Polyvocal Knowledge Modelling for Ethnographic Heritage Object Provenance

Anonymous authors

Institute
emails

Abstract. Many data sets within the Cultural heritage sector are available as Knowledge Graphs. The domain has produced reusable models that have become domain standards for representing object metadata. However, information about the biography of the object (*object provenance*) is often unavailable to researchers in this structured format and remains accessible only in unstructured textual provenance reports. This limits the findability and reusability of the provenance information for domain experts. We address the challenge of defining a data model to represent complex and ethnographic cultural heritage objects' polyvocal provenance data, which includes multiple interpretations of, and theories for, the object biography, chains of custody and context of acquiring. We elicit a set of requirements for the data model from five provenance experts through (semi-)structured interviews. These requirements are used to define a set of modelling choices for representing the object provenance, which reuses domain standards CIDOC-CRM, PROV-DM, and RDF-named graphs. Finally, we validate the model by modelling six ethnographic cultural heritage objects based on information from two provenance reports and addressing nine competency questions elicited from the domain experts. The results show that the adapted combined model allows us to express the heterogeneity and polyvocality of the object and provenance information, trace data provenance and link with other data sources for further analysis.

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 [10]. One of the reasons for adopting these technologies in this domain is that the graph structure allows to express heterogeneity of information [3], while facilitating interoperability. However, producing machine-readable knowledge graphs from existing structured data and unstructured sources is not a trivial operation [17]. 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. Especially in the context of post-colonialism and "decolonization of the database",

heritage institutions are seeking out how to incorporate previously underrepresented voices in their practice, collections and information systems (cf. [25]).

How to model such polyvocal knowledge is the topic of this paper.

Cultural heritage institutions, especially those with ethnographic collections, continuously (re)contextualise objects by learning new facts about objects' biographies [18]. 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 [13]. It can also be used to form an assessment of the authenticity of an artefact and identify any unlawfully appropriated works [5]. In many cases, the details of this extensive research are not available in a structured format but only reported in a narrative document in an unstructured textual format, which limits the findability, reusability and interoperability of such information. Typically, the person who conducts the research or the institute they are presenting reflects on the object metadata [24]. 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 categorisation has equal merit to be true, 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 [26].

The Semantic Web as an information architecture and Knowledge Graphs as the data model are promising technologies of such polyvocal knowledge representations [26]. Its dispersed and networked nature makes it ideally suited to handle diverse opinions. At the same time, it is also crucial to preserve competing views with the sources of its 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 preserve multiple perspectives in data.

This paper's contribution lies in examining how semantic web technologies can be used to meet these requirements specified by domain experts. We first identify domain requirements for representing multi-perspective ethnographic object provenance information. We propose how such information can be modelled and apply this model to six ethnographic objects demonstrating the expression of complex chains of custody of the object biographies while preserving data sources. 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 cultural heritage 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 [9]. 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 on the representation of cultural heritage metadata [10,17,4,14,21]. However, none of these works investigated how to model rich 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 [10,12]. The event-centric model represents knowledge through associated events, such as acquisition or production. An ISO standard since 2006, CIDOC-CRM [11,2] 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 [27] 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. Paper [27] argues 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 [19,6]. Dijkshoorn et al. [10] 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 [15] 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. [22], 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 [7]. Bizer et al. [7] 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. This analysis was done in the context of previous work [1][1], and the procedure and findings of that study are listed in this section. We present the approach and the resulting list of requirements and competency questions.

3.1 Approach

To collect data requirements for historical object collections and 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; they attempt to investigate current problems and concerns. We approach requirements engineering as an iterative process: the identified requirements and elicited Competency Questions from that are used both for model construction and validation.

For the focused interview, we recruited 5 participants from the domain of Museology, all currently involved with a project called Pressing Matter[2] in different capacities (see Table 1 for an overview of the interviewees). Pressing Matter is a Dutch project investigating the potentialities of artefacts collected during the colonial period in order to support societal reconciliation with the colonial past. The professionals are carefully chosen with the motivation to understand what is missing in the information system in use and what is necessary. All participants have experience working with the current museum information system and are also responsible for updating object metadata with object provenance identified in their research. These professionals can be considered the primary end-users for a developed data model from this research.

Each of the participants attended around one hour-long individual semi-structured interviews. Except for the first one, all interviews were via a web conferencing tool. The first interview was conducted as a pilot study to ensure all questions were clearly understood. Since the answers from that interview were insightful and guiding, it has been incorporated into the study. Each interview used the same interview guide, which served as a procedural protocol for directing the interview. The interview guide [8] is aligned with the objective of this research[3]. It has been compiled to address the requirements for proper representation of cultural heritage provenance data, highlighting four different aspects: 1) Provenance research: the process and the common challenges, which sources are used and their opinion on which provenance-related information is important to represent; 2) Documentation of their provenance research; 3) Representation of provenance information; 4) The utility of provenance information.

[1] name and citation removed for anonymity
[2] `https://pressingmatter.nl/`
[3] The interview guide can be found at `https://doi.org/10.5281/zenodo.7439306`

Table 1. Overview of 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

3.2 Findings

This section is divided into three different 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 The interview results imply that there is no standard goal with provenance research, some state it is about tracing the history of an object, where others think it is essential to understand in which ways objects were implicated in (colonial) history and what this implication means for their future. Common to all respondents was that the provenance research helps to gain a better understanding of where collections and objects come from and lead to better-documented collections.

When asked about how provenance research is conducted, different processes were described. Mainly for the reason that interviewees have different approaches; some heavily rely on archives while others utilises different sources such as libraries and web-search. All respondents, however, initiate their research from the museum collection management system (CMS)[4]. Two of the participant mentioned that it often contains missing information and observational bias as facts, depending on the previous researcher who documented the entry.

After the research is conducted, some form of documentation is made. The respondents shared that there are no guidelines on 'how' to represent provenance information, though some white guidelines for the documentation are available but specific details are missing. When the provenance information gets beyond the scope of a single field in the CMS, a separate provenance report is often written. Otherwise, the CMS will be updated with the necessary first-level metadata about the object. It is important for all the respondents to be able to trace the sources from these object reports to asses or retrace the information. 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 The main motivation for requirement engineering in this research is to understand the nature of the problem and the challenges of current information representation. In that response, all

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

participants agreed that the information in the CMS can be faulty, incomplete or unreliable. One of the problems out of many is the lack of digitisation of archival material (R5). On the other hand, the available materials for information are decentralised, which leaves museum authorities with a poor understanding of their objects' origin, as mentioned by one respondent (R3).

One of the respondents (R4) states that current representations of provenance in the CMS do not match the complexity of the provenance research, and some statements are difficult to represent using the tool. R3 states that since the CMS is object-centred, important relations between people, places and objects cannot be represented in a machine-readable way. Another problem identified by all the respondents is that the management system does not contain any data provenance information, making it difficult to trace previous provenance statements made about an object, forcing the respondents to base their research on previous research that is impossible to trace.

Opinions on multi-perspective representation cultural heritage data In response to multi-perspective representation, all respondents agreed that keeping the nuance in object provenance information is important. Changing times and ways of thinking lead to new perspectives of perceiving information about those objects. Primarily when current representations of an object in the museum database reflect the dominant perspective of that time, e.g., colonial representatives (R2, R4). Therefore, these records are influenced by their perception of the world, a perception that some may not agree with now. One respondent (R4) states that it is important to acknowledge how objects and their history were seen before since it can tell us something about the collections. It is not always about stating that an object was acquired in certain ways; it is about keeping track of how we think about collections. Respondent R5 says that provenance rarely can be proven, so if multiple versions exist, they should all be represented.

Two respondents (R1, R3), however, state that even if nuance is important it is not practical to preserve every piece of information, and it depends on the goal of provenance research. One respondent (R2) also notes the importance of making a distinction between information deemed as more correct now contra prior research. According to one respondent (R2), it is crucial to understand that even if multiple perspectives are presented, some statements may not be true if other evidence weighs heavier. The respondents further agreed that their research is also a view on the provenance of an object; it is impossible to say that their research is the correct 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).

1. Overall representation
   - *Digital representation:* Provenance research is easier to conduct if cultural heritage data is digitally preserved.
   - *Event-centered:* It is easier to identify relations between actors, objects and places when they are represented in an event-centred way.
   - *Machine-readable:* Machine-readable representations are easier to access compared to when the provenance research is presented in a written report.
2. Object provenance
   - *Object data:* Representation of object title, object number, object category, material, part of which collection and its origin.
   - *Object creation:* When, where, and by who was the object created
   - *Collecting actors:* To identify the network of people in which an object exists.
   - *Locations:* To identify and note places where objects were acquired and/or created, as well as understanding the route an object travels.
   - *Events and time periods:* A historical event can depict a historical context, or it can describe a time period where it was known that a lot of colonial objects are acquired unethically.
   - *Multiple descriptions:* If multiple views on an acquisition exist they should be noted to keep nuance.
   - *Comments:* An event may need to be described with further detailed comments and notes in natural language.
3. Data provenance
   - *Provenance statement source:* Users should be able to review the sources that explain and describe the object provenance.
   - *Source:* Users should be able to find the source if they desire to review it.
   - *Traceability to previous research:* Each version of object provenance research should include data provenance information.

Competency Questions Competency questions (CQs) are questions in natural language that outline the knowledge and specify the constraints for knowledge representation [29].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. It is also important to convey information about *dates and events*; for example, the date or year the object was obtained. This enables identifying networks of connected objects through historical events, which may collectively project on an objects' acquisition. *Geographic locations* are important to determine which items were bought or sold in particular regions or countries. 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. 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.

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

ID	Question	SPARQL query	Answers CQ?
CQ-1	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.
CQ-2	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.
CQ-3	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.
CQ-4	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
CQ-5	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
CQ-6	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
CQ-7	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
CQ-8	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
CQ-9	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 ?s ?p ?o the query return provenance report associated with it in descending publishing order.

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"[16] and "Provenance #2" [28]. These provenance reports are issued by the Dutch National Museum of World Cultures (NMVW)[5] to embed provenance of artworks 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" [28]. 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" [16], 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.

4.1 Model Description

The first step is to reconstruct the acquisition history 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.

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

[5] `http://wereldculturen.nl/`

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[6], with adjustments based on use-case requirements. The collection object itself is an instance of **E24 Physical Human-Made Thing**, 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)
- 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.

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

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

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**, 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[7] visualizer. This diagram only specifies the shape for the primary entities, less significant entities' detail is left out here for visual simplicity.

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[20] ontology. Named graphs can be used to attach provenance data and model context and scope assertions[8]. This provides the capacity to assess various assertions made in a graph by the information providers [7]. 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.

[7] `https://rdfshape.weso.es/`
[8] cf. `https://www.w3.org/2009/07/NamedGraph.html`

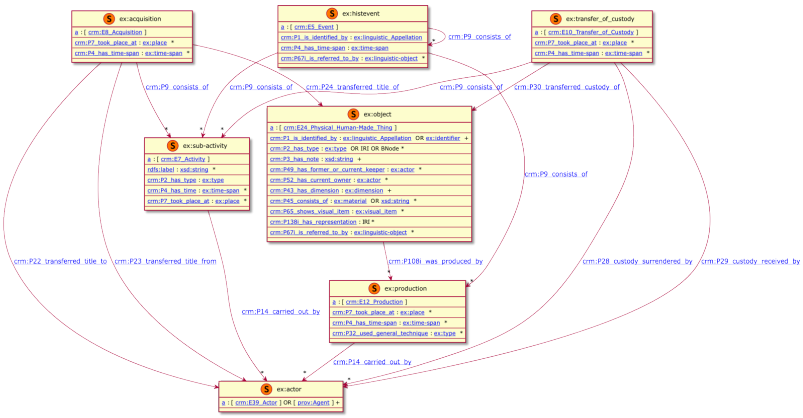

Fig. 1. Ontology overview of cultural heritage object provenance modelling

PROV-DM[20] is a conceptual model for modelling provenance, with PROV-O being a mapping to RDF[9] with proven applicability in the cultural heritage domain [22,23]. In our model it is used for attaching data provenance to a named graph IRI. A typical 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 (see Figure 2).

Polyvocal Modelling 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 and its location and/or time period to the view according to PROV-DM ontology; and, more importantly, it allows to tag sources of each statement. Each named graph or triple collection is represented as a **prov:Entity** that was derived from another **prov:Entity** which is a source.

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

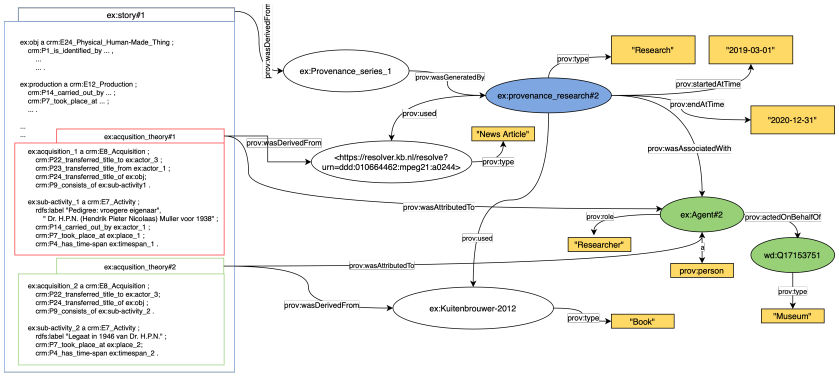

Fig. 2. Multiple views or theories of Object acquisition are separated into named graphs and tagged with sources following the PROV-DM ontology. The transparent elliptic shapes represent **prov:Entity**, blue elliptical shapes represent **prov:Activity**, and green elliptical shapes represent **prov:Agent**.

5 Results and Validation

In this section, we describe the result of modelling the six objects as knowledge graphs initially used to construct the model. We also randomly selected five more objects from different project reports to understand the generalisability of the proposed model. For this purpose, we used five provenance reports from the project called Pilotproject Provenance Research on Objects of the Colonial Era (PPROCE)[10]. Later, we validate the resulting knowledge graph against the competency questions to determine to which extent the current data model satisfies the domain requirements. All files for this section can be found in our Zenodo repository[11].

5.1 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 call these six object 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. In addition, further named graphs are created to represent different view's on the same entity with their source of information. The entire knowledge graph consists of 6 instances of **crm:E24 Physical Human-Made Thing**, with 12

[10] https://www.niod.nl/en/projects/pilotproject-provenance-research-objects-colonial-era-pproce
[11] https://doi.org/10.5281/zenodo.7437713

crm:E8 Acquisition, 7 **crm:10 Transfer of Custody** and 27 **crm:E39 Actor** instances. A total of 20 **prov:Agent** instances and 38 **prov:Entity** instances are used to describe the data provenance. More statistics can be found in the Zenodo repository(/data/entity_stat.csv) where the knowledge graph itself is also available as a set of RDF TriG files under the data folder.

We further model 5 unseen objects' report from another project and we are going to call these objects set as evaluation objects. To draw a comparison on triple statics, we present Table 3. The number of total rdf triples (1290) and named graphs (27) are comparable to the construct objects set as we modeled only 5 objects' provenance data into a new knowledge graph. This consistency also remained on the number of instance of different other classes, i.e., **crm:E8 Acquisition** (8), **crm:10 Transfer of Custody** (7), **crm:E39 Actor** (14), **prov:Agent** (18), and **prov:Entity** (51). All the TriG files of this object set and details statics can be found under the "evaluation" folder in the Zotero repo.

To maintain data consistency and to ensure the data shape, the conceptual model is converted into ShEx[12] rules. We specify the property restriction and expected cardinality for each notable class. By writing ShEx ShapeMap queries, we validated each entity to ensure data consistency against the chosen ontology. For the validation interface, we used the RDFshape[13] web tool. All the ShEx rule and ShapeMap queries can be found in our repository.

Table 3. Comparison of triples statistics generated from 6 construct provenance object report (left table) and 5 evaluation object report(right table)

Entity	Count
Total triples	1778
Named graphs	31
crm:E24_Physical Human-Made Thing	6
crm:E12_Production	7
crm:E10_Transfer of Custody	7
crm:E8_Acquisition	12
crm:E7_Activity	21
crm:E5_Event	3
crm:E39_Actor	27
crm:E33_E41 Linguistic Appellation	40
crm:E33_Linguistic Object	13
crm:E42_Identifier	7
crm:E52_Time-Span	57
crm:E53_Place	11
crm:E54_Dimension	7
crm:E55_Type	109
crm:E56_Language	26
crm:E57_Material	6
prov:Activity	2
prov:Agent	20
prov:Entity	38

Entity	Count
Total triples	1290
Named graphs	27
crm:E24_Physical_Human-Made_Thing	5
crm:E12_Production	5
crm:E10_Transfer_of_Custody	7
crm:E8_Acquisition	8
crm:E7_Activity	12
crm:E5_Event	1
crm:E39_Actor	14
crm:E33_E41_Linguistic_Appellation	54
crm:E33_Linguistic_Object	11
crm:E42_Identifier	5
crm:E52_Time-Span	50
crm:E53_Place	10
crm:E54_Dimension	5
crm:E55_Type	66
crm:E56_Language	7
crm:E57_Material	14
prov:Activity	5
prov:Agent	18
prov:Entity	51

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

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

5.2 Validation through Competency Question

We validate the resultant Knowledge Graph by addressing the 9 competency questions elicited from the domain experts. For each CQs, we can devise a SPARQL query to produce the desired results. The rightmost column of Table 2 lists the queries. The constructed CQs may have different meaning; therefore a discussion on CQs and how the constructed SPARQL reflect on the CQs is given below.

For **CQ-1**, the listed query matches if the intent is to find out actors involved in object biography as a formal keeper or owner. However, in which exact capacity these actors are involved in objects provenance will require a different query. Different activities are connected with objects both with incoming (i.e., Acquisition and Transfer of Custody) and outgoing links (i.e., Production). So, query must search for actors in all those activities and retrieve the property that connects the actor with activity. This alternate query is given in the supporting material.

The **CQ-2** is interested in retrieving all objects that are connected with person A through any activity except Producion. The query for **CQ-2** retrieves both legal and illegal collection activities for an object and involved collectors; therefore answers the CQ accurately.

The **CQ-3** can be considered rather broad. If the intent is to find out if two actor can be linked with the same object then from the constructed graph, this is answerable. The listed query for **CQ-3** retrieves two persons involved through a shared activity with the same object. Other relations are also possible but would require separate queries. Nevertheless, if we want to find out any possible path between two actor, we need more complicated query from current data.

All activities can list it's location using crm:P7_took_place_at properties. So, each collection activity is connected with the location if this information is know. The query given for **CQ-4** targets to retrieve location when it is connected with sub-activity of Acquisition or Transfer of custody activity. Similar, queries can be writen when the location is connected directly with activity with activity. For detail, see Zenodo repository(validation_sparql.txt).

We answer **CQ-5** by making multiple 'hops', since historical events are not directly connected with the objects, but rather can consist of activities concerning the object.

CQ-6 This is an extension of CQ-4, but with specific time-period. So, the query used for **CQ-4** can be reused for this one with with temporal information. However, the current query checks for exact match of time-period, more advanced query can be implemented to find the temporal match.

All data statement are represented within one or more named-graph depending on source(s) and all named graph is connected to corresponding information source(s) or responsible agent(s) acting on behalf of institution. The query for **CQ-7** retrieves the source that directly connects to data statements and sources that are connected through the associated activity.

Answering to the **CQ-8** from the given model was rather straightforward, as each research activity generating object provenance data is represented an

prov:Activity. Each **prov:Activity** is then connected to one or more **prov:Agent** according to PROV ontology. The **prov:Agent** can then be represent who/which institute the were acting on behalf of; therefore states the institution conducting research.

Given any statement, the query for **CQ-9** will retrieve the associated named graph and **prov:activity** responsible to generate these statement. The query further with list all different version of these activities on descending order of time. To retrieve the latest version of each research will further need filtering which is not shown in the given SPARQL query.

In conclusion, the results show that the data represented through the proposed model is indeed capable of answering all nine of these questions. However, some of the CQs are conceptually too broad, therefore needed further interpretation by the researchers to answer through SPARQL query. Moreover, some of the queries need to be further refined with advanced query or other computational techniques to answer the CQs concretely. The overview of the SPARQL queries and their ability to answer respective CQ(s) is given on the Table 2.

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, a written report 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 text data recorded 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.

Majority of the information from the provenance report summary can be recorded as per our model; however, we are going to highlight the shortcomings or the difficulties observed while modeling evaluation objects. The background or biography of the actors involved is left behind as it is not within the scope of the given ontology. The context of the collection was tried to preserve in the form of related historical events, dates and places of collection and through mentioning sub-activity to specify the form of acquisition, i.e., loot, donation or "found". Yet, the narrative of the text will contain more information.

The current model did not make any distinction between current custodian and current possessor. There is no way in CIDOC-CRM to represent this distinction. Moreover, it fails to address predecessor relationship, when one organisation is taken over by another organisation and start operating under different name

(renamed or reformed). For example, one institution was the predecessor of another institution or multiple institutions were merged into one. Modeling these organisational transformations will require more understanding of domain need; therefore left out of the scope of this paper.

Polyvocality can be observed with different theory of place of origin, manner or place of acquisition. These information are extremely important for cultural heritage object provenance where past truth was established through current research. Though the model supports polyvocal information representation, there is no way to place emphasis on one theory than the others. In the provenance report sometimes information is listed and questioned at the same time. For example, "given the circumstances, the acquisition method is unlikely to be just 'found'". At this moment, the model does not support existential quantifier; therefore it was not possible to place weight on different statements, whether contradicting or not.

Another important issue observed during the evaluation object modeling is that the current data model do not support information misrepresentation happened in the past. In order to, preserve that information, this model places such statement under different named graph and connect with agents who made this statement. However, there was no means to state that, this information is not considered as truth today. As a solution, one may consider stating time-period information with name graphs to represent that all the containing statement is true within certain time-period. According to **prov:Entity** can also have time when they are valid; though we did not use that information in the current model.

To manage the complexity in the chain of custody of a cultural heritage object and to reach abstraction over multiple objects, this research recommends using **E8 Acquisition** for known legal acquisition and **E10 Transfer of Custody** otherwise. Both entities can further have sub-activities to specify the "type" of such transfer of ownership. This allows institutions to model their specific notions of accession and deaccession. The museum guideline authored by ICOM [13], which is seen as a standard on documentation, highlights the importance of using controlled terms when representing provenance information to ensure consistent documentation. However, the domain-standard vocabulary, AAT, for declaring different types does not provide any terms for unethical ways of acquisition.

7 Conclusion

Previous research found that domain ontology, CIDOC-CRM, does not provide efficient solutions for storing multiple interpretations [27]. Our research extends that by storing various interpretations on an entity in named graphs and by incorporating PROV-DM, it is possible to attach a source to a graph to meet the identified requirements concerning traceability and overcome the issues with multiple-perspective representation.

Further, this research demonstrates the application of named graphs combined with PROV-DM to represent data provenance. 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 level 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. Anonymous: xyz. Master's thesis, X University (XXXX)
2. Bikakis, A., Hyvönen, E., Jean, S., Markhoff, B., Mosca, A.: Special issue on semantic web for cultural heritage. Semantic Web **12**(2), 163–167 (2021)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data: The story so far. In: Semantic services, interoperability and web applications: emerging concepts, pp. 205–227. IGI global (2011)
4. de Boer, V., Wielemaker, J., van Gent, J., Oosterbroek, M., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Amsterdam museum linked open data. Semantic Web **4**(3), 237–243 (2013)
5. Campfens, E.: The bangwa queen: artifact or heritage? International Journal of Cultural Property **26**(1), 75–110 (2019)
6. 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)
7. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: Proceedings of the 14th international conference on World Wide Web. pp. 613–622 (2005)
8. Castillo-Montoya, M.: Preparing for interview research: The interview protocol refinement framework. The qualitative report **21**(5), 811–831 (2016)
9. Choy, S.C., Crofts, N., Fisher, R., Lek Choh, N., Nickel, S., Oury, C., Ślaska, K., et al.: The UNESCO/PERSIST guidelines for the selection of digital heritage for long-term preservation (2016)
10. Dijkshoorn, C., Aroyo, L., Van Ossenbruggen, J., Schreiber, G.: Modeling cultural heritage data for online publication. Applied Ontology **13**(4), 255–271 (2018)
11. Doerr, M.: The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine **24**(3), 75–75 (2003)
12. 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)
13. 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)
14. 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 pp. 19–20 (2016)
15. Isaac, A., et al.: Europeana data model primer (2013)
16. Johnson, S., Veys, F.W. (eds.): *(Foreword by Henrietta Lidchi)* Provenance, vol. 1. Nationaal Museum van Wereldculturen (2020)
17. Knoblock, C.A., Szekely, P., Fink, E., Degler, D., Newbury, D., Sanderson, R., Blanch, K., Snyder, S., Chheda, N., Jain, N., et al.: Lessons learned in building linked data for the american art collaborative. In: International Semantic Web Conference. pp. 263–279. Springer (2017)
18. Modest, W.: Ethnographic museums and the double bind. Matters of Belonging (2019)
19. Modest, W., Lelijveld, R.: Words matter: An unfinished guide to word choices in the cultural sector. https://www.materialculture.nl/en/publications/words-matter (2018)

20. Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., et al.: Prov-dm: the prov data model technical reports. World Wide Web Consortium (2012)
21. 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. (eds.) The Semantic Web. Latest Advances and New Domains. pp. 637–651. Lecture Notes in Computer Science, Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-18818-8_39
22. Ockeloen, N., Fokkens, A., Ter Braake, S., Vossen, P., De Boer, V., Schreiber, G., Legêne, S.: Biographynet: Managing provenance at multiple levels and from different perspectives. In: LISC@ ISWC. pp. 59–71. Citeseer (2013)
23. Sandusky, R.J.: Computational provenance: Dataone and implications for cultural heritage institutions. In: 2016 IEEE International Conference on Big Data (Big Data). pp. 3266–3271. IEEE (2016)
24. Turner, H.: Organizing knowledge in museums: A review of concepts and concerns. Knowledge Organization **44**(7) (2017)
25. Turner, H.: Cataloguing culture: Legacies of Colonialism in museum documentation. UBC Press (2020)
26. van Erp, M., de Boer, V.: A polyvocal and contextualised semantic web. In: European Semantic Web Conference. pp. 506–512. Springer (2021)
27. 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 **2** (2017)
28. Veys, F.W. (ed.): *(Foreword by Henrietta Lidchi)* Provenance, vol. 2. Nationaal Museum van Wereldculturen (2021)
29. Wiśniewski, D., Potoniec, J., Ławrynowicz, A., Keet, C.M.: Analysis of ontology competency questions and their formalizations in sparql-owl. Journal of Web Semantics **59**, 100534 (2019)