Polyvocal Knowledge Modelling for Ethnographic Heritage Object Provenance

Anonymous AUTHOR [a,1]
ORCiD ID: Anonymous Author https://orcid.org/0000-0000-0000-0000

Abstract. Purpose: Information about biographies of museum object objects (*object provenance*) is often unavailable in a machine-readable format. This limits the findability and reusability of object provenance information for domain research. We address the challenges of defining a data model to represent ethnographic cultural heritage objects' provenance, which includes multiple interpretations (*polyvocality*) of, and theories for, the object biography, chains of custody and context of acquiring.
Methodology: To develop a data model for representing the provenance of ethnographic objects, we conducted (semi-)structured interviews with five provenance experts to elicit a set of requirements. Based on these requirements and a careful examination of six diverse examples of ethnographic object provenance reports, we established a set of modelling choices that utilise existing ontologies such as CIDOC-CRM (a domain standard) and PROV-DM, as well as RDF-named graphs..
Evaluation: Finally, we validate the model on provenance reports containing six seen and five unseen ethnographic cultural heritage object from three separate sources. The 11 reports are converted into RDF triples following the proposed data model. We also constructed SPARQL queries corresponding to nine competency questions elicited from domain experts in order to report on satisfiability.
Findings: The results show that the adapted combined model allows us to express the heterogeneity and polyvocality of the object provenance information, trace data provenance and link with other data sources for further enrichment.
Value: The proposed model from this paper allows publishing such knowledge in a machine-readable format, which will foster information contextualisation, findability and reusability.

Keywords. Cultural Heritage, Provenance information, Polyvocality, Domain modelling, Knowledge Representation.

1. Introduction

More and more cultural heritage organisations are producing and publishing their data as Knowledge Graphs [1]. One of the reasons for adopting these technologies in this domain is that the graph structure allows to express heterogeneity of information [2], while facilitating interoperability. However, producing machine-readable knowledge graphs from existing structured data and unstructured sources is not a trivial operation [3]. At the same time, museum professionals and external researchers continuously acquire new information about collection objects and existing information is then cast in a new light.

[1] Corresponding Author: anon@anon.an

Especially in the context of post-colonial challenges to ”decolonise the database”, heritage institutions are seeking out how to incorporate previously underrepresented voices in their practice, collections and information systems (cf. [4]).

Cultural heritage institutions, especially those with ethnographic collections, continuously (re)contextualise objects by learning new facts about objects’ biographies [5]. The traditional method is through dedicated research on individual objects, known as *provenance research*. Cultural heritage object provenance describes an object’s history of ownership and evidence of the legal status of an object [6]. It can also be used to form an assessment of the authenticity of an artefact and identify any unlawfully appropriated works [7]. In many cases, the details of this extensive research are not available as structured data but only reported in a narrative textual document, which limits the findability, reusability and interoperability of such information. Typically, the person who conducts the research or the institute they are representing reflects on the object metadata [8]. Once the research is done, the researcher might update a limited amount of metadata in the collection management system. In this process, potentially multiple views are reduced to a single perspective of truth. For example, the same object can be annotated with term “war loot” or “legally owned property”, depending on the researcher’s interpretation of the documentation on the war. Such an interpretation can be affected by personal, cultural or political context and is likely to change over time. More importantly, when more than one annotation has equal merit to be valid, it is necessary to preserve both interpretations in the metadata. To allow for future researchers and professionals to investigate these multiple perspectives, the institutions’ information systems will need to be able to preserve, maintain and deliver the different views of objects and their provenance data [9].

The Semantic Web as an information architecture and Knowledge Graphs as the data model are promising technologies of such polyvocal knowledge representations [9]. Its dispersed and networked nature makes it ideally suited to handle diverse opinions, while at the same time preserving competing views with the sources of their origin. We consider this research on representing multiple perspectives as an example of such polyvocal knowledge representation. We investigate to which extent existing Semantic Web solutions, such as named graphs and existing ontologies for provenance and heritage, are suitable to represent multiple perspectives in data.

This paper’s contribution lies in examining how Semantic Web technologies can be used to meet polyvocal provenance requirements specified by domain experts. We first identify these requirements for representing multi-perspective ethnographic object provenance information. We propose how such information can be modelled and apply this model to eleven ethnographic objects demonstrating the expression of complex chains of custody of the object biographies while preserving data provenance. The solution preserves the polyvocality of such information when multiple alternate theories are available. The resulting knowledge graph is validated against the Competency Questions constructed from the requirement analysis.

2. Related Work

High-quality metadata is necessary to increase the accessibility and reusability of digital content. The metadata of a museum object must include detail about the object before it enters a museum, as well as details that are generated while the object is in the museum [10]. When modelling cultural heritage data, it must be represented in a usable way for non-technical users, such as cultural heritage experts, to query, review and reuse it. There has been extensive research done on modelling of cultural heritage metadata [1,3,11,12, 13]. However, none of these works investigated how to model rich object provenance information, which typically ends up as textual reports only.

Both object-centric and event-centric ontologies have been developed to represent cultural heritage metadata. Research, however, found that an event-centric approach provides advantages for representing provenance or other temporal data [1,14]. The event-centric model represents knowledge through associated events, such as acquisition or production. An ISO standard since 2006, CIDOC-CRM [15,16] is an event-centric ontology which is designed for the cultural heritage sector to facilitate the integration and interchange. CIDOC-CRM can be used to model multiple instances of semantic information regarding a given reality by adding multiple information layers. However, research [17] has shown that by itself this is not an feasible solution for representing multi-perspective data as these multiple layers are simply information accumulation without mentioning data provenance. The authors argue that the data must be organised so researchers can easily find previous information and use it for new reasoning.

Conversations around multiple perspectives are taking place in the cultural heritage domain [18,19]. Dijkshoorn et al. [1] present six requirements for cultural heritage ontologies, one of these supports capturing multiple sources with possibly conflicting views while describing the same artefact. In their research, it has been shown that the Europeana Data Model [20] allows multiple records for the same object by using proxies. Proxies in EDM can, however, only depict objects on a general level by connecting a proxy to the object resource and not to a specific statement about that resource. A similar approach is adopted by Ockeloen et al. [21], who propose a proxy solution for representing biographical descriptions from different perspectives and sources.

Another solution for multi-perspective representation can be found using named graphs. Bizer et al. [22] state that information providers have different world views; therefore, a named graph allows different information providers to make different claims regarding the same entity. The advantage of named graphs is that it allows grouping a collection of triples to make statements on the whole set and can quickly be adopted when CIDOC-CRM is implemented in RDF. Having IRIs on the named graphs introduces the possibility of attaching data provenance to the graph itself.

While the need for multi-perspective representations of cultural heritage data is identified, the practical application is still challenging. This research identifies a possible solution for representing multi-perspective interpretations of cultural heritage object provenance that is based on the domain standards discussed above.

3. Requirement Analysis

This section describes the requirement analysis for representing multi-perspective representations of cultural heritage provenance. A more detailed account of this analysis is found in [23][2]We here present the main approach and the resulting list of requirements and competency questions.

[2] name and citation removed for anonymity

Table 1. Overview of museology expert interviewees

Respondent	Role	Expertise
R1	Postdoctoral researcher	Objects from East Africa
R2	Junior provenance researcher	Object combined with human remains
R3	Senior provenance researcher	Objects from Central and Southern Africa
R4	Postdoctoral researcher	Objects collected in Missionary context
R5	Senior provenance researcher	Objects from Asia

3.1. Approach

To collect data requirements for multi-perspective representations of cultural heritage provenance, we conduct a problem analysis through focused interviews with domain experts, which is concerned with developing an understanding of the nature of the problem. Focused interviews are a basic requirement engineering tool, to investigate current problems and concerns. After identifying the requirements for the data model, we utilised them to construct the model (Section 4). Additionally, we elicited nine Competency Questions from the interview which we use to validate the model.

For the focused interviews, we recruited five Museology professionals who are involved in the Pressing Matter project[3] in different capacities (see Table 1 for an overview of the interviewees). Pressing Matter is a Dutch project which investigates artefacts collected during the colonial period to support societal reconciliation with the colonial past. The professionals were chosen based on their varied experience, background, and working methods. Although they work on different collections of ethnographic objects, they all have experience with the current museum information system and are responsible for updating object metadata with provenance information. These professionals can be considered the end-users of the data model developed in this research.

Each participant completed a one-hour individual semi-structured interview, with all interviews following the same interview guide. The interview guide [24] is aligned with the objective of this research[4]. The interview addressed the proper representation of cultural heritage provenance data, covering (1) provenance research processes and challenges, (2) documentation of research, (3) representation of provenance information, and (4) the utility of such information. A pilot interview was conducted before the actual study, and its insights were incorporated in the next interviews. All interviews, except the pilot, were conducted via web conferencing.

3.2. Findings

We report on the main findings in three parts. First, how provenance research is conducted and documented. Second, the identified challenges and problems with current representations are presented. Third, the respondents' opinions on multi-perspective representations of cultural heritage provenance.

Provenance Research: Interview results suggest that there is no standard goal for provenance research. However, all respondents agreed that it helps in gaining a better understanding of where collections and objects come from, leading to better-documented collections. Respondents had different approaches to conducting research, with varying

[3] `https://pressingmatter.nl/`

[4] The interview guide can be found at `https://doi.org/10.5281/zenodo.7437713`

reliance on sources such as archives, libraries, and web searches. All respondents, however, begin their research from the museum's collection management system (CMS)[5]. Important to note that, two mentioned that the system often contains missing information and observational bias. Respondents also shared that there are no guidelines on how to represent provenance information. Typically, when the information goes beyond the CMS, a separate report is written. However, there is no efficient way to trace or find such information within the system except for the unstructured text report.

Problem with current representations: Participants agreed that current information representation in the CMS is problematic due to faulty, incomplete or unreliable information. Lack of digitisation of archival material (R5) and decentralisation of available materials for information (R3) are identified as major challenges. Current representations of provenance in the CMS do not match the complexity of provenance research (R4), and important relations among people, places, objects and event cannot be represented in a machine-readable way (R3). Another problem identified by all the respondents is that the current management system does not contain any data provenance information, making it difficult to trace provenance of statements previously made about an object.

Opinions on multi-perspective representation cultural heritage data: Respondents agreed that keeping nuance in object provenance information is important, as changing times and perspectives lead to new ways of perceiving information. Museum database records can be influenced by the dominant perspective of their time, such as colonial representations and language use, which may not align with current views (R2, R4). Acknowledging how objects were seen before can tell us something about collections (R4). If multiple versions of provenance exist, all should be represented, as provenance is rarely fully proven (R5). On the other hands, some respondents (R1, R3) argue that it's not practical to preserve all information, and it depends on the research goals. Another respondent (R2) notes the importance of distinguishing between information deemed more correct now versus prior research. They also agreed that their research is just one interpretation of an object's history; it is impossible to say that their research is the final interpretation of the history of an object.

3.3. Identified Requirements and Competency Questions

The overall requirements for a representation are presented in the list below, divided into three types: overall representation, object provenance (information identified as important related to the chain of custody of an object) and data provenance (about the cultural heritage provenance statements, such as sources used during the research).

Overall representation

- *Digital representation:* Provenance research is easier to conduct if cultural heritage data is digitally preserved.
- *Event-centeric representation:* It is easier to identify relations between actors, objects and places when they are represented in an event-centric way.
- *Machine-readable:* Machine-readable representations are easier to access compared to when the provenance research is presented in a written report.

Object provenance

[5] In this case, TMS `https://www.gallerysystems.com/solutions/collections-management/`

- *Object data:* Representation of object title, object number, object category, material, part of which collection and its origin.
- *Object creation:* When, where, and by whom the object was created.
- *Actors:* The network of actors involved in the object's collection.
- *Locations:* Object acquisition and creation places as well as travel route.
- *Events and time periods:* Historical events or time-period, may provide context, including unethical acquisition of colonial objects.
- *Multiple descriptions:* If multiple views on an acquisition exist they should be noted to keep nuance.
- *Comments:* An event may need detailed comments and notes in natural language.

Data provenance

- *Provenance statement source:* Users should be able to review the sources/author for provenance statements.
- *Source:* Users should be able to find the source materials of object provenance.
- *Traceability to previous research:* Each version of object provenance research should include data provenance information.

3.3.1. Competency Questions

Competency questions (CQs) are questions in natural language that outline the knowledge and specify the constraints for knowledge representation [25]. The concept of competency questions was explained during the interviews, and the respondents were requested to come up with specific questions based on their own requirements. The individual CQs were then aggregated in this study. All participants agreed that it is crucial to keep track of the *people* who were engaged in object acquisition (collectors, traders...) to identify networks of individuals involved in the acquisition of an object (**CQ1, CQ2, CQ3**). It is also important to convey information about *dates and events*; for example, the date or occation the object was obtained (**CQ4**). This enables identifying networks of connected objects through historical events, which may collectively project on an objects' acquisition (**CQ5**). *Geographic locations* are important to determine which items were bought or sold in particular regions or countries (**CQ6**). The respondents also identified *data provenance* as a crucial part in their competency questions. The participants unanimously agreed that each claim about the objects' provenance must be documented to track previous studies (**CQ7**, **CQ8**, and **CQ9**). They also mentioned the need to revisit earlier provenance versions to acquire a complete picture of all previous studies. The full list of aggregated competency questions from the text above is shown in the first two columns of Table 2. In Section 5.2, we describe how these are used for the purpose of validation.

4. Data Model

To further guide the modelling of object provenance information, we investigate this in the form of a case study concerning six ethnographic objects that are described in two different provenance reports "Provenance #1"[26] and "Provenance #2" [27], issued by the Dutch National Museum of World Cultures (NMVW)[6] to embed provenance of art-

[6] http://wereldculturen.nl/

Table 2. Competency questions (Section 3.3) and corresponding SPARQL queries (Section 5.2)

ID	Question	SPARQL query	Answers CQ?

CQ1	Which persons were involved in the provenance of this object?	SELECT * WHERE {
		?o a crm:E24_Physical_Human-Made_Thing .
		?o crm:P49_has_former_or_current_keeper ?p.
		?p rdfs:label ?lab}
			Yes, demonstrated query answers if the intent is to find out actors involved in object biography as a formal keeper or owner.

CQ2	Which objects are collected by person A?	SELECT * WHERE { ?p a crm:E39_Actor . {?act crm:P29_custody_received_by ?p. ?act crm:P30_transferred_custody_of ?o} UNION {?act crm:P23_transferred_title_from ?p. ?act crm:P24_transferred_title_of ?o}}	**Yes**, query retrieves all objects ?o if associated with actor ?p through any collection activity.
CQ3	Is there a relationship between person A and person B?	SELECT ?p1 ?p2 WHERE { ?p1 a crm:E39_Actor . ?p2 a crm:E39_Actor . ?act1 a crm:E7_Activity. ?act1 ?prop1 ?p1. ?act1 ?prop2 ?p2. FILTER (?p1 != ?p2)}	**Yes**, query demonstrates retrieval of two persons, involved through a shared activity with the same object.

CQ4	Which objects were collected in this geographical location?	SELECT ?o ?p WHERE{
		?o a crm:E24_Physical_Human-Made_Thing .
		{?act crm:P30_transferred_custody_of ?o}
		UNION
		{?act crm:P24_transferred_title_of ?o}
		?act crm:P9_consists_of ?sub .
		?sub crm:P7_took_place_at ?p. }
			Yes, query demonstrates how to retrieve object with location when location

CQ5	Which objects were collected during this event?	SELECT DISTINCT ?obj ?event WHERE { ?event a crm:E5_Event . ?event crm:P9_consists_of ?act . ?m_act crm:P9_consists_of ?act . {?m_act a crm:E8_Acquisition .} UNION {?m_act a crm:E10_Transfer_of_Custody .} ?m_act ?p ?obj . ?obj a crm:E24_Physical_Human-Made_Thing. }	**Yes**, given historical event ?e, this query returns all the objects whose collection activity is relevant to this event
CQ6	Which objects were collected in this geographical location during this time period?	SELECT ?o ?p ?b_time ?e_time WHERE{ ?o a crm:E24_Physical_Human-Made_Thing . ?act crm:P9_consists_of ?sub . ?sub crm:P7_took_place_at ?p. ?sub crm:P4_has_time-span ?time . ?t crm:P82a_begin_of_the_begin ?b_time . ?t crm:P82b_end_of_the_end ?e_time. {?act crm:P30_transferred_custody_of ?o} UNION {?act crm:P24_transferred_title_of ?o}}	**Yes**, the query returns all the objects with known geographic location and time-period

CQ7	Which source states this statement?	SELECT * WHERE{
		Graph ?g {?s ?p ?o . }
		?g (a—!a)+ ?g_o .
		?g_o a prov:Entity .
			Yes, the query returns all statements with their associated named graph and data provenance for the named graph

CQ8	Who or which institution conducted this research?	SELECT ?r ?a1 ?a2 WHERE { ?r a prov:Activity . ?r prov:wasAssociatedWith ?a1 . OPTIONAL {? prov:actedOnBehalfOf ?a2}}	**Yes**, given an research activity, the query returns agents or institution
CQ9	Which is the latest version of the provenance research?	SELECT ?act ?date WHERE { Graph ?g {?s ?p ?o .} ?g prov:wasDerivedFrom ?entity . ?entity prov:wasGeneratedBy ?act . ?act prov:endedAtTime ?date . filter not exists { ?act prov:endedAtTime ?date1 filter (?date ?date1) } } ORDER BY DESC(?date)	**Yes**, given a triple the query returns provenance report associated with it in descending publishing order.

works into its practice and policy. They describe objects with rich provenance information elicited by extensive provenance research on these objects.

One object, RV-1148-1 is an elephant tusk that was part of the extensive provenance research on collections from Benin City. This provenance research aimed to assess the strength of connection to the military campaign led by British forces against Benin City in early February 1897; the research can be found in the report “Provenance #2” [27]. The rest of the objects (RV-2584-169a, RV-2334-1, RV-2334-2, RV-2334-3 and RV-2334-1) are from “Provenance #1” [26], where two different types of provenance information are found. On one side, RV-2584-169a is an interesting case-study because of having different possible theories of the acquisition and/or origin of the object, which left the researcher incapable of concluding on one theory with a high degree of certainty. On the other hand, the objects with id RV-2334-* are insightful because of their complicated chain of custody and links to different archival material. Despite the diversity in information, what is common in all the objects is that there have been discovered possible links to important historical events or time-periods, which projects unethical acquisition in the objects’ chain of custody. The diversity in information, possible links to different historical events and time-periods, and available connections to different archival materials make these objects an ideal case study for the current research. These objects not only refer to the requirements identified by the previous section but also represent the complex nature of such knowledge. In the following subsection, we will describe our modelling choice for the proposed data model.

The first step is to reconstruct the information of these objects to identify essential statements related to the object’s provenance from the textual report. This information is translated to provenance data that illustrates the key components considered necessary for representing the artefact’s provenance. For this research, we investigated the use of CIDOC-CRM to ensure reusability.

4.1. Object

Our requirements indicate there are various types of knowledge about the object to represent. First-level object information is essential even if detailed provenance is unknown. As one respondent (R1) mentioned, the origin can be identified by knowing the materials used or the creation or collection date. CIDOC-CRM allows an object- and event-centric approach to co-exist by connecting instances directly to an object and also without an intermediary event. The domain experts call for representations of objects’ inventory number, title, category or classification, material, collection and origin. Additionally, they request the possibility of attaching comments and descriptive notes.

Based on the participants’ requirement of object description, this research reuses the specification mentioned for the function to express object collection information by CIDOC-CRM guideline[7], with adjustments based on use-case requirements. The collection object itself is an instance of **E24 Physical Human-Made Thing**[8], with the following properties (and object classes).

- Inventory Id & Title: *P1 is identified by* (**E41 Linguistic Appellation** & **E42 Identifier**)
- Object classification: *P2 has type* (**E55 Type**)
- Textual Description: *P3 has note* (Literal)

[7] `https://www.cidoc-crm.org/FunctionalUnits/object-collection-information`

[8] In the following, we list ontology classes in **bold** typeface and properties in *italics*

- Related Person or Organization: *P49 has former or current keeper*, *P52 has current owner* (**E39 Actor**)
- Object Dimension: *P43 has dimension* (**E54 Dimension**)
- Material: *P45 consists of* (**E57 Material**)
- Represented Visual Concept: *P65 shows visual item* (**E36 Visual Item**)
- Image: *P138i has representation* (IRI)
- Collective Name/ Group: *P67i is referred to by* (**E33 Linguistic Object**)

Additional object information is best represented through events, i.e., **E7 Activity**, **E8 Acquisition**, **E12 Production** or **E10 Transfer of Custody**. Therefore, such information is connected with representative activities that are themselves connected to the artefact. For example, *P14 carried out by*, *P7 took place at*, and *P4 has time-span* are properties of the **E12 Production** event, where the object connects this activity with property *P108i was produced by*. On the other hand, properties such as *P28 custody surrendered by*, *P29 custody received by*, and *P30 transferred custody of* are not mentioned due to the complexity of such information and are modelled as part of provenance information.

4.2. Provenance Information

In addition to representing basic object information, the domain experts desired representations of detailed provenance information with known actors, locations and events. An object's provenance can be seen as a series of events where the custody of an object is transferred between different actors during time and places. The provenance of an object is mainly represented using two different entities in CIDOC-CRM. **E8 Acquisition** comprises the transfer of legal ownership from one or more instances of **E39 Actor** to another. In contrast, **E8 Acquisition** refers to legal ownership, thus the view that the change of owners is interpreted as a legal right, for example, object is purchased.

Common to all six objects considered here, there is at least one transfer of custody in their biography that is not seen as a legal right, namely when it was looted during the military campaign, purchased from illegal authority or receiving questionable gift. Therefore, using E8 Acquisition for modelling unethical ways of acquisition where the legal right is questioned may not be a appropriate. CIDOC-CRM separates legal ownership and physical custody. **E10 Transfer of custody** can be used to represent non-legal ways of acquisition, where a specific type of acquisition, such as theft, loot or gift, can be declared.

For the current modelling choice of selected objects, we used **E10 Transfer of Custody** to represent any illegal transfer of ownership and **E8 Acquisition** for legal cases. Any such activity (both E8 and E10) can further contain other activity(-ies) as the sub-activity(-ies) falls within the space-time volume of the main activity. This sub-activity is an instance of **E7 Activity**, where the type of activity is further mentioned with *P2 has type* property and itself being connected with main activity with *P9 consists of* property. Consider, an object acquisition that occurred as a result of a government representative receiving a diplomatic gift and later transferring it to the museum. In that case, the acquisition of the collection consists of the activity of "receiving gift". This distinction between the main activity and "typed" sub-activity is made to achieve a level of abstraction across all objects, even when the transfer method of ownership is unknown.

Common to both E8 and E10 there is possibility to include a time-span, a location to the event and actors involved. Since both of them are sub-class of **E2 Temporal Entity**,

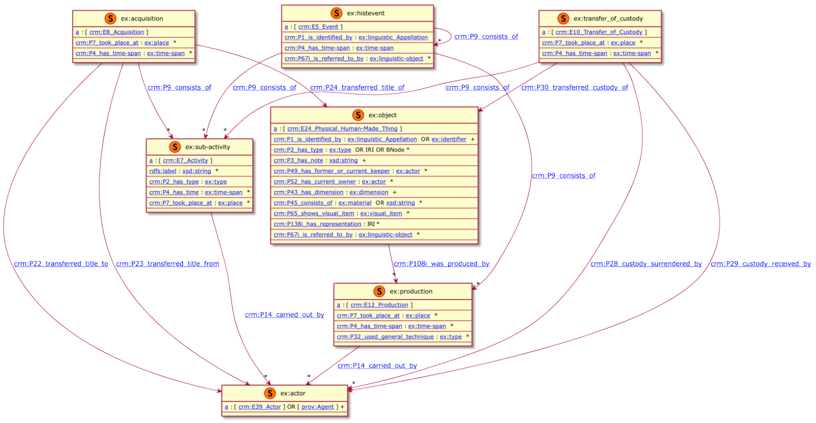

ex:acquisition
a : [crm:E8_Acquisition]
crm:P7_took_place_at : ex:place *
crm:P4_has_time-span : ex:time-span *
ex:histevent
a : [crm:E5_Event]
crm:P1_is_identified_by : ex:linguistic_Appellation
crm:P4_has_time-span : ex:time-span
crm:P67i_is_referred_to_by : ex:linguistic-object *
ex:transfer_of_custody
a : [crm:E10_Transfer_of_Custody]
crm:P7_took_place_at : ex:place *
crm:P4_has_time-span : ex:time-span *
crm:P9_consists_of
crm:P9_consists_of
crm:P24_transferred_title_of
crm:P9_consists_of
crm:P30_transferred_custody_of
crm:P9_consists_of
ex:sub-activity
a : [crm:E7_Activity]
rdfs:label : xsd:string *
crm:P2_has_type : ex:type
crm:P4_has_time : ex:time-span *
crm:P7_took_place_at : ex:place *
ex:object
a : [crm:E24_Physical_Human-Made_Thing]
crm:P1_is_identified_by : ex:linguistic_Appellation OR ex:identifier +
crm:P2_has_type : ex:type OR IRI OR BNode *
crm:P3_has_note : xsd:string +
crm:P49_has_former_or_current_keeper : ex:actor *
crm:P52_has_current_owner : ex:actor *
crm:P43_has_dimension : ex:dimension +
crm:P45_consists_of : ex:material OR xsd:string *
crm:P65_shows_visual_item : ex:visual_item *
crm:P138i_has_representation : IRI *
crm:P67i_is_referred_to_by : ex:linguistic-object *
crm:P9_consists_of
crm:P22_transferred_title_to
crm:P23_transferred_title_from
crm:P108i_was_produced_by
crm:P28_custody_surrendered_by
crm:P29_custody_received_by
crm:P14_carried_out_by
ex:production
a : [crm:E12_Production]
crm:P7_took_place_at : ex:place *
crm:P4_has_time-span : ex:time-span *
crm:P32_used_general_technique : ex:type *
crm:P14_carried_out_by
ex:actor
a : [crm:E39_Actor] OR [prov:Agent] +

Figure 1. Ontology overview of cultural heritage object provenance modelling

time specification can be mentioned by *P4 has time-span* property. As subclasses of **E4 Period**, they can have *P7 took place* at properties with a **E53 Place** instance as object. For instances of **E8 Acquisition**, the properties *P23 transferred title from*, *P22 transferred title to* and *P24 transferred title of* are used to connect actors and objects to the activity. Fo instances of **E10 Transfer of Custody** the properties *P28 custody surrendered by*, *P29 custody received by* and *P30 transferred custody of* play that role.

Figure 1 visualizes the main entities of the ontology created, generated using the RDFShape[9] visualizer. This diagram only specifies the shape for the primary entities, less significant entities' detail is left out here for visual simplicity.

4.3. Data Provenance

Besides the *object provenance* information, the domain experts also requested representation of *data provenance*. The representation of the data provenance is required for the traceability of research. First, it concerns sources linked to each claim regarding the objects' provenance. Secondly, it relates to data provenance regarding the provenance research itself, including details on who did the study, for which institution, and when it was done.

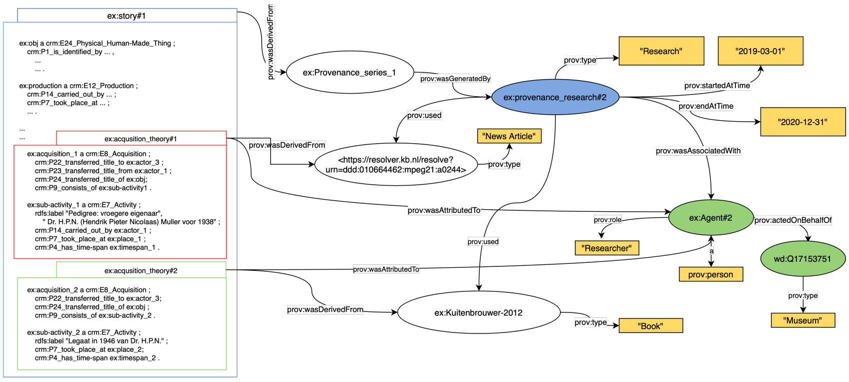
Our solution for such representations is to use named graphs in combination with the PROV-DM [28] ontology. Named graphs can be used to attach provenance data and model context and scope assertions[10]. This provides the capacity to assess various assertions made in a graph by the information providers [22]. In this case, a provenance researcher can identify a single source of knowledge containing various statements. It is up to information consumers to decide whether or not they can trust the information provider and how reliable the information is.

PROV-DM [28] is a conceptual model for modelling provenance, with PROV-O being a mapping to RDF[11] with proven applicability in the cultural heritage domain [21,29]. In our model it is used for attaching data provenance to a named graph IRI. A typical

[9] `https://rdfshape.weso.es/`

[10] cf. `https://www.w3.org/2009/07/NamedGraph.html`

[11] `https://www.w3.org/TR/prov-o/`



```
ex:obj a crm:E24_Physical_Human-Made_Thing ;
   crm:P1_is_identified_by ... ,
      ...
      ... .

ex:production a crm:E12_Production ;
   crm:P14_carried_out_by ... ;
   crm:P7_took_place_at ... ;
   ... .

...
...
```

```
ex:acquisition_1 a crm:E8_Acquisition ;
   crm:P22_transferred_title_to ex:actor_3 ;
   crm:P23_transferred_title_from ex:actor_1 ;
   crm:P24_transferred_title_of ex:obj;
   crm:P9_consists_of ex:sub-activity1 .

ex:sub-activity_1 a crm:E7_Activity ;
   rdfs:label "Pedigree: vroegere eigenaar",
      " Dr. H.P.N. (Hendrik Pieter Nicolaas) Muller voor 1938" ;
   crm:P14_carried_out_by ex:actor_1 ;
   crm:P7_took_place_at ex:place_1 ;
   crm:P4_has_time-span ex:timespan_1 .
```

```
ex:acquisition_2 a crm:E8_Acquisition ;
   crm:P22_transferred_title_to ex:actor_3;
   crm:P24_transferred_title_of ex:obj ;
   crm:P9_consists_of ex:sub-activity_2 .

ex:sub-activity_2 a crm:E7_Activity ;
   rdfs:label "Legaat in 1946 van Dr. H.P.N." ;
   crm:P7_took_place_at ex:place_2;
   crm:P4_has_time-span ex:timespan_2 .
```

Figure 2. Multiple views or theories of Object acquisition are separated into named graphs and tagged with sources following the PROV-DM ontology. The transparent elliptic shapes represent **prov:Entity**, blue elliptical shapes represent **prov:Activity**, and green elliptical shapes represent **prov:Agent**.

use case for PROV-DM is to achieve data quality, traceability and trustworthiness. The **Entity**, **Activity** and **Agent** classes are the building blocks for the model. According to the model, any physical, digital or conceptual things can a **prov:Entity**, where an **prov:Activity** is any action that occurs over a time-period, and **prov:Agent** is any actor who is responsible for the action. Therefore, in our case, the named graph containing all triples from the provenance research is of type **prov:Entity**, the provenance research activity itself is of type **prov:Activity**, and the institution or the person who are involved with this research is a **prov:Agent**. Representing provenance research and derived statements is one example of how we modelled different statements generated from different resources and activities. In Figure 2, the named graph ex:story#1 includes all data triples associated with object RV-2584-169a, which were generated during the research activity ex:provenance_research#2. This activity is an instance of **prov:Activity** in accordance with the PROV-DM ontology, enabling data provenance tracing for the generated triples.

4.4. Polyvocal Modelling

Data provenance modelling and named graphs can be used to group (CIDOC-CRM) triples that conform to a particular view of acquisition and distinguish them from other statements that conform to another view. Figure 2 shows how named graphs separate different (CIDOC-CRM) triples representing specific views of acquisition and how data provenance of such named graphs is specified using the PROV-DM ontology. When querying for a particular object provenance with SPARQL 1.1, triples stating object provenance can be returned without making any distinction in acquisition theory. By using the GRAPH keyword, information from only specified (provenance) graphs can be returned. This allows attaching a source, location and/or time period to that view and responsible agents/sources for a group of statements. Each named graph or triple collection is represented as a **prov:Entity**, typically derived from (*prov:wasDerivedFrom*) another **prov:Entity**(i.e., source) or generated by (*prov:wasGeneratedBy*) a **prov:Activity** (i.e., a domain research activity).

5. Results and Validation

This section reflects on the RDF triples generated from converting object provenance reports using the proposed model. Six objects were initially used for modeling decisions, and five more were randomly selected from the Pilotproject Provenance Research on Objects of the Colonial Era (PPROCE)[12] project to test the model's generalizability. The resulting knowledge graph was validated against competency questions to ensure it adheres to domain requirements. All relevant files can be found in the Zenodo repository[13].

5.1. Statics of Resulting knowledge graph

Following the modelling choices stated in the previous section and based on the information from the provenance reports, we first model the descriptions and provenance data of the six selected objects those were initially chosen to construct the model. For convenience, we are going to refer to these six objects as *construct object* in the rest of the paper. The resultant knowledge graph contains 1,786 triples spread across 31 named graphs. These named graphs contain either entire object metadata triples or triples generated by a single source of information. Additional named graphs were created to represent different views of the same entity with their source of information. The statistics of the resultant graph is given in the second column of Table 3. More statistics can be found in the Zenodo repository(/construct/entity_stat.csv) where the knowledge graphs itself is also available as TriG files in the data folder.

We also modeled the reports of 5 unseen objects from the PPROCE project , which we refer to as the ”evaluation objects.” The resulting knowledge graph contained 1,290 RDF triples spread across 27 named graphs, which is comparable to the construct objects set as we modeled only 5 objects' provenance data. The number of instances of different classes, such as **crm:E8 Acquisition**, **crm:10 Transfer of Custody**, **crm:E39 Actor**, **prov:Agent**, and **prov:Entity**, were consistent with the construct objects set. All TriG files and detailed statistics for this object set can be found in the ”evaluation” folder of the Zotero repository.

The conceptual model is converted to ShEx[14] rules to maintain data consistency and shape. ShapeMap queries are used to validate each entity against the proposed ontology using the RDFshape[15] web tool. ShEx rules and ShapeMap queries can be found in the repository.

5.2. Validation through Competency Questions

We validate the Knowledge Graphs using SPARQL queries to answer 9 competency questions provided by domain experts. The queries are listed in the third column of Table 2. Interpretation of the CQs and their corresponding SPARQL queries are discussed below.

For **CQ1**, the listed query only matches if the intent is to find formal keepers or owners involved in the object's biography, but an alternate query is required to determine

[12] https://www.niod.nl/en/projects/pilotproject-provenance-research-objects-colonial-era-pproce

[13] https://doi.org/10.5281/zenodo.7437713

[14] ShEx, shape expressions, see https://shex.io/

[15] https://rdfshape.weso.es

Entity	**Construct**	**Evaluation**
Named graphs	31	27
crm:E24_Physical Human-Made Thing	6	5
crm:E12_Production	7	5
crm:E10_Transfer of Custody	7	7
crm:E8_Acquisition	12	8
crm:E7_Activity	21	12
crm:E5_Event	3	1
crm:E39_Actor	27	14
crm:E52_Time-Span	57	50
crm:E53_Place	11	10
prov:Activity	2	5
prov:Agent	20	18
prov:Entity	38	51
Total triples	1778	1290

Table 3. Number of triples for both the initial six construct objects and the five additional objects used for validation

the exact capacity in which actors are involved in the object's provenance, as different activities are connected to objects with incoming and outgoing links. The alternate query is available in the supporting material.

CQ2 is interested in retrieving all objects that are connected with person A through a collection activity, i.e., **E8 Acquisition** or **E10 Transfer of Custody**. The query in Table 2 for **CQ2** retrieves both collection activities for an object and involved collectors; therefore answers the CQ accurately.

CQ3 can be answered in multiple ways. The listed query in the table retrieves two persons involved through a shared activity of the same object. If the intent is to find two actors linked with the same object, a separate query is needed (provided in supporting material). Nevertheless, computing all possible paths between two actors can be computationally expensive.

All activities can list its location using *crm:P7_took_place_at* properties. So, each collection activity is connected with the location if this information is known. The query given for **CQ4** targets to retrieve location when it is connected with sub-activity of Acquisition or Transfer of custody activity. Similar, queries can be written when the location is connected directly with activity. For detail, see Zenodo repository(validation_sparql.txt).

The query listed for **CQ4** retrieves location information when it is connected with a sub-activity of **E8 Acquisition** or **E10 Transfer of Custody** activity. Other queries can be written to retrieve location information when it is connected directly with the activity. More details can be found in the Zenodo repository under "validation_sparql.txt".

We answer **CQ5** by making multiple 'hops', since historical events are not directly connected with the objects, but rather can consist of activities concerning the object.

CQ6 is an extension of **CQ4** with a time-period specification. The same query used for **CQ4** can be reused for this one with temporal information. However, more advanced queries can be implemented to find temporal matches. A query for finding which object was collected from a given location withing a specific time-period is provided in the supplementary material."

All data statement are represented within one or more named-graph depending on source(s) and named graphs are connected to corresponding source(s) or responsible agent(s) acting on behalf of institution(s). The query for **CQ7** retrieves the source that directly connects to data statements and sources that are connected through the associated activity. The alternate query to find out who/which institution makes this statement is given in the supplementary document.

The answer to **CQ8** is straightforward, as each research activity generating object provenance data is represented as an **prov:Activity**. Each **prov:Activity** is then connected to one or more **prov:Agent**, according to the PROV ontology, which can answer who/which institution is conducting the research.

The query for **CQ9** retrieves the associated named graph and **prov:activity** responsible for any given statement. It lists all versions of these activities in descending order of execution time, and the latest version can be retrieved with filtering or by specifying *LIMIT* 1.

In conclusion, we demonstrate that the proposed model can answer all nine questions, although some of them are too broad and require further interpretation to be answered through SPARQL queries.

6. Discussion

The query results and the implementation of real-world object provenance (both from seen and unseen report) confirm that the combination of CIDOC-CRM, PROV-DM and named graphs can be used to model the representation of object and data provenance. From a technological feasibility perspective, we did not observe particular obstacles in representing ethnographic cultural heritage objects' provenance information in an interoperable manner. Nonetheless, it is essential to note that provenance research produces a mass of information; thus, unstructured data, i.e., written narrative report on a single object's biography may contain richer information. Additionally, there will always be a trade-off between expressivity and efficiency in digital humanities. Therefore, the representation of object provenance in a Knowledge Graph might not contain all the information recorded in the textual format regarding the provenance of an object. However, because the model supports representations of complicated networks between objects, people, places, and events utilizing the model, it projects an valuable overview to contextualize objects.

The provenance report summary can mostly be recorded using our model; however, we would like to highlight the interesting findings including the limitations and challenges encountered when modeling evaluation objects. Actor background and biography were not included in the ontology scope. The context of the collection was preserved using related historical events, dates, and places of collection, as well as specifying the form of acquisition, but the textual narrative may contain more information. The model does not distinguish between current custodian and current possessor and does not address predecessor relationships between organizations, such as mergers or renaming. These issues are beyond the scope of this paper and would require a deeper understanding of domain needs.

Polyvocality, observed through various theories of origin and acquisition, is crucial in determining the provenance of cultural heritage objects. Although the model supports polyvocal information representation, it does not prioritize one theory over another, especially when provenance reports may list and question information simultaneously. The model lacks the ability to assign weight to different statements, even if they contradict each other, due to the absence of an existential quantifier. Another important issue is that the current data model do not support information misrepresentation happened in the past. To preserve this information, the model places such statements under different named graphs and connects them with agents who made the statements. However, there is no means to indicate that this information is no longer considered valid. One solution is to use time-period information with named graphs to be able to refer to historical (event-based and provenance based) context.

To simplify the management of a cultural heritage object's chain of custody and to provide an abstraction over multiple objects, this study suggests using **E8 Acquisition** for legally recognized acquisitions and **E10 Transfer of Custody** for all other transfers. These entities can be further specified with sub-activities to define the type of ownership transfer, enabling institutions to model their specific notions of accession and deaccession. The International Council of Museums' documentation standard emphasizes the importance of using controlled terms to ensure consistent documentation [6], but the domain-standard vocabulary, AAT, lacks terms for unethical acquisitions.

7. Conclusion

Previous research identified a need for extending the domain ontology CIDOC-CRM to provide effective solutions for modelling multiple interpretations of cultural heritage object [17]. This study identifies requirements for modelling multiple perspectives on biographies of cultural heritage objects. After analyzing six distinct examples of ethnographic object provenance reports and considering the requirements, this paper proposes a data model that utilizes existing ontologies, i.e., CIDOC-CRM and PROV-DM, along with RDF-named graphs. Validation on six seen and five unseen objects confirms that the proposed model addresses complex chain-of-custody, data provenance, and multi-perspective representation requirements. We therefore conclude that the proposed data model allows to express cultural heritage object provenance in an interoperable manner for domain use.

In the field of heritage and humanities, and especially in the context of "decolonization" of the museums' databases, it is crucial that multiple (temporal, cultural and geographical) views from researchers, source communities and others, can be represented in the data structures. Although we focus in this research on ethnographic heritage collection's provenance information, the findings have implications on a more general provenance report to express such data polyvocality. Future work should incorporate information extraction tools to automate data conversion from textual reports of such knowledge that is inherently complex. The other possibility is facilitating the domain expert with easy tooling support to allow data modelling by themselves. Additionally, the provided model can be extended with methods to assign degrees of certainty to statements to allow data modellers to indicate the confidence levels of those statements.

Acknowledgements

The authors would like to thank the domain experts for their invaluable contributions. This work is partially based on previous work conducted in collaboration with X (*name removed for anonymity*).

References

[1] Dijkshoorn C, Aroyo L, Van Ossenbruggen J, Schreiber G. Modeling cultural heritage data for online publication. Applied Ontology. 2018;13(4):255-71.

[2] Bizer C, Heath T, Berners-Lee T. Linked data: The story so far. In: Semantic services, interoperability and web applications: emerging concepts. IGI global; 2011. p. 205-27.

[3] Knoblock CA, Szekely P, Fink E, Degler D, Newbury D, Sanderson R, et al. Lessons learned in building linked data for the American art collaborative. In: International Semantic Web Conference. Springer; 2017. p. 263-79.

[4] Turner H. Cataloguing culture: Legacies of Colonialism in museum documentation. UBC Press; 2020.

[5] Modest W. Ethnographic Museums and the Double Bind. Matters of Belonging. 2019.

[6] Grant A, Nieuwenhuis J, Petersen T. International guidelines for museum object information: the CIDOC information categories. International Committee for Documentation of the International Council of . . . ; 1995.

[7] Campfens E. The Bangwa queen: artifact or heritage? International Journal of Cultural Property. 2019;26(1):75-110.

[8] Turner H. Organizing Knowledge in Museums: A Review of Concepts and Concerns. Knowledge Organization. 2017;44(7).

[9] van Erp M, de Boer V. A polyvocal and contextualised semantic web. In: European Semantic Web Conference. Springer; 2021. p. 506-12.

[10] Choy SC, Crofts N, Fisher R, Lek Choh N, Nickel S, Oury C, et al.. The UNESCO/PERSIST Guidelines for the selection of digital heritage for long-term preservation. UNESCO Digital Library; 2016. https://unesdoc.unesco.org/ark:/48223/pf0000244280.

[11] de Boer V, Wielemaker J, van Gent J, Oosterbroek M, Hildebrand M, Isaac A, et al. Amsterdam museum linked open data. Semantic Web. 2013;4(3):237-43.

[12] Hyvönen E. Cultural heritage linked data on the semantic web: Three case studies using the sampo model. VIII Encounter of documentation centres of contemporary art: open linked data and integral management of information in cultural centres Artium, Vitoria-Gasteiz, Spain, October. 2016:19-20.

[13] Mouromtsev D, Haase P, Cherny E, Pavlov D, Andreev A, Spiridonova A. Towards the Russian Linked Culture Cloud: Data Enrichment and Publishing. In: Gandon F, Sabou M, Sack H, d'Amato C, Cudré-Mauroux P, Zimmermann A, editors. The Semantic Web. Latest Advances and New Domains. Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 637-51.

[14] Goerz G, Albers L. Representing place in space and time-methodological aspects in modelling the provenance of cultural heritage knowledge. In: Provenance of Knowledge. Proceedings CIDOC Conference; 2018. .

[15] Doerr M. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine. 2003;24(3):75-5.

[16] Bikakis A, Hyvönen E, Jean S, Markhoff B, Mosca A. Special issue on Semantic Web for Cultural Heritage. Semantic Web. 2021;12(2):163-7.

[17] Van Ruymbeke M, Hallot P, Billen R. Enhancing CIDOC-CRM and compatible models with the concept of multiple interpretation. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2017;2.

[18] Modest W, Lelijveld R. Words Matter: An Unfinished Guide to Word Choices in the Cultural Sector. National Museum of World Cultures; 2018. https://www.materialculture.nl/en/publications/words-matter.

[19] Captain E, Osbourne A, Somers Miles R, Tzialli E. Inward Outward, Critical archival engagements with sounds and films of coloniality. 2020 Inward Outward Symposium. 2020.

[20] Isaac A, et al.. Europeana data model primer. Europeana; 2013.

[21] Ockeloen N, Fokkens A, Ter Braake S, Vossen P, De Boer V, Schreiber G, et al. BiographyNet: Managing Provenance at Multiple Levels and from Different Perspectives. In: LISC@ ISWC. Citeseer; 2013. p. 59-71.

[22] Carroll JJ, Bizer C, Hayes P, Stickler P. Named graphs, provenance and trust. In: Proceedings of the 14th international conference on World Wide Web; 2005. p. 613-22.

[23] Anonymous. xyz. Master thesis. XX University; 2022.

[24] Castillo-Montoya M. Preparing for interview research: The interview protocol refinement framework. The qualitative report. 2016;21(5):811-31.

[25] Wiśniewski D, Potoniec J, Ławrynowicz A, Keet CM. Analysis of ontology competency questions and their formalizations in SPARQL-OWL. Journal of Web Semantics. 2019;59:100534.

[26] Johnson S, Veys FW, editors. *(Foreword by Henrietta Lidchi)* Provenance. vol. 1. Nationaal Museum van Wereldculturen; 2020.

[27] Veys FW, editor. *(Foreword by Henrietta Lidchi)* Provenance. vol. 2. Nationaal Museum van Wereldculturen; 2021.

[28] Moreau L, Missier P, Belhajjame K, B'Far R, Cheney J, Coppens S, et al. PROV-DM: the PROV data model technical reports. World Wide Web Consortium. 2012.

[29] Sandusky RJ. Computational provenance: Dataone and implications for cultural heritage institutions. In: 2016 IEEE International Conference on Big Data (Big Data). IEEE; 2016. p. 3266-71.