, depending on the use-context and discourse space. We illustrate below a case where the unanalysed referent ‘photograph’ coalesces into two reference entities (the real-world objects about which statements are made and documentation gathered) and two models developed to reflect these. In this case, we identified two reference entities, one describing the physical support and the other describing the intellectual content reproduced across its multiple manifestations (Figure 2). The model documenting the physical support may thus contain information about where it can be found, what are its dimensions, its current condition, the technique used to produce it and so forth. These are the fields that fall within the scope of the object under consideration qua physical object. One might call this model, for example, “Photographic Support” The model documenting the intellectual content may contain information about the event, place, object or subject depicted therein, the iconographical status, but also about the creator of the image. These fields fall under the scope of the object qua visual representation. One might call this model, for example, “Photographic Content”.

The SRDM modelling process consists of constructing a coherent set of fields that define the qualities typically ascribed to referenced entities within the project’s domain, as models,

providing them an identity and semantic definition. Fields can then be further bundled together into mid level patterns, called collections, which document fields typically tied together in documentation contexts, which are also given a fixed ontological scope. A typical example of this would be ‘name’ (in CIDOC, scoped to any E1 CRM Entity) or ‘timespan’ (scoped to E2 Temporal Entity), in which a canonical set of fields that are commonly found together (e.g.: the name itself, its type, its language, etc). Collections allow us to identifying common patterns that go beyond individual models and save us from having to create unique fields for different models that have the same semantic structure for, e.g., a person’s name, an event’s name, a thing’s name, etc., as the ‘name’ collection itself is scoped to the more generic E1 CRM Entity class, which applies across all of its subclasses.

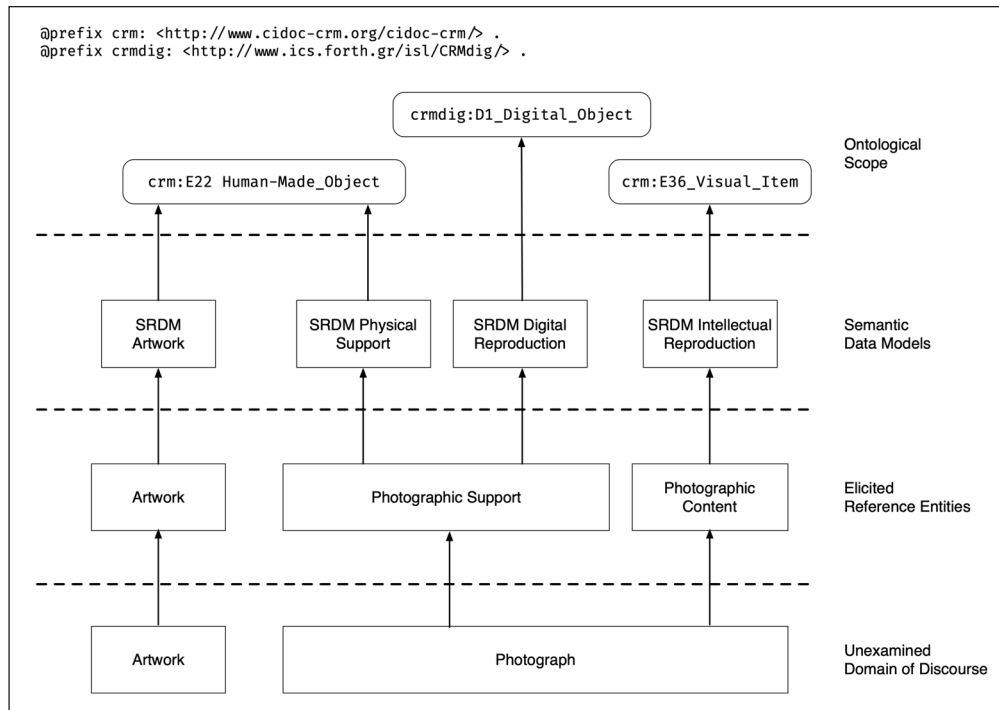


Figure 2 Relationship between reference entities, semantic reference data models and ontological scope. The analysis of a domain of discourse elicits a set of reference entities which reflect the documentation needs. Each reference entity is then described using one or more reference data models.

The purpose of the SRDM is to describe each of these units of information – field, collection, and model – using terms familiar to the domain specialists, alongside the abstract ontological classes typically found in a formal ontology, and thus provide a semantic definition of how the relevant information in a given dataset or domain is to be represented in the target ontology. The core attributes for SRDM fields are summarised in [Table 1](#). These include a uniquely identified semantic path along with an ontological validity scope, backend and frontend naming conventions, and prose description of the intended use of the pattern. All of this aims to create a recipe for the consistent representation of semantic expressions that speak to a variety of users in their own terms. In the table below we see the metadata requirements for the field pattern in the SRDM documentation standard. Similar requirements are specified for collections and models. Key within these is the distinction between those field specifications, which, once declared and documented, can be reused across multiple reference models, and those which vary between models ([Figure 4](#)).

ATTRIBUTE NAME	FIELD DESCRIPTION	EXAMPLE
Identifier	A unique, stable identifier to identify the field across usage contexts.	fie_92_Coordinates
System Name	A developer-friendly name for the field to be used in data modelling/mapping contexts.	wkt_coordinates
UI Name	A user-friendly name for the field to be used in user interface. Naming conventions may change to reflect intended interpretation of the field as employed within a specific reference model as distinct from another.	Coordinates
Description	A user friendly prose description of the intended use/function of the field.	This field is used to indicate the coordinates of the documented geographic place.

Table 1 Semantic Reference Data Model Metadata Specifications.

ATTRIBUTE NAME	FIELD DESCRIPTION	EXAMPLE
Ontological Scope	An ontological class that provides the maximal ontological scope for the field according to its defined function. Whenever the field is to be employed in a reference model, the ontological scope of the model must match or fall within the ontological scope of the field (i.e., the scope of the field must be equal to or wider than the reference entity captured in the reference model).	crm:E53_Place
Semantic Path	Edge and node representation of a defined semantic path syntax representing the meaning of the field in the target ontology. Uses a more readily human-readable notation form for class and property abbreviations.	E53 → P168 → geosparql:wkt
RDF Encoding	An RDF representation of the Semantic Path using Turtle Syntax.	<https://ex.org/place/ fie_75_1> a crm:E53_ Place; crm:P168_ place_is_defined_by “^^^geosparql:wkt.
Expected Value Type	The kind of data value the field expects (e.g.: string, integer, date, concept, collection, reference model, URI)	Well-known text (WKT)

3.2 THE DOCUMENTATION METHOD

The goal of the SRDM documentation protocol is to create a common space for collaboration between domain specialists and knowledge engineers, which will result in a multipurpose set of models, collections and fields that are sufficiently representative of the domain in question, semantically accurate according to the target ontology and amenable to practical implementation. Having outlined the requisite documentation structure above, it is still necessary however to define a method for populating it and so deriving some set of fields, collections and models that are adequate to a given project or domain. The method proposed here is iterative (See [Figure 3](#)), but requires an initial setup which consists in a collaboration between domain experts and knowledge engineers, who together circumscribe the domain of interest and determine the scope of the project. Identifying the scope of the project consists in gathering the primary materials (typically non-semantic) needed to determine the entities that are of interest and the broad contours of the information space in which they are documented. With this at hand, a target ontology can be selected which provides sufficient coverage for the given modelling project, at which point a list of reference entities to be modelled should be proposed and the data sources should be analysed to ascertain typical assertions made for each of them. The goal of this activity is to derive a representative list of typical properties used to describe the reference entities, enabling the final SRDM models to function as a recipe by which to render typical data in the domain into the target semantic form.

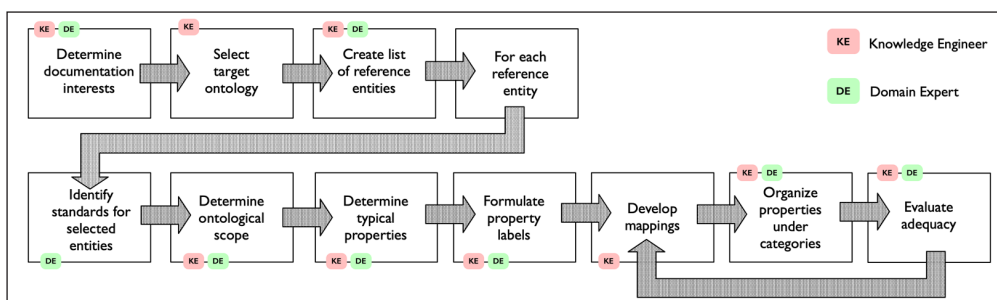


Figure 3 The process of creation of the Semantic Reference Data Models follows a series of iterative steps, shared by domain experts and knowledge engineers, to determine the relevant reference entities. Following this process, each entity is collaboratively examined to identify and determine the more comprehensive and reusable ontological patterns for describing it.

The selection of typical assertions is ideally done on the basis of a comparative analysis of the fields deployed in sources having similar intensional content. There is no hard and fast rule to determine which properties from the selected ontology can and should be used to flesh out the reference model. At its most basic level, properties with the same intension, as identified by the semantic modeller and domain specialist, that are repeated across many schemas are more likely to be of interest. On the other hand, properties that occur only rarely but capture an interesting or important element for the domain may be the result of a specialist schema that would be of wider use were it raised to a more generic level. The model should not be based on quantitative analysis alone. This process is non-deterministic and depends on the general ambition of the project for

which a set of SRDM patterns is being made. As such, the SRDM method still closely follows the basic strategy of formal ontology development with regard to faithfulness to data schemas. It differs, however, in that it does not seek to exhaustively describe the entities within the research space with regards to all its potential properties, but focuses on strictly defining a semantic representation for typical properties using an extant ontology. The SRDM documentation protocol is at an intermediate level between extant data and the ontology itself, providing an interface by which to understand how to apply an ontology in typical cases and make decisions on how to consistently model the same data to avoid the heterogeneity problem identified above. For each reference entity then, an iterative process should be followed that includes:

1. Identify and analyse standards, guidelines, protocols or schemas documenting an entity. (see section 5.1.1).

This step selects from the overall data structures gathered and identifies relevant documentation about the selected entity. We must select the widest scope of documents relevant to the reference entity in the specified domain. Finding extant data standards relative to the reference entity provides lists of properties of interest relevant to it.

2. Determine the appropriate ontological scope for the documented entity (see section 5.1.2).

Once the reference entity has been identified it should be matched to one or more classes in the target ontology.

3. Determine the typical properties used in the documentation of the entity (see section 5.1.3).

The task of identifying the attributes of interest associated with the reference entity is a question of abstracting, in a pre-ontological way, identical attributes for identical things. For example, sources a and b may use different labels to identify the property 'production date', but the informational content is the same. Or indeed, may use identical labels to document semantically divergent attributes. Consulting common data structures and standards provides insight into what are the more and less frequently used properties and motivates the decision of what fields need to be declared for the entity.

4. Formulate property labels and definitions relevant to the domain (see section 5.1.4).

On the basis of the identification of typical properties, the SRDM modeller can declare a list of fields (in the SRDM sense) that apply to the entity giving them appropriate names for the domain user. This declaration should also include the attribution of an identifier and a description of the intended interpretation of the field.

5. Develop a mapping to a target ontology (see section 5.1.5).

This step redounds to the semantic modeller who should assign a unique semantic path from the target ontology to each declared field in the SRDM project. As the modelling proceeds, common fields which had been specified for a specific entity (e.g.: text creation) but which can be specified more generally in the target ontology (e.g.: conceptual creation), may be generalised in both naming and semantics in order to cover the broadest possible set of use cases and reduce the need for field declaration. Care should be taken here not to follow the path of the target ontology and potentially diverge too heavily from the familiar terms by which entities can be described.

6. Organise the typical properties under common categories (see section 5.1.6).

SRDM model users need clear guidance on the types of information a model contains and where it is located. Organising properties under specific information categories helps users locate and understand the fields by grouping them into meaningful and relevant categories.

7. Evaluate the adequacy of the resulting model (see section 5.1.7).

In this phase we can return to the documentation used as a source for abstracting the common attributes and assess the overall coverage of the SRDM model.

Following the above process iteratively for each target reference entity will create the set of models that define the project and describe the typical set of attributes employed by the domain specialist within it. The outcome is a reusable documentation structure that facilitates communication of the semantic data the domain users wish to produce and consume and serve as a guide to understanding, implementing and eventually querying the model. While SRDM

models are meant to be beneficial in multi-institutional projects that involve data exchange or the creation of shared data spaces, they are equally applicable for use by individual researchers or within institutions. Additionally, they can be repurposed across various projects or organisations with comparable documentation needs. After a set of SRDM models has been established for the entities in a specific domain or project scope, the fields defined there can be shared among institutions, allowing them to align with, adopt, or expand an SRDM model, fostering a more efficient process for semantic data generation. At this stage of development, the SRDM modelling method has been developed as a practical solution to, rather than a formal theory over, the problem of creating reusable and understandable semantics for domain experts. The key structures proposed and the approach to creating such structures are outlined above. Below we will look at the first concrete application of this method by way of using this as a demonstration of the approach itself.

4 RESULTS AND DISCUSSION

From the outset, the SRDM method was designed with two primary aims. First, it sought to provide digital humanists with stable modelling patterns for basic entities using a formal ontology, focussing specifically on CIDOC-CRM in the first instance. This was intended to help domain experts bootstrap semantic projects and create truly **interoperable** data (Wilkinson et al., 2016). Second, the method aimed to develop a comprehensible semantic metadata documentation protocol that would allow the models themselves to be easily understood by diverse stakeholders, and thus reused, extended and expanded over time. The overarching vision is to foster a distributed and accessible repository of semantic data models, ready to be deployed by domain specialists and so empower data producers to take greater control over the expression of their own data. To this purpose, the SARI SRDM documentation has been made available freely online² and serves as a reference set of data modelling patterns. It provides SRDM style documented modelling for 14 models, 55 collections and 279 fields meant to support the application of CIDOC-CRM to art historical modelling contexts. Compared to the abstract definitions and explanations of the ontology itself, the SRDM documentation protocol aims to provide more approachable, understandable paths into the semantic representation of humanities data. With regards to uptake, a survey of the use and usability of the models by end users would provide valuable feedback for refining and moving forward with this programme. With regard to the spread of the method, an additional effort is required. The initial elaboration of the proposed documentation structure and method of deriving SRDM models is part of such an effort. Supporting the method also requires the full elaboration of the semantic documentation data model of which an initial documentation is provided in this article. For both, workshops and tutorials need to be organised in order to demonstrate and explain the procedures and empirically verify their adoptability by the broader community, adjusting the method based on these inputs. Furthermore, the elaboration of a software platform that would enable the controlled production and display of such semantic data documentation in a repeatable and comparable way would also strongly facilitate the uptake of the method. Moreover, while conceived to potentially be used with any ontology the method has only been tested with regards to the CIDOC-CRM model and its extensions. While there is no principled reason why it could not be used for any other target ontology, proof of its effectiveness with other ontologies is something that remains for future research and demonstration. The methodological foundations outlined in this article provide the necessary groundwork to be undertaken in order to support this next step.

5 APPLICATION

The first application of this method was in the SARI semantic data modelling project. This section outlines the initial SRDM setup phase and demonstrates the method using a single reference entity as an example.

The project goal was to create a set of standardised models for entities typically referenced by art historians. Although broadly scoped to encompass art and architectural history, the initial focus narrowed to key reference datasets used by scholars in these fields. From these datasets, a list of reference entities was defined, reflecting commonly documented real-world entities relevant

² <https://docs.swissartresearch.net/>.

to scholarly discourse. These included: Person, Group, Artwork, Built Work, Bibliographic Item, Digital Object, Event, and Place. Next, a suitable target ontology was identified to represent this information, that being CIDOC-CRM. The choice was made for reasons that include the fit between the scope of the ontology and that of the project, its use for information exchange by many cultural research infrastructures,³ and its adoption by many software development groups (Enriquez et al., 2018; Scholz & Goertz, 2012; Oldman & Tanase, 2018). With the ontology in hand, the effort to develop the SRDM in terms of the relevant models, fields and collections could begin.

5.1 SRDM DEVELOPMENT

To explore the development of an individual SRDM model and demonstrate the methods and strategies outlined above, we will look at the example of the model derived for the entity **Person**.⁴

5.1.1 Step 1. Identify and analyse standards, guidelines, protocols or schemas documenting an entity

Research into the development of the person model began with the decision to create a model for ‘artists’ to accompany ‘artworks’. Relevant data models from within the domain of art and architecture or related fields were sought out as evidence for the properties to be modelled (See Table 2). These were chosen for analysis if they were dedicated to or significantly overlapped with common metadata schemas for artists. Given the frequent need to reference persons in art historical contexts, numerous schemas were available for analysis. However, despite their diversity, the number of relevant cross-schema properties was relatively small. This likely reflects the fact that “Person” is often treated as a secondary referent in relation to a primary object of interest, such as artworks, and is therefore modelled with a fairly consistent set of typical properties, including name, birth and death dates, place of residence, and group memberships.

ACRONYM	SOURCE NAME	MAINTAINED BY
Agrelon	Agrelon, an Agent Relationship Ontology	Deutsche Nationalbibliothek
SIKART	Dictionary of Art in Switzerland	SIK-ISEA
MARC 21	Marc 21 – Bibliography Heading Fields	Library of Congress
VIAF	Virtual International Authority File	OCLC
ULAN	Union List of Artist Names	Getty
Schema.org	Schema.org	Schema.org
CDWA	Categories for the Description of Works of Art	Getty
CCO	Cataloging Cultural Objects	CCO Commons
VRA Core	Visual Resources Association core categories	Visual Resources Association

Table 2 Data Sources and Access Points. Source: <https://docs.swissartresearch.net/et/persons/>.

5.1.2 Step 2. Determining the appropriate ontological scope for the documented entity

The selection of the reference entity leads to a decision point regarding the definition of the ontological scope for the SRDM model being developed. In this case, the immediate and obvious move with regard to this model of artists in particular was to widen the scope to persons in general. Ontologically speaking, the attribute of ‘artist’ is non-essential for the description of a person and the range of persons of interest relative to art history goes well beyond the artists themselves, to include a myriad of actors who interact with artworks in diverse ways. Thus while we began with a model for ‘artist’, we quickly changed the intended reference entity to person. From this change of perspective it became simple to choose an appropriate class in the target ontology, E21 Person to anchor this model. While this may seem obvious at first glance, we note in passing that we followed the alternate strategy in the case of the SRDM models for ‘artwork’ and ‘builtwork’, both of which are scoped to the CRM class E22 Human-Made Object. In this case the reference entities were deemed to be different enough, and documented differently enough in the source data, to warrant separate models.

³ To mention a few: Ariadne (<https://ariadne-infrastructure.eu/>), Parthenos Project (<https://www.parthenos-project.eu/>), Pharos Project (<https://www.parthenos-project.eu/>).

⁴ <https://docs.swissartresearch.net/et/persons/>.

5.1.3 Step 3. Determine the typical properties used in the documentation of the entity

The next step in setting up the SRDM model is to determine the properties of general interest with regard to the entity being modelled. For the purpose of our Person reference model, typical properties were chosen with regard to relevance and include, among others, naming convention (e.g., primary name, alternative names, name used only for a period), date and place of birth, date and place of death, group affiliations (e.g., national, cultural, institutional), gender, occupation, social relationships (generic or specific), performed activities (e.g., period and field of activity), knowledge (e.g., language), education, documentation (e.g., citation, images, biography documents), and so forth. One can see from this selection that many of the fields defined in order to express the relevant data semantically will potentially have applications to other reference models beyond person alone – e.g., naming conventions for artworks, activities performed by groups, bibliographic documentation of built works, etc – and so benefit from the SRDM strategy of composability and reuse (See [Figure 4](#)).

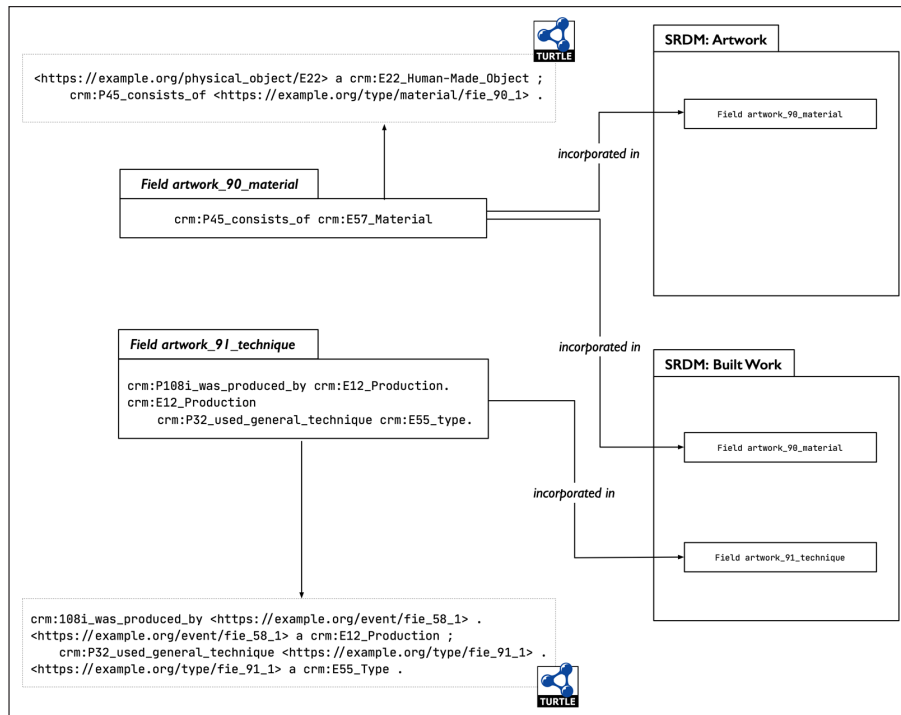


Figure 4 Interconnection between fields and SRDMs. A field (example: artwork_90_material) can be interlinked with multiple SRDMs or with only one (Pattern artwork_91_technique).

5.1.4 Step 4. Formulate property labels and definitions relevant to the domain

At this stage we assigned identifiers, picked readily understandable labels for the model's fields and defined simple, guiding definitions for the understanding and application of the selected properties. This can be illustrated, for example, by the declaration of the field “person_75_birth_location” (See [Figure 5](#)), and its attribution of the label ‘Birth Location’ and the attribution of the simple definition “This field is used to indicate the place of birth of the documented person”. Such field definitions were crafted to be readily understandable to domain experts with no further explanation required. Compared to the scope notes of the relevant properties and classes of the target ontology, this field is directly understandable and applicable. As noted above, providing the fields with identifiers and scoping them to their most broad application has made them reusable units for documentation in diverse contexts and projects.

5.1.5 Step 5. Develop a mapping to a target ontology

Once the common attributes have been identified, named and described in natural language, we assign them a semantic path, an ideal propositional form in the target ontology which has the same semantic form as the meaning of the described field in natural language. A semantic path takes the form of a chain of classes and properties from the target ontology, here CIDOC-CRM, required to express the semantics of a field. It is encoded both in RDF and in a human-readable syntax, formed with only classes and properties identifiers. For example, the semantic path:

E21 → P98i → E67 → P7 → E53 → P1 → E33_E41 → P190 → rdfs:Literal

expresses, in a human-readable syntax, the CRM mapping used for indicating the birthplace of the documented person. As different fields may document different attributes of the same event, for example birthplace and birthdate, we introduced a variable in the human-readable syntax of the semantic path. For example, the ones below

E21 → P98i → E67[x] → P7 → E53 → P1 → E33_E41 → P190 → rdfs:Literal

E21 → P98i → E67[x] → P4 → E52[y] → P82a → xsd:dateTime

E21 → P98i → E67[x] → P4 → E52[y] → P82b → xsd:dateTime

describe, in a human-readable syntax (a turtle translation is provided in [Listing 2](#)), the birthplace of the documented person as well as their earliest and latest known date of his birth. Note, that as the nodes representing the classes E52 and E67 are used across different fields, they are coupled with a variable indicating that the expression refers to the same node.

```
<https://example.org/actor/E21> a crm:E21_Person ;
  crm:P98i_was_born <https://example.org/event/fie_73_1> .

<https://example.org/event/fie_73_1> a crm:E67_Birth ;
  crm:P4_has_time-span <https://example.org/time_span/fie_73_2> ;
  crm:P7_took_place_at <https://example.org/place/fie_75_1> .

<https://example.org/place/fie_75_1> a crm:E53_Place ;
  crm:P1_is_identified_by <https://example.org/name/fie_10_1> .

<https://example.org/name/fie_10_1> a crm:E33_E41_Linguistic_Appellation ;
  crm:P190_has_symbolic_content ""^^rdfs:Literal .

<https://example.org/time_span/fie_73_2> a crm:E52_Time-Span ;
  crm:P82a_begin_of_the_begin ""^^xsd:dateTime ;
  crm:P82b_end_of_the_end ""^^xsd:dateTime .
```

Listing 2 Turtle representation of the semantic paths used for the documentation of a birthplace and birthdate. Prefixes used are listed on [prefix.cc](#).

The mappings produced at this stage were used to guide the data transformation phase. For such a task, SARI has utilised the X3ML Framework and Engine Mapper ([Marketakis et al., 2017](#)). Nevertheless, other RML-based mappers, such as YARRRML ([Van Assche et al., 2021](#)) and MORPH-KG ([Arenas-Guerrero et al., 2024](#)), can also be employed, as the methodology is independent of the specific mapping language used.

5.1.6 Step 6. Organise the typical properties under common categories

A major part of the functionality that an SRDM aims to enable is the understandability of the model between the domain user and knowledge engineer at a common level. For this reason the presentation and ordering of properties within a model such that they group together in epistemic categories that hang together for the end user are important. Thus we see for example here that the fields for the lifespan of a person are clustered together for easy reference by all users ([Figure 5](#)).

identifier	Name	Description	CRM Path	used by
person_73_ birth_date_ _earliest	Birth Date - Earliest	This field is used to record the earliest possible date for the birth of the documented person.	→ P98i → E67[73_1] → P4 → E52[73_2] → P82a → xsd:dateTime	Swiss Art Research Infrastructure
person_74_ birth_date_ _latest	Birth Date - Latest	This field is used to indicate the latest known date for the birth of the documented person.	→ P98i → E67[73_1] → P4 → E52[73_2] → P82b → xsd:dateTime	Swiss Art Research Infrastructure
person_75_ birth_locati on	Birth Location	This field is used to indicate the place of birth of the documented person.	→ P98i → E67[73_1] → P7 → E53[75_1]	Swiss Art Research Infrastructure
person_76_ death_date_ _earliest	Death Date - Earliest	This field is used to indicate the earliest known date for the death of the documented person.	→ P100i → E69[76_1] → P4 → E52[76_2] → P82a → xsd:dateTime	Swiss Art Research Infrastructure
person_77_ death_date_ _latest	Death Date - Latest	This field is used to indicate the latest known date for the death of the documented person.	→ P100i → E69[76_1] → P4 → E52[76_2] → P82b → xsd:dateTime	Swiss Art Research Infrastructure
person_78_ death_locati on	Death Location	This field is used to indicate the place of death of the documented person.	→ P100i → E69[76_1] → P7 → E53[78_1]	Swiss Art Research Infrastructure

Figure 5 Example of the cluster of properties grouped under the category 'Existence' within the SRDM of a Person.

5.1.7 Step 7. Evaluate the adequacy of the model

The adequacy of the model was then tested to determine its capacity to accurately express semantically the scope of information recorded for persons. Adequacy was tested both by a review of the input data to the modelling process, to verify that fields in the source data structures can be translated, and through an iterative process of engagement with domain specialists. During the elaboration of the Person model for SARI we determined that three identified fields could not be expressed using CIDOC-CRM: the language(s) known by the documented person, the occupations they held, and the different social relationships they established over the course of their lives. As no extensions of CRM, or compatible ontologies, were found, we created a small-ontology to cover the project's needs. We note that at such decision points, the designers of an SRDM model will have to choose between (i) the creation of a new property or class, declaring it as an extension of the chosen ontology, or (ii) adopting pre-defined properties from an extant ontology in a manner that manages to cover the missing property, although perhaps in an unintended manner. Both options should be undertaken with care, as they potentially limit future interoperability alongside any potential gains derived in the present.

6 CONCLUSION

In this paper we have proposed a documentation protocol to support the production of semantic data modelling documentation by building a bridge between existing data management practices and semantic data-based model. This method proposed was illustrated through the use-case of the Swiss Art Research Infrastructure (SARI) project's creation of a set of reference data models for the domain of art and architectural history using the CIDOC-CRM ontology, particularly the SRDM Person.

FUNDING STATEMENT

This work has been funded by the Swiss Art Research Infrastructure (SARI).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

George Bruseker: Conceptualisation, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Nicola Carboni: Conceptualisation, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Matthew Fielding: Writing – original draft, Writing – review & editing, Validation.


Denitsa Nenova: Investigation, Validation.

Thomas Hänsli: Funding acquisition, Project administration.


AUTHOR AFFILIATIONS

George Bruseker  orcid.org/0000-0001-7519-1970
Takin Solutions Ltd., Plovdiv, Bulgaria

Nicola Carboni  orcid.org/0000-0003-4912-4947
School of Information Sciences University of Illinois Urbana-Champaign, USA

Matthew Fielding  orcid.org/0009-0001-5543-1372
Takin Solutions Ltd., Plovdiv, Bulgaria

Denitsa Nenova  orcid.org/0000-0003-3138-1689
Takin Solutions Ltd., Plovdiv, Bulgaria

Thomas Hänsli  orcid.org/0000-0001-7818-5605
Institute for the History and Theory of Architecture, ETH, Zurich, Switzerland; Swiss Art Research Infrastructure, University of Zurich, Zurich, Switzerland

- Arenas-Guerrero, J., Chaves-Fraga, D., Toledo, J., Pérez, M. S., & Corcho, O. (2024). Morph-KGC: Scalable knowledge graph materialization with mapping partitions. *Semantic Web*, 15(1), 1–20. <https://doi.org/10.3233/SW-223135>
- Borgo, S., Galton, A., & Kutz, O. (2022). Foundational Ontologies in Action. *Applied Ontology*, 17(1), 1–16. <https://doi.org/10.3233/AO-220265>
- Bruseker, G., Carboni, N., & Guillem, A. (2017). Cultural Heritage Data Management: The Role of Formal Ontology and CIDOC CRM. In *Heritage and Archaeology in the Digital Age* (pp. 93–131). Cham: Springer. https://doi.org/10.1007/978-3-319-65370-9_6
- Crofts, N., Doerr, M., & Gill, T. (2003). The CIDOC Conceptual Reference Model: A standard for communicating cultural contents. *Cultivate Interactive*, 9.
- Enriquez, A. L., Myers, D., & Dalgity, A. (2018). The Arches Heritage Inventory and Management System for the Protection of Cultural Resources. *Forum Journal*, 32(1), 30–38. <https://doi.org/10.1353/fmj.2018.0004>
- Gangemi, A., & Presutti, V. (2009). Ontology Design Patterns. In *Handbook on Ontologies*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-92673-3_10
- Guarino, N., Oberle, D., & Staab, S. (2009). What Is an Ontology? In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 1–17). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-92673-3_0
- Hammar, K., Blomqvist, E., Carral, D., van Erp, M., Fokkens, A., Gangemi, A., ... Svatek, V. (2016). Collected Research Questions Concerning Ontology Design Patterns. In *Ontology Engineering with Ontology Design Patterns* (pp. 189–198). IOS Press. <https://doi.org/10.3233/978-1-61499-676-7-189>
- Hitzler, P. (2021). A Review of the Semantic Web Field. *Communications of the ACM*, 64(2), 76–83. <https://doi.org/10.1145/3397512>
- Lozano-Tello, A., & Gomez-Perez, A. (2004). ONTOMETRIC: A Method to Choose the Appropriate Ontology. *Journal of Database Management*, 15(2). <https://doi.org/10.4018/jdm.2004040101>
- Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., ... Doerr, M. (2017, November). X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, 18(4), 301–319. <https://doi.org/10.1007/s00799-016-0179-1>
- Oldman, D., & Tanase, D. (2018). Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. In D. Vrandečić et al. (Eds.), *The Semantic Web – ISWC 2018* (pp. 325–340). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-00668-6_20
- Ristoski, P., & Paulheim, H. (2016, January). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics*, 36, 1–22. <https://doi.org/10.1016/j.websem.2016.01.001>
- Scholz, M., & Goerz, G. (2012). WissKI: A Virtual Research Environment for Cultural Heritage. In *ECAI 2012* (pp. 1017–1018). Amsterdam: IOS Press. <https://doi.org/10.3233/978-1-61499-098-7-1017>
- Trojahn, C., Vieira, R., Schmidt, D., Pease, A., & Guizzardi, G. (2022, January). Foundational ontologies meet ontology matching: A survey. *Semantic Web*, 13(4), 685–704. <https://doi.org/10.3233/SW-210447>
- Van Assche, D., Delva, T., Heyvaert, P., De Meester, B., & Dimou, A. (2021). Towards a more human-friendly knowledge graph generation & publication. In *International semantic web conference (ISWC) 2021: Posters, demos, and industry tracks*.
- Westerinen, A., & Tauber, R. (2017, January). Ontology development by domain experts (without using the “O” word). *Applied Ontology*, 12(3–4), 299–311. <https://doi.org/10.3233/AO-170183>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016, March). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. <https://doi.org/10.1038/sdata.2016.18>
- Zeng, M. L. (2008). Knowledge Organization Systems (KOS). *Knowledge Organization*, 35(2–3), 160–182. <https://doi.org/10.5771/0943-7444-2008-2-3-160>

TO CITE THIS ARTICLE:

Bruseker, G., Carboni, N., Fielding, M., Nenova, D., & Hänsli, T. (2025). The Semantic Reference Data Modelling Method: Creating Understandable, Reusable and Sustainable Semantic Data Models. *Journal of Open Humanities Data*, 11: 22, pp. 1–14. DOI: <https://doi.org/10.5334/johd.282>

Submitted: 14 November 2024

Accepted: 20 January 2025

Published: 12 March 2025

COPYRIGHT:

© 2025 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.