



On using the CIDOC CRM to model archaeological datasets

Olivier Marlet, Théo Roulet, Florian Hivert, Béatrice Markhoff, Xavier Rodier, Gaël Simon

► To cite this version:

Olivier Marlet, Théo Roulet, Florian Hivert, Béatrice Markhoff, Xavier Rodier, et al.. On using the CIDOC CRM to model archaeological datasets. CAA2021 - Digital Crossroads, Jun 2021, Limasolle (virtual), Cyprus. halshs-04199967

HAL Id: halshs-04199967

<https://shs.hal.science/halshs-04199967>

Submitted on 8 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CAA2021 – Digital Crossroads

S2. Hic sunt dracones – Improving knowledge exchange in the Semantic Web with Linked Open and FAIR data (Standard)

Authors :

- Olivier Marlet (CITERES-LAT, Tours, Consortium MASA)
- Théo Roulet (LIFAT, Tours)
- Floriant Hivert (MSH Val de Loire, Tours)
- Béatrice Markhoff (LIFAT, Tours)
- Xavier Rodier (CITERES-LAT, MSH Val de Loire, Tours, Consortium MASA)
- Gaël Simon (UMR CITERES-LAT, Tours)

On using the CIDOC CRM to model archaeological datasets

The rise of the Semantic Web and the opening of research data in the Open Science dynamic are encouraging French archaeologists to apply FAIR principles to make their data sustainable, interoperable and reusable. This is one of the main missions of the MASA consortium (Memories of Archaeologists and Archaeological Sites): to support archaeologists in this process by relying on the CIDOC CRM, a standard ontology massively shared by a large part of the Cultural Heritage actors at the international level.

This paper's goal is to present how modeling with the CIDOC CRM can be adjusted according to the issues. The richness of this ontology allows indeed to set up models at different granularities in order to reach the according objectives.

We propose to compare two experiences made in two research projects: the first one relies on a generic model for several archaeological datasets within the MASA Consortium; the second one requires a more specific model, for excavation data and architectural data, within the SESAMES project (Semantization and Spatialization of Multi-Scale Heritage Artifacts).

It is important to notice that both of these models are extracted from the CIDOC CRM and its extensions, without defining any new concept or property. This is a choice that we believe is required in order to be consistent with the best practices of the semantic Web and the FAIR principles, and especially to ensure the data's interoperability.

1. OpenArcheo generic model for archaeological semantic data.

The MASA consortium has developed a SPARQL query platform for archaeological datasets, named [OpenArcheo](#) (Fig. 1). These datasets are quite heterogeneous (field recordings, epigraphic inventory or catalogue of artifacts). Each of these datasets has of course a different format and data structure. To pass these datasets into the OpenArcheo triplestore, RDF triples became the obvious format. For the semantic structure, we naturally chose the ontology of CIDOC CRM now widely shared in the cultural heritage domain.

The screenshot displays the OpenArcheo Explorer interface. At the top, there's a blue header with the OpenArcheo logo, 'OpenArcheo Explorer', and navigation links 'Welcome' and 'Explore'. A language selector shows 'en'. The main section is titled 'Explore' with 'Sources queried : Base ARSOL' and a link to 'Select other sources'. Below this is a visual query builder with two rows of concepts: 'Burial' (with a tomb icon) 'found in' 'Site' (with a location pin icon), and 'Site' 'studied by' 'Actor' (with a person icon) 'Elisabeth Lorans' (with a plus icon). A 'Where' label is connected to the 'Site' in the first row. An 'EXECUTE QUERY' button and a dropdown for 'Table' are at the bottom of the builder. A green status bar indicates 'Query successful ! - View/hide SPARQL query'. Below this is a SPARQL query editor showing a query that selects distinct values for '?this' and '?thisLabel' from the 'ARSOL' source, with a WHERE clause defining relationships between 'Burial', 'Site', and 'Actor' using CIDOC CRM URIs. At the bottom, there's a toolbar with options like 'Response', 'Table' (selected), 'Pivot Table', 'Geo', 'Google Chart', and 'Timeline'. It also shows 'Showing 1 to 50 of 722 entries', a search bar, and a 'Show 50 entries' dropdown. The results table has two columns: 'this' and 'thisLabel'. The first three rows show URIs for 'Burial' and their corresponding labels 'Sepulture AJ000001', 'Sepulture AJ000002', and 'Sepulture AJ000003'.

this	thisLabel
http://arsol.univ-tours.fr/4DACTION/WFICHEWEB/isepuAJ000001	Sepulture AJ000001
http://arsol.univ-tours.fr/4DACTION/WFICHEWEB/isepuAJ000002	Sepulture AJ000002
http://arsol.univ-tours.fr/4DACTION/WFICHEWEB/isepuAJ000003	Sepulture AJ000003

Fig. 1: OpenArcheo, a semantic Web platform for archaeological data.

We therefore analyzed the most common concepts in several datasets and made a selection from the CIDOC entities that best matches these concepts. Of course, a few purely archaeological concepts did not find an entity

that could represent them finely enough in the base CRM. Thus, we use the CRMarchaeo which provides us with the turnkey concept of the Stratigraphic Unit and also the CRMba extension for archaeological buildings.

It would have been irrelevant to fill in the gaps in the CRM with a new ontology specific to our project, especially for data interoperability. Therefore, we selected a few entities within these extensions when the CRMbase concepts were not precise enough for our modelling needs (Fig. 2).

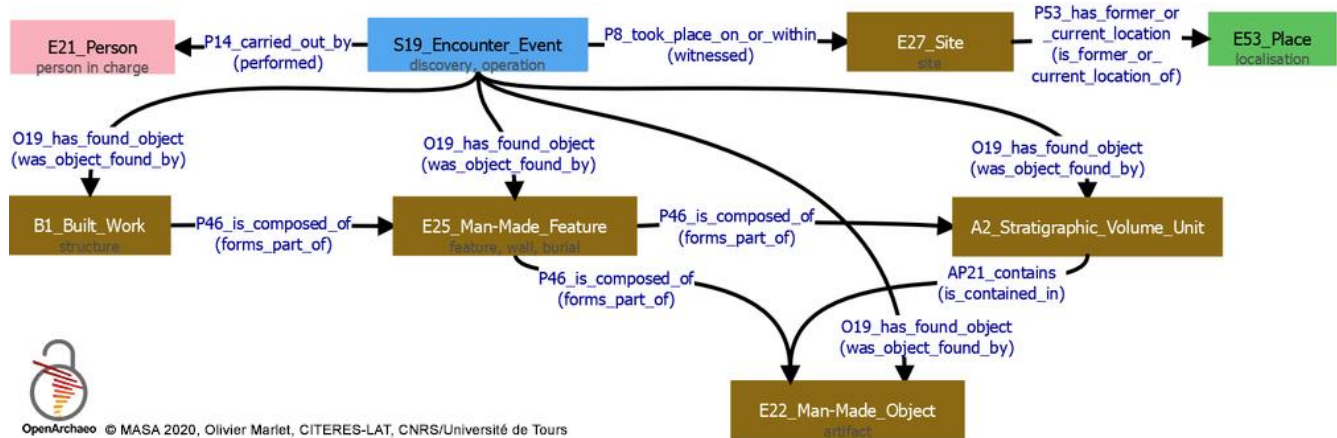


Fig. 2 : Main entities from CIDOC CRM used in the OpenArcheo generic model for archaeological data

The two main entities are the site (E27) and the associated operation (S19_Encounter_Event), whether it is an excavation or a discovery. With CIDOC, a distinction is made between the Site, as a physical element (that is, the remains), and the place where it is located (E53_place). The Encounter Event constitutes an important node in the model because this event can be placed under the responsibility of a researcher (himself part of an institution) and it can be dated. Above all, Encounter Event is the context for the archaeological discoveries.

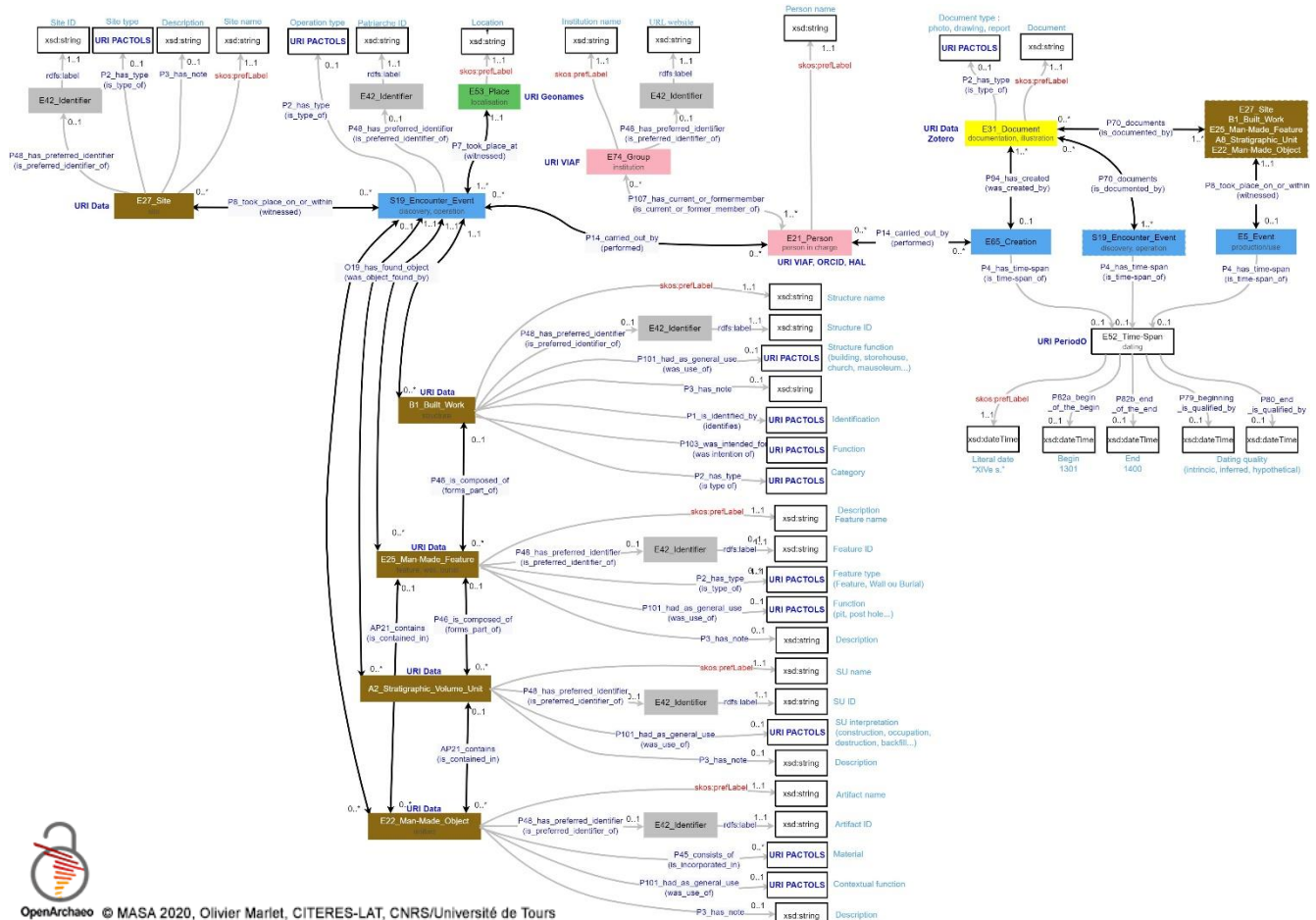
For the Stratigraphic Unit, we take advantage of the existence in the CRMarchaeo extension of the perfectly adapted concept A2_Startigraphic_Volume_Unit, which can contain artifacts. The archaeologist groups some SU according to spatial and functional criteria in order to identify pits, walls, burials, etc. These concepts match with E25_Human-Made_Feature, the traces left by Human. The archaeologist can then group these features into larger functional groups (necropolis, buildings, fortified enclosure). These structures match with B1_Built_Work of the CRMba extension.

Finally, the artifact is another essential entity in archaeology, especially insofar as it can help to interpret the context in which it was discovered and possibly to date it. The model therefore enables to identify the functions but also the materials. This artefact is mapped using E22_Human-Made_Object.

For each of these entities, we can associate a title or a label, and above all an URI, which enables to identify the resource without ambiguity on the Web (Fig 3). These sustainable identifiers concern both the resources themselves and the instances linked to gazetteers: GeoNames is used for locations; VIAF for individuals or institutions; PACTOLS multilingual standardized thesaurus for thematic types.

Each entity can be associated to documentation (E31_Document). Each document is typed (photo, drawing, paper record, book) and can include metadata elements on the author and the date of creation of the resource. Each of these main entities can be associated with a datation (E52_Time-Span). This datation is detailed with a literal datation and dating boundaries (beginning and end), in a necessarily numeric version to be exploited by a computer.

All databases do not use all of these concepts, depending on their scale: a dataset inventorying sites on a regional scale will of course not specify the SU, or even the artifacts. The site and the operation are the common denominators of the large part of the datasets. Here are two examples. You can see that for the field recording database we use every entities of the model while for an artifact inventory far fewer entities are involved.



However, on these same datasets, it is perfectly possible to consider a more detailed ontological modelling thanks to the CIDOC, but with another objective.

This second part takes place in the context of the [SESAMES ANR project](#) (Semantization and Spatialization of Multi-scale Heritage Artifacts : 3D annotation, sonification and formalization of reasoning). In the framework of this comprehensive interdisciplinary project regarding “multiscale built features”, it specifically focuses on the means to confront and combine two different views of the same heritage items: the architects’ view and the archaeologists’ view.

To do so it associates the teams of 3 research labs

- a team of specialists of architectural heritage (from the MAP laboratory)
- a team of archaeologists (from the CITERES-LAT)
- and a team, from the LIFAT, of computer scientists specialized on knowledge representation.

Our test corpus is the guesthouse building of the [Marmoutier Abbey](#), near Tours. With, on one hand, the excavation data that is stored in [ARSOL](#), the LAT's archaeological Database. And on the other hand, the architectural description of the elements of this partially still standing building by the MAP. Concerning these datasets, our practical goal is to conceive an ontology which can query and retrieve the knowledge produced by both archaeologists and architects about this same building. To formally describe the archaeologists' excavation data, we made the choice to rely only on the CIDOC CRM. Because with all its extensions, we thought it was a comprehensive enough framework to express all the archaeological knowledge we wanted to. So leaning on the work done for the OpenArchaeo Model, we extended it to represent finer-grained concepts: that is to say, to specify that the items being described are architectural elements part of a still standing building. And also to include the data relative to their chronological analysis.

How to Apply the CRMba to the excavation data of a still standing building?

The first step of our work was all about using the CRMba, a CIDOC extension that completes the CRMarchaeo and documents Archaeological Buildings. We exploited the CRMba for its ability to link a stratigraphic unit to an architectural element and represent how these elements are associated with each other inside a building (Fig. 4).

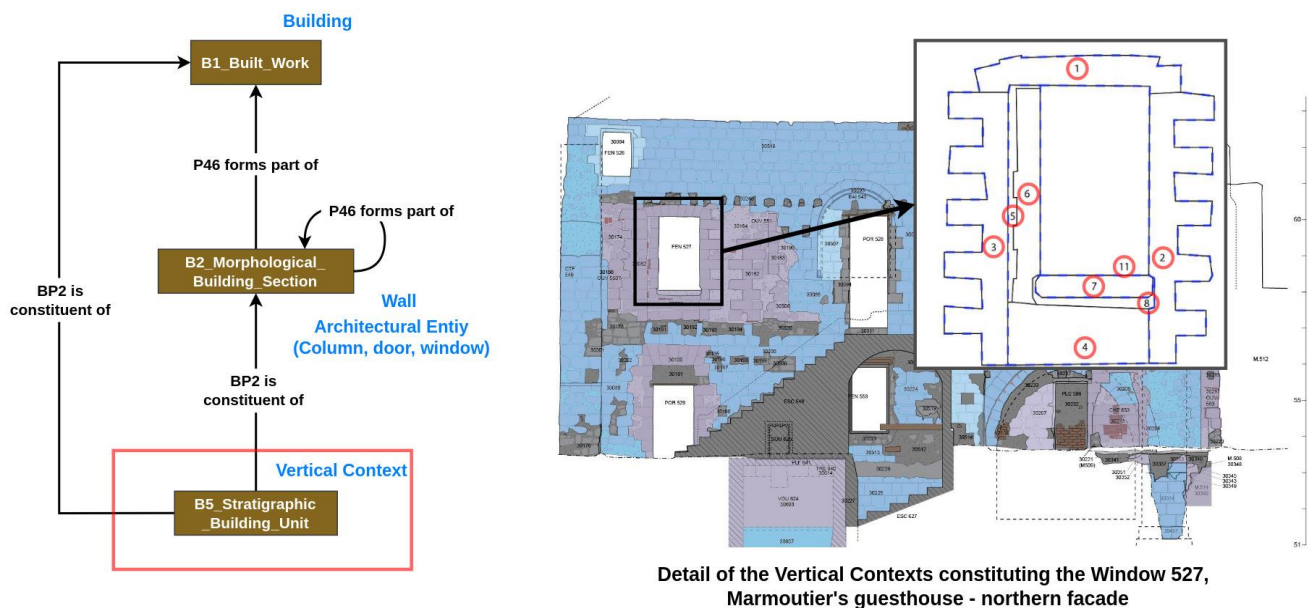
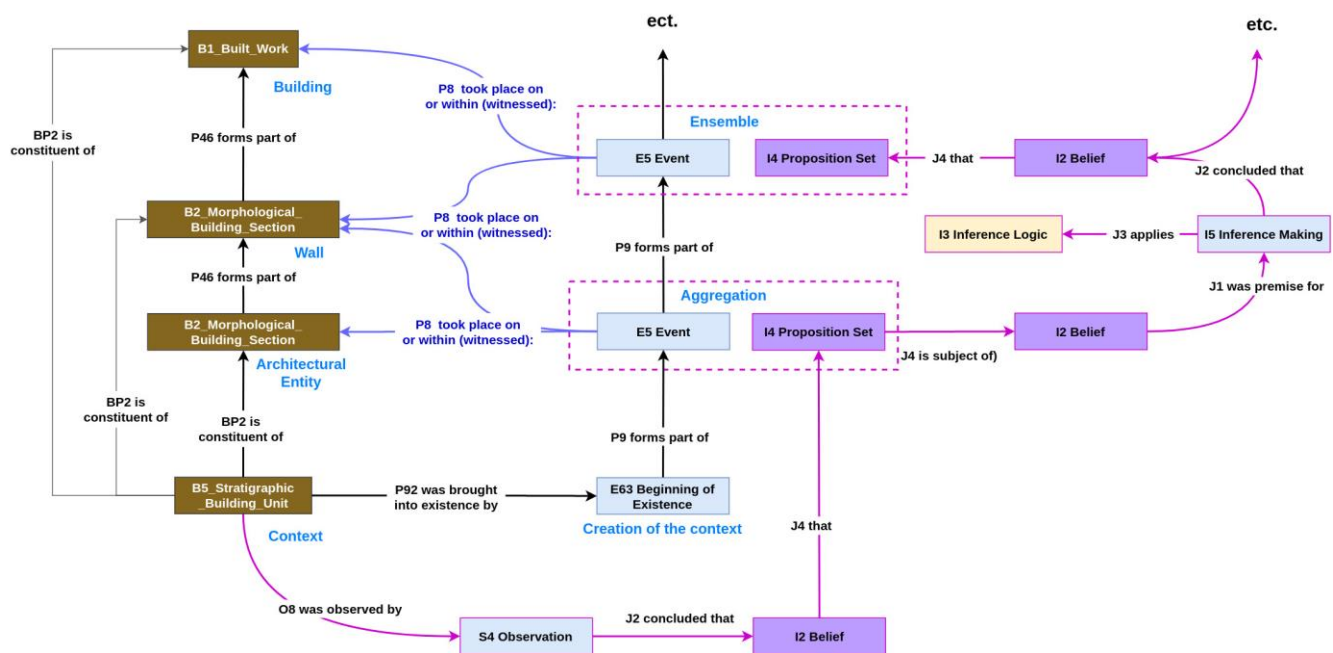


Fig. 4: The vertical contexts, B5_Stratigraphic_Building_Units

First, the CRMba gave us a class to represent the vertical stratigraphic units (B5 Stratigraphic Building Units), and to formally distinguish them from the horizontal ones, which are also present in our corpus. When grouped up, these stratigraphic building units constitute a functional architectural element, represented here by the class "B2 Morphological Building Section" (Fig. 5).

How to represent the chrono functional grouping of the contexts?



Representing an individual creation event for each stratigraphic unit enables us to model their gradual grouping according to a chrono functional logic, just like a Harris Matrix being progressively synthesized. This model presents the hierarchy of these ever spatially and chronologically larger groups of contexts, modeled as imbricated E5 Events, then E4 Periods.

In order to accurately represent the status of this data, we decided to model the interpretative dimension inherent to this grouping process. We did so through a path of CRMsci and CRMinf entities which represent the inductive reasoning leading to grouping of the contexts. The observation of several contexts leads to the belief

that their creations were close in time. So, their corresponding creation events can be regrouped to form a larger coherent event. And in turn, the assertion of this event becomes one of the premises leading to the recognition of a larger coherent event, etc.

This modeling proposition remains for now limited to the most basic inductive reasoning: it's only here to document the nature of some information. But it could be further developed, to provide more details on the process, or to include other forms of reasoning and other data sources which can be involved when establishing the most global phases.

Conclusion

We have seen two examples of modelling with two distinct purposes in the field of archaeology: a generic model for a referencing service of various resources; a more detailed and extensive model with a very specific research objective. In both cases, the use of CIDOC CRM and its extensions proved to be very effective despite the complexity of the ontology.

For these projects, we benefited from the work carried out by the European ARIADNEplus program to which we contribute. For SESAMES project, when we shared our attempts with the working party of ARIADNE in charge of the conception of an Application Profile for Excavation Data, we realized that the representation of processes of archaeological analysis and interpretation is a trending research topic in the context of data integration projects. So, we encourage everybody interested by the subject to get in touch with the ARIADNE Application Profile Working Party. For us, we will follow with great interest the presentation on the modeling of interpretations taking place later this morning.

Katsianis, Markos, Kostas Kotsakis and Filippos Stefanou. 2021. "Reconfiguring the 3D excavation archive. Technological shift and data remix in the archaeological project of Paliambela Kolindros". In Journal of Archaeological Science: Reports, Volume 36. <https://doi.org/10.1016/j.jasrep.2021.102857>.

Marlet, Olivier, Thomas Francart, Béatrice Markhoff and Xavier Rodier. 2019. "OpenArchaeo for Usable Semantic Interoperability" In Proceedings of First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH) co-located with the 31st International Conference on Advanced Information Systems Engineering (CAiSE 2019), edited by Antonella Poggi. Rome: Sapienza University of Rome. <http://ceur-ws.org/Vol-2375/paper1.pdf>.

Ronzino, Paola, Franco Niccolucci, Achille Felicetti et al. 2016. "CRMba a CRM extension for the documentation of standing buildings". Int J Digit Libr 17. <https://doi.org/10.1007/s00799-015-0160-4>.

