Versioning: Representing Cultural Heritage Evidences on CIDOC-CRM via a case study

Ariele Câmara[1,2], Ana de Almeida [1,2,3], João Oliveira [1,2,4]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Centro de Investigaçao em Ciências da Informaçao, Tecnologias e Arquitetura, Lisboa, Portugal
[3] Centre for Informatics and Systems of the University of Coimbra (CISUC), Portugal
[4] Instituto de Telecomunicações, Lisboa, Portugal

Ariele_Camara@iscte-iul.pt

Abstract. Understanding the elements that allow the recognition of archaeological structures is an essential task for the identification of cultural heritage. On the other hand, recording these elements is necessary for the historical study, evolution, and recognition of these types of structures. One of the challenges presented for digital representation of this information and knowledge relates with the fact that there are results from different surveys and records on the status of the same monument, that can be considered as separate versions of knowledge. In this paper, we describe a schema to represent versioning data about archaeological heritage dolmens using the CIDOC-CRM model as a basis. The versioning schema will work as a database model for the development of a knowledge-graph to aid with automatized dolmen recognition in images. The intended model efficiently stores and retrieves event-driven data, exposing how each update creates a new "version" of a new event. An event-driven model based on versioning data makes it possible to perform comparisons of versions produced at different times or people and allows for the creation of complex version chains and trees.

Keywords: archaeological structures, versioning, CIDOC-CRM, knowledge-graph, event-driven model

1 Introduction

Knowledge graphs have emerged as a technology that aims to provide semantic information capable of being understood and interpreted by machines about real-world entities (Souza Alves et al., 2018). To represent the knowledge about entities and processes, it is necessary to differentiate real things (phenomenal) from those described through information (declarative) (Hiebel et al., 2017). Mapping both the phenomenal and declarative knowledge of cultural heritage according to a common standard model such as the one provided by the International Committee for Documentation - Conceptual Reference Model (CIDOC-CRM), is a key to support interoperability.

The representation of historical, cultural, and archaeological data has traditionally been carried out by different specialists and maintained by institutions such as libraries, archives, and museums (Hiebel et al., 2017). The multiple sources and different researchers' backgrounds has led, through the years, to a disparity between data sources and formats, and different historical versions of declarative information, for example, data derived from interpretations of the same object (Doerr et al., 2011). Handling these metadata as a unique set is vital for different purposes such as information retrieval.
This paper explores the development of a schema to model a graph-based data model to represent the different versions of the knowledge acquired about dolmens - using the information about the structural elements that may help their recognition in satellite images. In order to achieve this goal, we adopted the CIDOC-CRM.

2 Representing data from heterogeneous sources

2.1 CIDOC-CRM

CIDOC-CRM is a formal ontology for the integration, mediation, and exchange of cultural heritage information with multi-domain knowledge. Its development started in 1996 and in 2006 it became an ISO 21127:2006 standard (Bekiari et al., 2021). Although it started as an ontology for museums, it is not limited by this usage and has been used for different purposes (Faraj & Micsik, 2021; Hiebel et al., 2017; Roman Bleier, 2019; Velios & Pickwoad, 2016). The CIDOC ontology (version 7.2.1) consists of 81 classes taxonomically organized and 160 unique properties to model knowledge. The most general classes are represented as a E77 Persistent Item, E2 Temporal Entity, E52 Time Span, E53 Place, E54 Dimension, and E92 Spacetime Volume.
As an event-centric model that supports the historical discourse, the CIDOC-CRM allows for the description of entities that are themselves processes or evolutions on time. Using the E2 Temporal Entity and its subclasses enables the description of entities that occurred at some point (or points) over time. One of these entities focus on the description of a temporal duration and documents the historical relationships between objects that were described using subclasses of E77 Persistent Item. CIDOC-CRM enables the creation of a target schema to join all the varied knowledge about a domain, since CRM provides a definition and a formal structure to represent the concepts and relationships of cultural heritage.

2.2 Definition – Schema Versioning

Schema evolution requires keeping the complete change history for the schema - so it is necessary to retain all previous definitions. Versioning mechanisms can be useful to support scholars in making research more transparent and discursive. The schema versioning idea was introduced in the object-oriented database system development context, as systems are implemented to deal with multiple schemas and evolution of information (Roddick, 1995).
We found few examples for versioning using the CIDOC-CRM as a way to track different versions. The authors in Velios & Pickwoad (2016) use an event-centric ap-

proach, where each entry represents a different version of bookbinding with all temporal classes directly connected to the same entity equivalent to each of the records on the cover of a given binding. Despite making it easier to understand from a human point of view, from a computational point of view having a unique entity that relates to several temporal instances will add cycles in the information graph. (Carriero et al., 2021) presents ArCo on how to develop and validate a cultural heritage knowledge graph – discussing an approach to represent dynamic concepts which may involve or change over time. Every change generates a new record version of the same persistent entity that is represented by the catalogue record and its versions are related to it. Each version is associated with a time interval and has temporal validity. Another work that we can mention is that in (Bürgermeister, 2020), which shows how Version Control Systems have different benefits in the field of digital humanities, thus proposing an implementation of versioning in collaboration with version history. However, it is not shown how to work with this model when using CRM to structure heritage data.
In a database there several versions of information can coexist, even more so when we talk about temporal databases, which is the case when we deal with information about cultural heritage derived from several investigations on a monument or about the analysis data generated about it.

2.3 Data Modelling Issues: Event-Version

Archaeological reasoning is supported by the multiple interpretations and theories stated, published, re-examined, and discussed over the years. Archaeology contains a rich and complicated example of argumentation used in a scientific community, showing how different fact-based theories were developed and changed over time (Doerr et al., 2011). The standard inferences, the sequence of factual observations, and the change of belief occurring over time can be represented using knowledge graphs. Despite this, there are some limitations to this representation of this data, such as (i) correctly grouping components from different periods/versions and (ii) scalability (Roman Bleier, 2019).
An event represents a single episode in the data collection or recording. This single event can only consist of an investigative technique and is therefore a unique entity in time and space. Different events may have new results over the same object. Thus, cultural property could be interpreted by various agents (e.g., researchers) at different moments in time, resulting in different interpretations. Event-based models are already well established (Bekiari et al., 2021; Guan et al., 2022; McKeague et al., 2020; Velios & Pickwoad, 2016). However, models as CIDOC were not designed to directly support the different pieces of information representing different perspectives interpretated by different agents, or new pieces of information generated through it (Carriero et al., 2021).
Using an event-centric model based on versions to structure data, we can 1) describe any number of component versions and 2) identify the components belonging to different versions (Roman Bleier, 2019). When we take into account the knowledge about the various phases of the thing being analyzed, we can link what we observe with the related events.

3 Recording Cultural Heritage: a Study Case Using Archaeological Monuments

The study case here presented deals with the representation of information about megalithic monuments classified as Dolmens in Pavia (Portugal) built in the Neolithic-Chalcolithic. These structures are one of the most representative and ubiquitous cultural features of prehistoric landscapes in Western Europe. The first systematic works about dolmens in Pavia were carried out by Vergílio Correia who published his research in 1921 (Câmara, 2017). His work is considered a benchmark for the knowledge of megalithic (Rocha, 1999). There are different records concerning research carried out in the area and on the same monuments.

For the development of the schema, we collected and analyzed all data available through the DGPC Digital Repository[1] regional archaeological map (Calado et al., 2012), the information provided by experts, and data obtained through photo interpretation. Unlike CIDOC-CRM, which is an event-based model, the data in most archaeological records on cultural heritage speaks implicitly about events. In this sense, we can use the CRM model to capture non-existent semantics and make explicit what is implicit. It helps to capture what lies between the existing semantic lines and structures into a formal ontology. At the same time, it serves as a link between heritage data, allowing to represent all this knowledge in a way that can be understood by people but also processed by machines, and thus allows the exchange, integration, research, and analysis of data. By making what is implicit explicit, we are able to use existing data to ask new questions and consequently obtain new results.

3.1 Object Based Record

Dolmens are persistent physical structures built by man, with a relatively stable form, which occupy a geometrically limited space and establish a trajectory in space-time throughout their existence. Following the hierarchy of classes defined by CIDOC, the E22 Human-Made Object is the most specialized class within the hierarchy of man-made persistent items. The dolmen, as a general term, is characterized here as a CIDOC-CRM entity ‘E22 Human-Made Object, which, according to the CIDOC-CRM class specification, “... comprises all persistent physical items of any size that are purposely created by human activity” (Bekiari et al., 2021).

3.2 Versioning Based Record

Since we work with data produced by different specialists at different times on the same object, the question of how to deal with such a rich and diverse number of primary sources is not simple, especially if the authorship and origin are not always clear (Roman Bleier, 2019). In order to do so we must focus on the content, with each record

[1] The DGPC is the State department responsible for managing archaeological activity in Portugal. Management of heritage is achieved through preventive archaeology and research, and records are provided viathe Archaeologist’s Portal: https://arqueologia.patrimoniocultural.pt/

being seen as a unique version of the same monument. As a result, we consider abstracts, records, or metadata that represent knowledge about the same entity as documents expressing a unique version about the monument. First a single instance is created representing where the knowledge was obtained. A new related entity is created for representing the information about the dolmen found in the document and, finally, these separate instances are connected through a new entity, as shown in Figure 1.

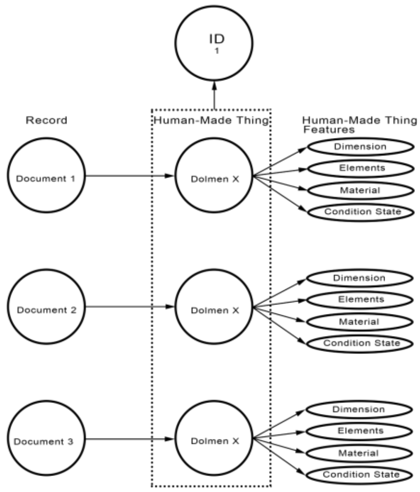

Figure 1: Schema representing the relationships in our versioning-based record to connect the item with their features by document.

We assign an ID to represent several E22 Human Made Thing instances that contain knowledge about the same human-made object. An ID is characterized as an instance of the CIDOC-CRM E42 Identifier and is used to group E22 Human Made Thing instances, each representing data about the same item but obtained from different documents. This model will record whatever activity over the object that generated the first record, whenever it is generated and acquired, while maintaining all previous knowledge to easily access it using the same class, creating a simple and non-cyclical model. A branch in this context means that N parallel versions can be developed separately at a certain point in the development model.

Since the goal is the representation of relevant information for posterior analysis, interpretation, and classification of images, in this case, for recognizing dolmens, the focus is on the dolmens structure representation and all the related elements that may assist in its recognition. To represent the structural information, E22 Human Made Thing instances can be used as output for new entities that allow the characterization of the elements it represents.

3.3 Event-Version Based Record

E2 Temporal Entity and its subclasses usage allows for the description of entities that occurred at some different point(s) over time. These entities focus on a temporal duration description and in records of the constant chronological relationships between the objects - to represent the information as an activity, beginning or end of something. However, this constant is not always respected. For example, when we talk about the beginning of the existence of a dolmen, we are talking about a phase of time described as Neolithic-Chalcolithic, semantically the existing connection properties would lead us to infer that the object description refers to its structure during this event and not its structure at the time it was analyzed and recorded, as is the case. Still, the initial structure is mostly unknown, since the structures may have been created, modified, and reused, and there are no records about these activities. In any case, this information would not help to identify structures in images. For our use case, we need a class to understand the actions of making claims about an object property and that allows us to access the date and place where the knowledge was obtained - or at least know all the characteristics of the object at the time of data collection.

The E13 Attribute Assignment class comprises the actions of making assertions about a property of an object or any unique relationship between two items or concepts, allowing to describe the people's actions making propositions and statements during scientific procedures, for example, who or when a condition statement was made. Note that any instance of properties described in a knowledge base such as this is someone's opinion - which in turn should not be recorded individually for all instances in favour of avoiding an endless resource whose opinion was the description of another opinion (Bekiari et al., 2021). However, for the present case, as the description obtained by different entities sometimes contain contradictory data, a model that works with the different views is necessary. Thus, these fragmented reports can be seen as versions that can enrich and complement our knowledge of the monuments and their relations, but they can also present conflicting information and narratives and multiple E13 instances can potentially lead to a collection of contradictory values. This redundant modelling of alternative views is preferred here because when talking about structural features, they all become relevant for a better perception of the object and how it may have been affected and affect its surrounding environment - which can help recognition. In this sense, we use the E13 Attribute Assignment entity to record the action of describing the dolmen and connect the event to the object with its descriptions. Using records as events to deal with different pieces of information of dolmens status, and unique IDs to group instances concerning the same dolmen, made it possible to overcome the issue.

To associate the action of describing the dolmen to where the information was obtained, we use the E31 Document entity. This class allows for the representation of information on identifiable material items that originated propositions about the object under analysis. Thus, the relationship with the E31 Document entity is described based on the type of document that records the information. In addition to the document with the object description, we recorded the date of the information using the E52 TimeSpan entity. This information is relevant to prioritize the most current knowledge and enable the analysis of the chronological order of events that led to the description of the dolmen represented at that time. The schema model described is shown at Figure 2.

By using records as events and by considering each new record on the same monument as a unique version of the same, we create a model capable of dealing with the fact that different research works were, are being, or can be performed on the same monument, resulting in different outcomes since they can be made by different researchers, with diverse approaches and at different periods in time. Therefore, we manage to keep all the information that can later be relevant for the recognition of these or of similar structures.

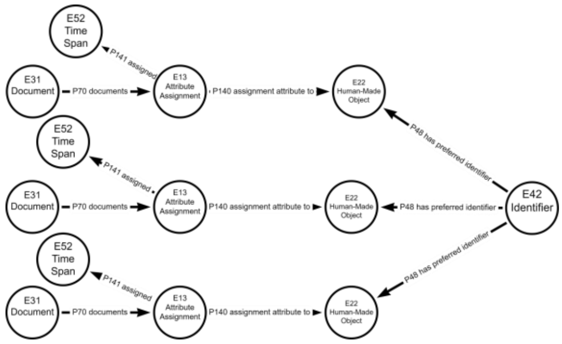

Figure 2: The archaeological monuments described in a record are represented as instances of Human-Made Object. When different documents report the same object, they are represented as unique entities (Human-Made Object) and related by a local ID. The local ID relates entities about the same monument and each Human-Made Object is related to the record activity and document, where knowledge is represented as acquired.

4 Conclusion

This article proposes the implementation of versioning in a model defined by CIDOC-CRM. We defined a new scheme model to represent different versions of information about the same monument, keeping all the previous and new knowledge without the need of merging, that could lead to incongruent information on the same entity due to different approaches in time and methodology. Thus, we developed an interoperable model capable of storing, analyzing, and retrieving data quickly and effortlessly, allowing cross-reference information, identifying patterns, and assisting in automated classification and recognition methods of these or similar structures in images. The next phases of the project involve the development of the schema model to represent the physical and geographical characteristics of the structure and the surrounding landscape to generate a knowledge graph capable of contextualizing all the elements that allow the identification of dolmens in the territory.

ACKNOWLEDGMENTS

This work was partially supported by the Fundação para a Ciência e a Tecnologia, I.P. (FCT) through the ISTAR-Iscte project UIDB/04466/2020 and UIDP/04466/2020, through the scholarship UI/BD/151495/2021.

References

Bekiari, C., Bruseker, G., Doerr, M., Ore, C. E., Stead, S., & Velios, A. (2021). *Volume A: Definition of the CIDOC Conceptual Reference Model*.

Bürgermeister, M. (2020). Extending Versioning in Collaborative Research. In *Versioning Cultural Objects Digital Approaches* (pp. 171–190). http://dnb.d-nb.de/

Calado, M., Rocha, L., & Alvim, P. (2012). *O Tempo das Pedras. Carta Arqueológica de Mora*. Câmara Municipal de Mora. https://dspace.uevora.pt/rdpc/handle/10174/7051

Câmara, A. (2017). *A fotointerpretação como recurso de prospeção arqueológica. Chaves para a identificação e interpretação de monumentos megalíticos no Alentejo: aplicação nos concelhos de Mora e Arraiolos* [Universidade de Évora]. https://dspace.uevora.pt/rdpc/handle/10174/22054

Carriero, V. A., Gangemi, A., Mancinelli, M. L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2021). Pattern-based design applied to cultural heritage knowledge graphs. *Semantic Web*, *12*(2), 313–357. https://w3id.org/arco

Doerr, M., Kritsotaki, A., & Boutsika, K. (2011). Factual argumentation-a core model for assertions making. *ACM J. Comput. Cult. Herit*, *3*(8). https://doi.org/10.1145/1921614.1921615

Faraj, G., & Micsik, A. (2021). Representing and Validating Cultural Heritage Knowledge Graphs in CIDOC-CRM Ontology. *Future Internet 2021, Vol. 13, Page 277*, *13*(11), 277. https://doi.org/10.3390/FI13110277

Guan, S., Cheng, X., Bai, L., Zhang, F., Li, Z., Zeng, Y., Jin, X., & Guo, J. (2022). What is Event Knowledge Graph: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 1–20. https://doi.org/10.1109/TKDE.2022.3180362

Hiebel, G., Doerr, M., & Eide, Ø. (2017). CRMgeo: A spatiotemporal extension of CIDOC-CRM. *International Journal on Digital Libraries*, *18*(4), 271–279. https://doi.org/10.1007/S00799-016-0192-4/FIGURES/6

McKeague, P., Corns, A., Larsson, Å., Moreau, A., Posluschny, A., Daele, K. van, & Evans, T. (2020). One Archaeology: A Manifesto for the Systematic and Effective Use of Mapped Data from Archaeological Fieldwork and Research. *Information 2020, Vol. 11, Page 222*, *11*(4), 222. https://doi.org/10.3390/INFO11040222

Rocha, L. (1999). Aspectos do Megalitismo da área de Pavia, Mora (Portugal). *Revista Portuguesa de Arqueologia*, *2*(1). https://dspace.uevora.pt/rdpc/handle/10174/2248

Roddick, J. F. (1995). A survey of schema versioning issues for database systems. *Information and Software Technology*, *37*(7), 383–393. https://doi.org/10.1016/0950-5849(95)91494-K

Roman Bleier, S. M. W. (Ed.). (2019). *Versioning Cultural Objects: Digital Approaches*.

Souza Alves, T. de, de Oliveira, C. S., Sanin, C., & Szczerbicki, E. (2018). From Knowledge based Vision Systems to Cognitive Vision Systems: A Review. *Procedia Computer Science*, *126*, 1855–1864. https://doi.org/10.1016/J.PROCS.2018.08.077

Velios, A., & Pickwoad, N. (2016). Versioning materiality: Documenting evidence of past binding structures. . In *Versioning Cultural Objects Digital Approaches* (pp. 103–126). http://dnb.d-nb.de/