On using the CIDOC CRM to model archaeological datasets

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

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. OpenArchaeo generic model for archaeological semantic data.

The MASA consortium has developed a SPARQL query platform for archaeological datasets, named OpenArchaeo (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 OpenArchaeo 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.

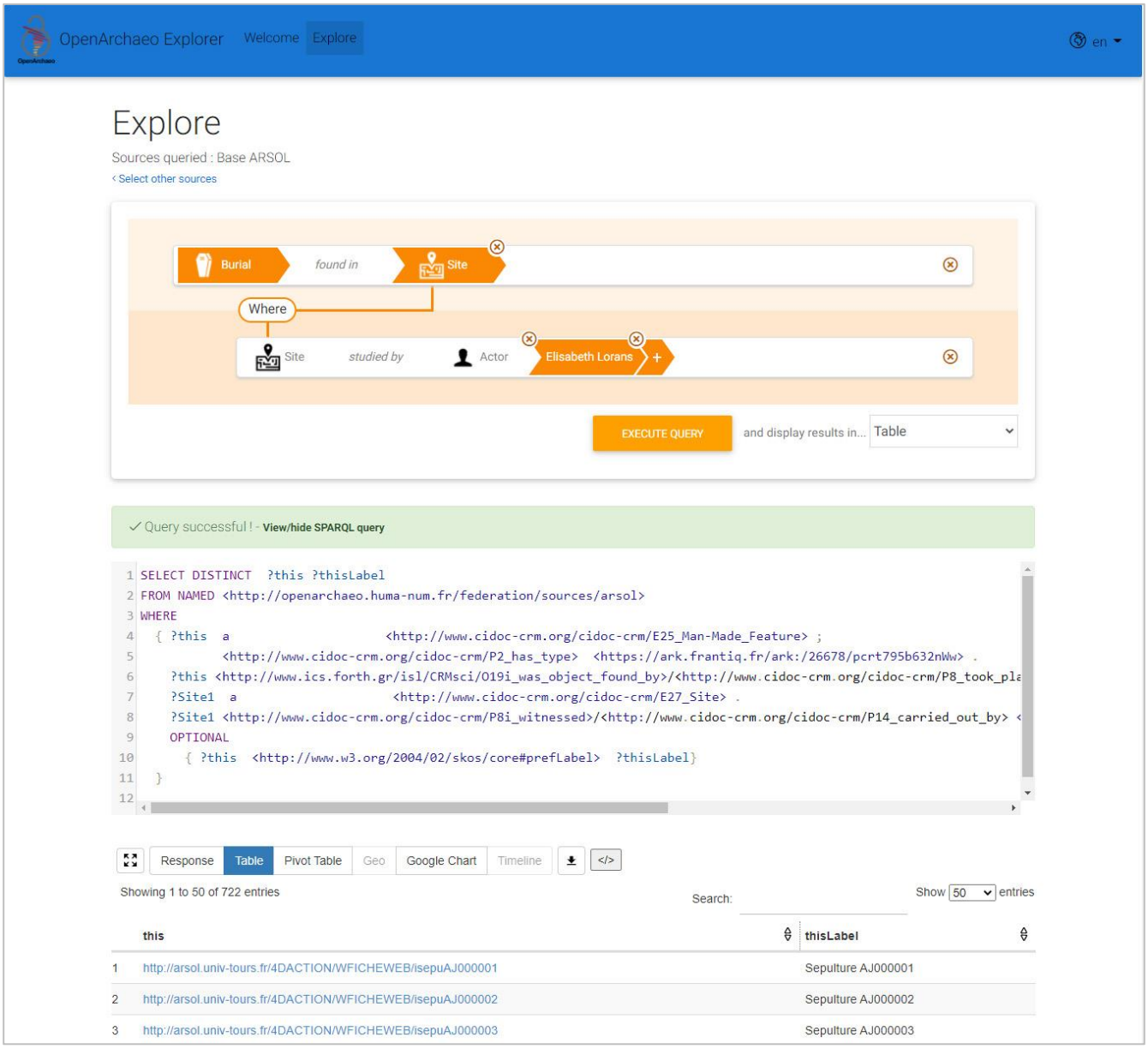

Fig. 1: OpenArchaeo, a semantic Web platform for archaeological data.

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).

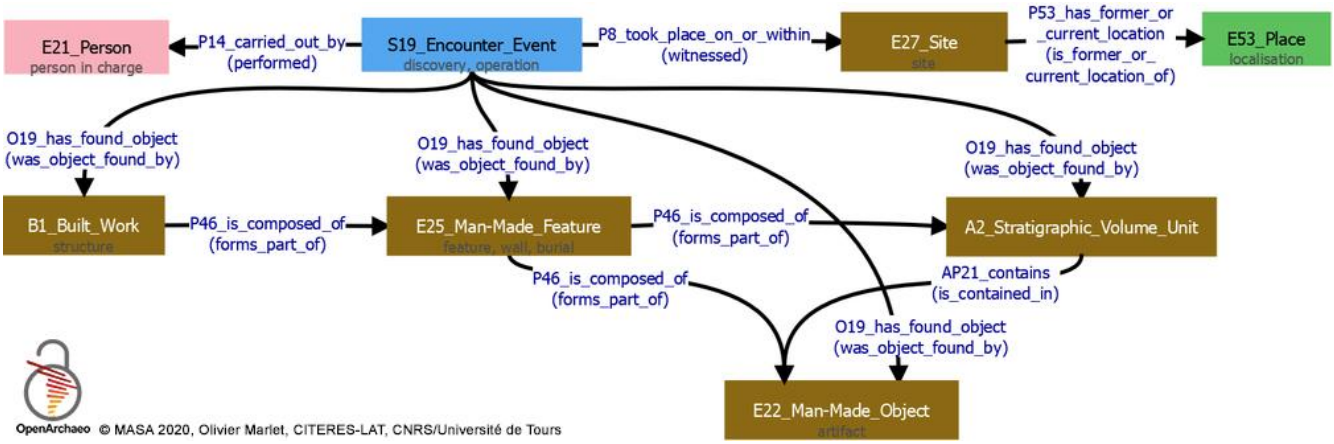

Fig. 2 : Main entities from CIDOC CRM used in the OpenArchaeo generic model for archaeological data

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.

The OpenArchaeo model is resolutely generic in order to facilitate federated searches within various sources. Access to more specific information is reserved to the online datasets publishing these resources.

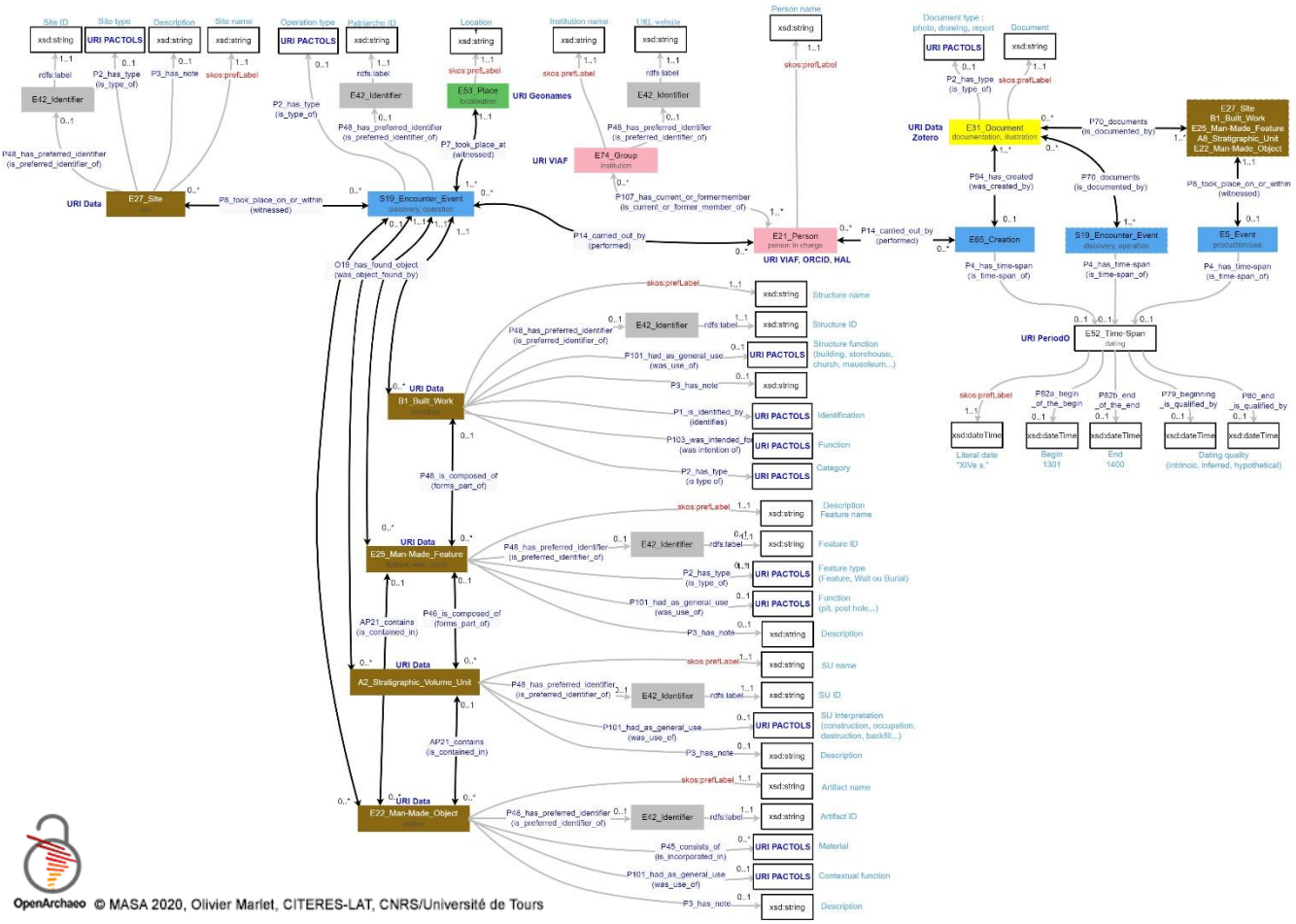

Fig. 3: OpenArchaeo generic model for archaeological data.

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

2. Modelling together architectural and archaeological views

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).

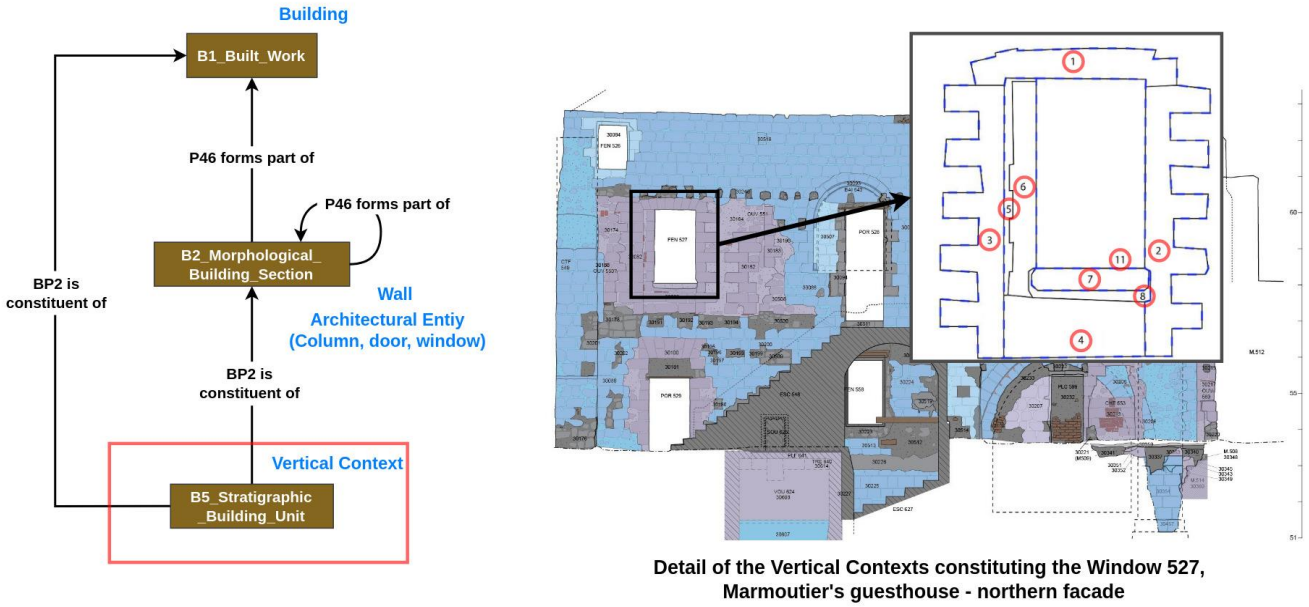

Fig. 4: The vertical contexts, B5_Stratigraphic_Building_Units

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).

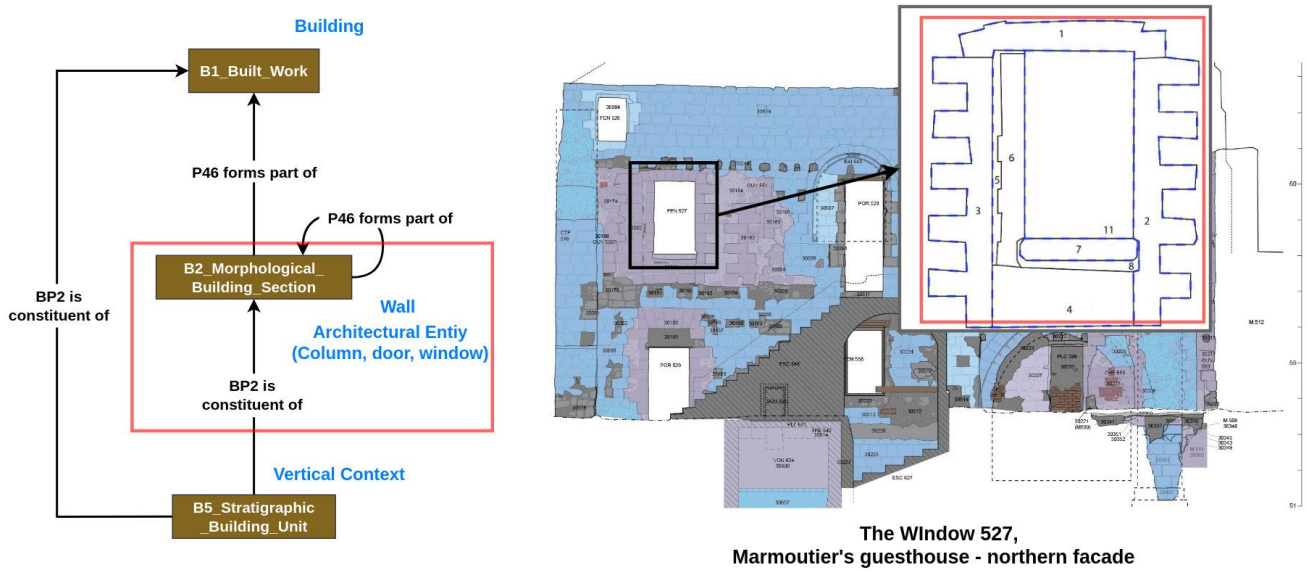

Fig. 5: Contexts constituting Architectural Elements, B2_Morphological_Building_Sections

We use this B2 Class to refer to any type of architectural element that can be found in a Building: a window, an arch, a staircase, and even a wall.

However, since those architectural elements can have different spatial scales, some of the smaller ones will be contained by the larger ones. On the illustration on the right you can see that in fact, the wall contains them all. So, we choose to imbricate several instances of B2, using the P46 property forms part of to indicate how some architectural elements can be part of a larger functional element (Fig. 6).

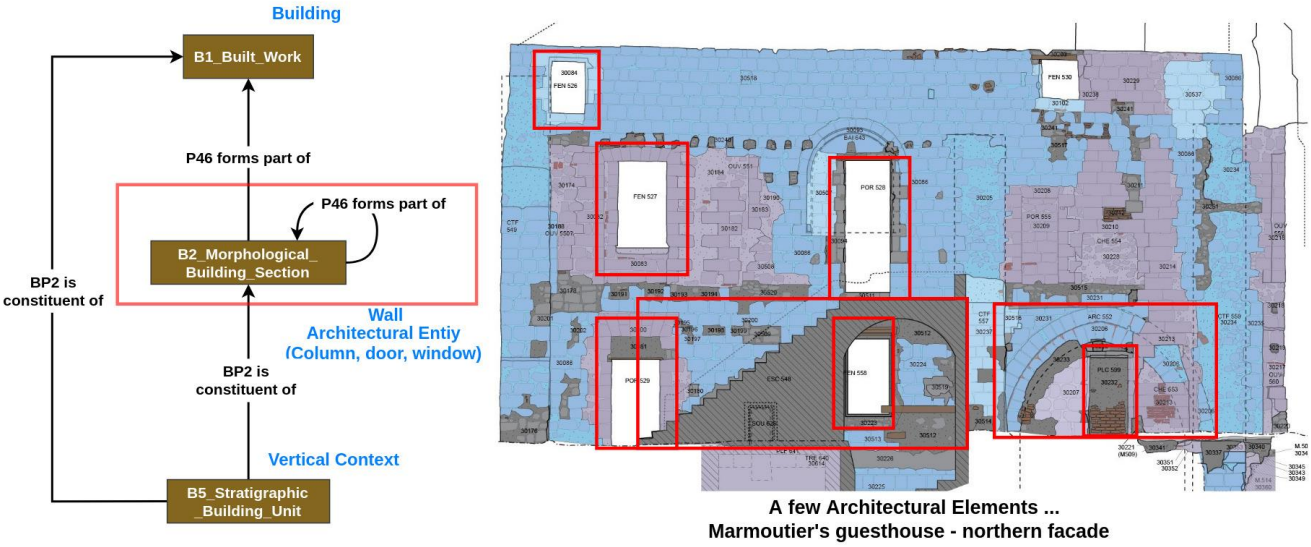

Fig. 6: B2 Morphological Building Sections represent elements of any scale and type

By doing so, we designed a generic model, but a modular one, and that was what we were aiming for. We only introduced the couple of archaeological concepts we needed. The rest will be further specialized, with the architects' finer description.

How to represent the chrono functional grouping of the contexts?

Another invaluable piece of archaeological knowledge we had to represent is the chronological analysis of the horizontal and vertical stratigraphic units. Because of the endurant / perdurant fundamental distinction within the CIDOC, all the events affecting physical objects are modelled as separate temporal entities, beginning with the moment of their creation. So, if we want to locate the stratigraphic units in time and chronologically order them, we can do it through the events representing “the beginning of their existence”. The CRMba and the CRMarchaeo provide “shortcuts”, simpler ways to model chronological relationships between two individuals. However, they lack obvious solutions to assert more complex associations, and this is what we tried to do (Fig. 7).

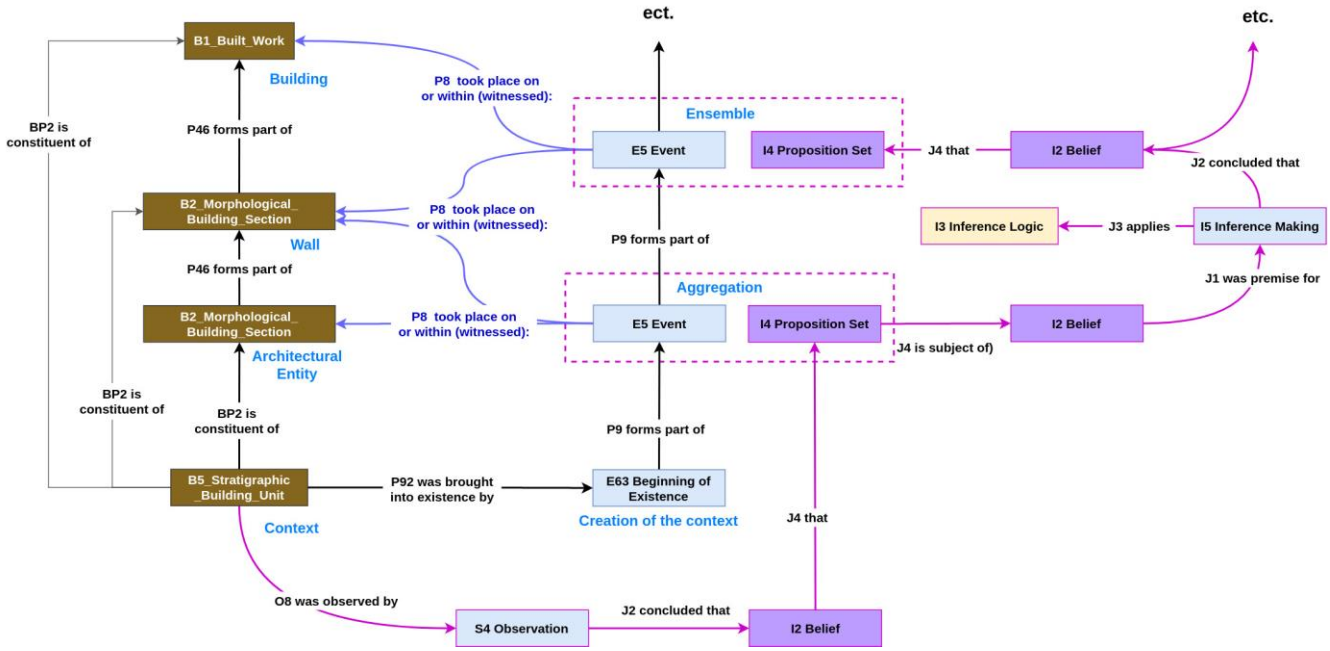

Fig. 7: Representing the interpretative dimension of grouping

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.

The first level could for example regroup the creation events of a lintel, a sill, and other pieces of stone frame, all grouped up into a larger event: the construction of a window. Next, this window construction could be declared to be a part of a larger stage of construction. And this stage would in turn be part of a certain phase of the history of the building. The P8 property in blue is a shortcut to remind more easily to which architectural elements these defined events apply to. These groups provide a very detailed description of the chronology of the building’s elements, but they do not represent factual past events: they are only the interpretative grouping of the traces that remain.

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.

Huma-Num MASA Mémoires des Archéologues et des Sites Archéologiques ARIADNEplus Sesames imar cnrs PRISM citeres UMR 7324 LIFAT ANR PROJET FINANCÉ PAR L'ANR PROJECT FUNDED BY THE ANR