Ontological Modelling of the work of the Centre for Archaeology

September 2004

Paul Cripps, Anne Greenhalgh, Dave Fellows, Keith May, David Robinson

1. Introduction

This project developed directly from work carried out as part of the Centre for Archaeology's (CfA) Revelation project (Cross, 2004). The assessment stage of Revelation included several pieces of work that investigated, analysed and reported on the existing state of the data systems and the inter-relationships between various data resources constituting the information management systems of the CfA. These included a review of the existing systems (Cross, 2004 Appendix A) and the production of data-flow diagrams (for explanation of terms see 6. Glossary) and entity relationship models to represent how information is collected, managed and distributed by CfA staff in their work.

The resulting picture showed the CfA systems as a rather disparate grouping, or 'archipelago', of diverse, specialised, but rather isolated and independent information systems and databases. In many cases, due to their age, these systems do not have very clear mechanisms to enable the sharing of data either between the different data 'islands' within the CfA or with the outside world. Another outcome of this initial work from Revelation was the recognition that, whereas the conventional modelling work had proved quite successful in revealing gaps existing between systems, it did not readily enable the modelling of likely solutions, i.e. how the information held in different systems could be shared.

What was needed was an approach to modelling which would produce a more conceptual overview of all the information being created. Such a model needed to include how existing data items would continue to be represented. But it should also show the conceptual relationships that pertained between data, thus allowing construction of a more complete picture of how all the data fitted together. It was at this point that the idea of using an ontological approach to modelling was considered and attention turned to the International Committee for Documentation of the International Council of Museum's Conceptual Reference Model, in short the CIDOC CRM (Crofts et al..2003) as a tool for producing an ontological model.

An ontology is an explicit formal declaration (with a standardised vocabulary) of how to represent object concepts and other classes assumed to exis t in some area of interest (a domain) and the relationships between them. In this sense an ontology is a specification of a conceptualization.
Ontologies provide a shared and common understanding of data and, in some cases, services and processes that exist within a domain (in this case the Centre for Archaeology). This facilitates communication between people and information systems and an enhanced ability to search for information across different knowledge repositories. The common understanding allows mapping of the concepts within an ontology to information and processes within the organisation being represented. Using the terms defined in ontologies enables application designers to understand fully the meaning and context of the information being modelled. This helps represent data in a meaningful and consistent way, enabling better integration of data across applications.

The CIDOC CRM ontology is an emerging international standard, created in the first instance for the museums world (via ICOM), but which has found applications across the broader heritage sector. As the CRM ontology is event-based, rather than data-driven, it appeared well suited to modelling the core of the archaeological process, by which archaeologists attempt to record and document the results of past events through a series of events or activities in the present.

1.1 Aims and objectives

When constructing an ontological model it is important to decide and define clear boundaries for the domain that is to be modelled. The scope of this project was established at an early stage to be the information domain of the archaeological work of the Centre for Archaeology. The main aim of the project was to define a conceptual framework for this information domain. The specific objectives were:

- Produce a flexible, open and readily extensible model of the information philosophy at the CfA as a basis for systems development via the Revelation project.
- Facilitate discussion and continued development of CfA information systems by achieving a common understanding and the definition of a shared language.
- Ensure that the ontological model incorporates the knowledge management embodied in existing systems while enabling the required improvements to be modelled.
- Identify where existing terminologies and word lists are currently used and to inform where standard terminologies can be used or will need to be defined for systems development.

This report sets out the main results and the methods used to achieve all the above objectives.

2. Methods

This section explains the overall methods adopted and gives details of the different techniques used for modelling.

2.1 Overall methods

A small project team carried out the bulk of the work. This team consisted of a project manager (KM) and a principle investigator (PC) who was responsible for the CRM data modelling, along with two other project team members from the archaeological and scientific teams of the CfA respectively (DF & DER). An external consultant with archaeological and systems design experience (AG) was recruited to carry out the bulk of information gathering about the domain through interviews with CfA staff, and also gave support for the CRM modelling and produced the UML (Unified Modelling Language) model. At the outset, there was a need to provide introductory training for the project team, as none of them was familiar with the CIDOC CRM. In addition, a more general introduction had to be given to the rest of CfA staff. Introductory presentations were delivered by Matthew Stiff (EH Data Services Unit (DSU)) and a couple of workshops on the CRM were held with the rest of the DSU. The project also benefited from consultancy input from two CIDOC CRM experts. Steve Stead ran three separate workshops for the project team, initially to give training on using the CRM (see Appendix G) and later to provide feedback and discussion on the developing models. As part of the final verification of the model Martin Doerr was engaged in the role of professional referee to review and comment on the final drafts of the CRM model diagrams and textual explanations (see Appendices A, B and C).

The overall methodology used for the CfA's ontological modelling project derived from other generally used approaches to ontology building. The principle is to encapsulate the broad concepts used by Domain Experts in their work. As such, the ontology development and modelling is not driven by Information Science specialists but by experts in the domain being modelled. In this case the Domain Experts were CfA staff and the Domain covered was the CfA archaeological process.

Sources	***Contacts***
Environmental archaeology	David Robinson, Polydora Baker, Andy Hammon, Gill Campbell
Geoarchaeology	Jen Heathcote, Gianna Ayala
Conservation	Karla Graham
Geophysics	Andy Payne
Computing	Brian Attewell
Archiving	Claire Jones
Survey	Tom Cromwell
Finds	Joern Schuster and comments by Sarah Jennings
Administration	Mary Walkden
Scientific Dating	Peter Marshall, Derek Hamilton, Amanda Grieve
Graphics	Eddie Lyons
Archaeologists - workshop	Brian Kerr, Tony Wilmott, Sarah Cross, Tom Cromwell, Dave Fellows, Joern Schuster, Vicky Crosby , Sarah Reilly, Jon Last, Fachtna McAvoy.

The CRM uses an object-oriented approach and defines a relatively small number of object Classes (i.e. global concepts) and Properties (roles & relationships between classes). These classes of objects and their associated properties and relationships are the building blocks that can be used to describe formally a particular knowledge domain and to model explicitly the less easily represented semantic relationships that exist between the different classes used by that domain. The CRM currently contains a listing of approximately 80 declared Classes, denoted by the pre-fix 'E' followed by a numeric identifier and the appropriate conceptual entity, for example, <<*E39 Actor(s)*>>. There are currently about 136 Properties, identified by a numeric identifier preceded by the letter 'P', for example <<*P14 carried out by (performed)*>>*.* See Appendix F for terminologies.

The broad method for building an ontological model for the Domain can be summarized in the following five main stages (Denny, 2002):

2.2 Project methodology

Acquire domain knowledge

The limits of the Domain to be modelled were defined as the archaeological work of the CfA. Crucially, this meant we were not trying to map all archaeological systems to the CRM but rather focusing specifically on work carried out by the CfA. Acquiring Domain knowledge principally meant holding discussions and interviews with CfA staff and collecting information on all available systems and procedural documentation and then collating what was relevant. Following consultation with CRM experts, it was decided not to tackle details of areas such as project management and administration which are business processes that other types of data modelling could cover more appropriately. A decision was also made at an early stage to model the existing data sources using Unified Modelling Language (UML) diagrams as an aid in explaining to CfA staff members how their specific data can be represented and how it relates to other data entities.

Organize the ontological model

This requires two basic operations:

- Identify global concepts (Classes) that best match the data being created.
- Identify the Properties (the roles & relationships between the classes).

The CRM itself does not contain specific methods for how to go about representing formally the Classes or Properties, although the models that are given as examples in the CRM were drawn up using the TELOS data model and there are some mapping tools available on the CRM website (http://cidoc.ics.forth.gr/tools.html). For practical reasons, and to enable shared use around the CfA, most of the project diagrams were drawn up using EH corporate standard Windows-based graphical and spreadsheet software such as MS Word, Visio and Excel.

Flesh out the ontological model

Graphical representations of the various models were needed to help explain the modelling within the project team and to CfA staff whose data we were attempting to depict. The CRM uses mapping statements composed of text string triplets in the form of Class – Property – Class (for example, *<<E17 Type assignment>> – <<P14 Carried out by>> – <<E39 Actor>>*). It is very difficult to do this directly from real life interviews without having some form of intermediary diagram to embody a common understanding between the person trying to model and the person explaining their use of data. Draft representations of the CRM and UML models were therefore created using Excel or Visio and texts were documented in Word, with final versions being saved as PDFs. In addition, text-based descriptive documents were created to give a more detailed description of each Class and Property and to show their relationships as depicted in the CRM diagrams. Attempts were made, with variable success, to reach a comparable level of granularity (i.e. equivalent degrees of resolution and detail) across the model so that each of the main information areas within the CfA could see their activities defined. Some areas of the model, in particular the context recording system which was central to many aspects of the CfA's process, were developed in more detail.

Check the work

Considerable revision and re-working of the models took place on the basis of a number of group discussions with Domain Experts, workshops and feedback from CRM consultants and by simply checking and re-checking the drafts of diagrams and texts with the CfA data users themselves.

Commit the ontological model

The final version of the model will first be verified by CRM experts (Martin Doerr and Steve Stead) and then disseminated to a wider audience within the CRM and archaeological communities. Further plans for publication and dissemination will… (*to be agreed as part of a dissemination review at the end of the ontological modelling project)*. Although the product is primarily a CfA-based model, it is hoped that the core, dealing with the archaeological recording system, may find broader usage, where appropriate, in the wider archaeological community.

2.3 CRM modelling methods

The usual approach to working with the CRM is to take a well defined data model, generally extracted from existing database structures, and map data items to CRM entities. Unfortunately, very few of the systems in use within the CfA have suitable design documentation to enable this and many 'systems' are not computerised or rely heavily on manual input. As a consequence, a slightly different approach was adopted.

The initial intention was to take the results from the Review of Existing Systems produced as part of the assessment stage of the Revelation Project (Cross, 2004), supported by a first round of interviews with members of staff, in order to gather enough information to produce a series of draft models. These models could then be taken round to CfA staff in an iterative process, refining and enhancing them to capture additional detail and check for misinterpretations. It soon became apparent that for this process to work, both interviewer and interviewee needed to be familiar enough with the CRM for them to discuss their work in terms of CRM constructs. Accordingly, the initial interviews were used to collect notes and produce draft diagrams without using CRM constructs. This resulted in a series of UML diagrams representing detailed aspects of the information domain. The process of capturing detailed Domain Information using UML techniques is discussed later (section 2.4).

The next step was the compilation of an overall model built on these UML diagrams to present CfA concepts using CRM entities and properties in a graphical form. This allowed for ease of understanding of the model as it developed, a graphical representation being much easier to work with than a list of mapping statements. The event-driven nature of the CRM also facilitated the identification of gaps in our understanding of the Domain; where objects exist, they must be the product of an event, hence if there are objects without associated events, there must be events missing from the model.

Once this diagrammatical document had been circulated, it became apparent that the intent and semantic clarity embodied in the model through the work of the project team was not presented clearly enough. To resolve this, a secondary supporting document was produced in the form of a table describing salient points relating to particular entities and properties. A central feature of this document is a statement outlining each concept as used in the model, similar in form to CRM scope notes.

2.4 CRM statements in Excel

Although the diagrammatic representation of the model was important for an overview it became quite large and awkward to represent in an easily readable form. It was found very useful to transpose the statements represented in the diagram into a text-based form that was more manageable and made it easier to present the details of particular sections of the process.

This format uses an Excel spreadsheet to set out each of the principle entities in the diagram. Each horizontal line in the spreadsheet shows the class-relationship-class triplets that go together to make up the chains of statements as depicted in the diagram. This format is based on a template produced originally by Steve Stead for mapping MIDAS to the CRM as part of the FISH Tools project. The MIDAS project involved mapping of specific existing data fields to the CRM. The CfA diagram depicts a more conceptual level that includes some specific data items alongside the conceptual framework for additional information items. As such, the statements depicted in the spreadsheet do not necessarily map directly to existing data items.

The structure of the statements model developed as the different elements of the CRM diagram were constructed. In producing the declaration, an attempt was made to structure the statements in a broadly temporal sequence according to events as they occur in the archaeological process. However, due to the iterative nature of archaeological recording and analysis this process cannot be represented by a simple chain of events.

2.5 Using the CRM ontology with UML

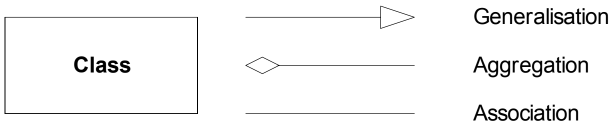
The CRM uses an object-oriented modelling technique which represents knowledge in the form of a conceptual model, and leaves the details of implementation (i.e. systems design and software solutions) to the platform-specific phase of the development. A platform is the technology on which a system runs, e.g. Windows, Unix, ORACLE, Access. In order to determine whether the CRM can be used to model archaeological systems, another object-oriented modelling technique, UML was employed. UML is a standard tool for object-oriented modelling within the Information Technology sector and is particularly suitable for taking the object-oriented concepts inherent in the CRM and adding more detail so that the model can be used as a basis for implementing new systems.

The CRM ontology has been used to constrain the way in which UML is applied. This means that the CRM ontology has been used to create UML profiles for the archaeology domain and constrains the use of UML in line with the set of definitions that make up the CRM. This allows us to see the build-up of UML patterns that can then be used as standard throughout the CfA, and EH if desired.

Seen from a different angle, we could be said to have instantiated UML, a purely representational language, with the CRM, which refers to real world concepts. We have then constrained the CRM to a subset of its range, the CfA domain.

When new systems are designed the patterns discerned from the ontological model can be applied to the new systems, allowing joining up of components of defined functionality. The archaeological systems are depicted as a conceptual model in UML and the depiction is more applicable to direct systems implementation than that of the CRM.

UML has a number of diagrammatic tools that can be used to represent Domain Knowledge. The UML modelling technique used here is the Class Diagram that depicts a static (i.e. data) view of the system and shows the commonality between objects (Classes) within the system. This relates effectively to the object-oriented technique used in the CRM. The CRM has declared (i.e. defined) a number of Classes with Properties specific to those Classes, which represent the relationships between them, and constrain their use, thus making this more precise. The Properties defined in the ontology have been used to name relationships between Classes in UML models. Each Class within an object model has defined behaviour and data. Classes collaborate with each other and with external actors to fulfil a function. The Class collaborations are shown by relationships between Classes. The various components within the UML Class diagram used in the CRM mapping are shown below.



Class
Within an object-oriented view of an “enterprise” (e.g. the CfA), classes are declared that encapsulate some piece of required data and behaviour within a system. The CRM declares classes that define knowledge within the Cultural Information Domain being studied (i.e. the CfA). In order to utilise the concepts modelled in the CRM, the UML model uses the CRM ‘Classes’ as Class stereotypes, thus mapping the behaviour of the object in the CRM to the real world representation as gleaned from investigations of current practices, in this case the CfA staff interviews etc. The UML concept of stereotype lets us use the behaviour defined for each Class in the CRM within the UML model, thus reflecting the CRM behaviour within the UML model. The use of the word stereotype in this instance is specific and means that all the behaviour and properties, as defined in the CRM, are inherited by the Classes in the different models. The stereotype is used to map Classes between the CRM and UML models and is shown in the following way:

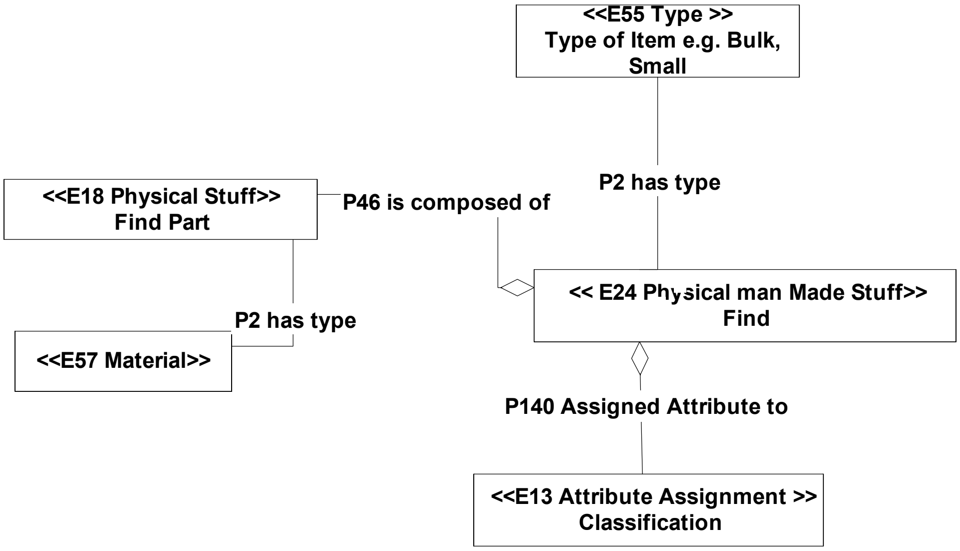
<<CRM Class Name >>
UML Class Name

The CRM Class name is held between the brackets and indicates that the Class Name below it shows all the behaviour declared for the CRM Class Name. Stereotypes are used to extend the UML notational elements, to give more detail in specifying associations, inheritance relationships and classes. In essence a stereotype is a form of inheritance in the metamodel.

Relationships
The lines show the relationships between Classes, and they are often named in order to clarify these relationships. Within the representation of the CRM-based model in UML, the properties defined in the CRM are used to name the relationships, thus making the specification more precise.

Generalisation
This indicates sub-typing. Referring to the Finds Example given below, a Weapon is a kind of find, and inherits all the (general) information - processes and data - associated with a find. Generalisation has not been used extensively in the UML models. Instead the CRM concept of <<*E55 Type*>> has been used together with the <<*E13 Attribute Assignment*>> (or one of its CRM sub-types) to model the process and the data. Where sub-typing is part of the CRM, and adds clarity to the model, the stereotype from the CRM is used. This is the case within the section on Bone Recording, where a sub-type <<*E20 Biological Object*>> of <<*E18 Physical Stuff*>> is used to show the concept of bone.

Finds Example



3. Research results

3.1 CRM modelling results

The CRM modelling exercise resulted in the production of a high-level, conceptual model of the CfA information domain. This was presented in the form of a diagram with an accompanying descriptive text. The model is object-oriented, event-based and makes extensive use of object-oriented techniques such as class inheritance, polymorphism and stereotyping. Class inheritance implies that all instances of a Class inherit all properties of that Class. Polymorphism implies that a Class can inherit from multiple Super-classes. For example, a bus can be seen as a “diesel-engined vehicle” and a “passenger vehicle” at the same time. Stereotyping involves using one Class of objects as a template for another. All properties are defined as being optional and repeatable, according to the CRM specification. Additional detail is provided by the UML models for individual sections of the CfA Information Domain.

In terms of content, the model is remarkably simple. Two groups of events have been identified: one represents events that happened in archaeological time, resulting in the formation of the archaeological record as excavated, and the other represents events carried out by archaeologists in order to document and make inferences about the first group. These two groups of events are related by way of the place in which they occur and by any physical remains existing at that place. Fig. 1 shows this high-level relationship between events, including mapping to research questions and formal procedures. Mapping in this sense means making an explicit link between one set of concepts and another, for example between this model and MAP2 Assessment. If we take the first group, it can be seen that it is through the sequence of these events that the archaeological record is formed. The group includes events which lead to the formation and transformation of contexts, to the production, use and deposition of finds and to the construction, use and disuse of structures. It is by means of the temporal relationships between context formation and transformation events that the stratigraphic sequence is formed, the stratigraphic sequence being the cumulative result of all formation and transformation events. It is through events relating to the production, use and loss of finds that we can build up a picture of the dating for a site. Broader scale events enable us to build up an overall sequence, i.e. “phase” the site. Features and finds preserved in the archaeological record are exclusively the products of such events.

Taking this in a bit more detail, what the model has done is to relate a number of pieces of disparate information within a common, event-based framework. The fieldwork recording system currently records which context a find comes from. A specialist finds assessment may indicate that the find was produced within a certain timeframe, possibly even at a known location. The finds information can be used to provide a spot-date for the context if the evidence meets a number of criteria. The model provides an explicit path for the relationship between the find and the context, so that rather than storing data as attributes of either the find or the context, it is possible to store data linked directly with the events to which it relates. Fig. 2 is a derived diagram showing finds production, use and deposition in conjunction with context formation and excavation, illustrating the logical path from object production to small find. Furthermore, the stratigraphic record can be seen as the product of the archaeologists' interpretation of the various site formation events, which resulted in the observed stratigraphy. The current CfA recording system is based around this notion, but the events are hidden, it is simply the observed relationships which are recorded, the events being then inferred from the record. For example, we currently record the cut of a ditch as being straigtraphically below the fills of the ditch, implying the action of cutting the ditch occurred *before* the ditch was filled. On a larger scale, matrix compilation may inform us that the cutting of the ditch also occurred stratigraphically before the cutting of another ditch. By referring to the actual events associated with such context formation processes, we can use temporal operators to manipulate the sequence of these events and obtain an understanding of the observed relationships between them rather than simply assigning before/after relationships to contexts. Indeed, given two contexts X and Y, when we state X is stratigraphically above Y, what we are actually saying is the formation event which led to the formation of X occurred sometime after the formation event which led to the formation of Y. Expanding this, it is possible to apply any of the temporal operators (after Allen, see section 6) found in the CRM to reason about these events, providing much greater semantic clarity in terms of resultant documentation.

The second group of events comprises those not related directly to the formation of the archaeological record, but rather to our attempts at understanding it. This group includes such events as excavation, survey, scientific analysis, drawing and the various activities referred to collectively as post-excavation, assessment and analysis. This second set of events can be characterised by the participation of an archaeologist or specialist in some role. In terms of the model, these events allow us to re-populate the archaeological process. Archaeological data is exclusively the product of one (single piece of data), or more (multiple pieces of data), of these events. Events can involve Actors, but these are often invisible in archaeological systems, the only person being associated with the records being the one responsible for data entry. By using activities to represent processes, other Actors in the process are reintroduced.

Following on from this, the process of documenting events in the past using events in the present allows us then to be critical of our own documentation. This is a very powerful construct within the model as it allows us to have multiple accounts of the same situation, and be able to distinguish between them in terms of their validity relative to the available source information. It may be, for example, that two different phasing schemes are produced based on different interpretations of the finds evidence and the stratigraphic sequence. The model supports such multiplicity of views.

A corollary to this is that, via the model, the concepts of analysis and interpretation become central to the recording system rather than being peripheral to it. Each activity carried out by an archaeologist that produces data involves a degree of interpretation, or at least some conscious thought, and this interpretative information can now be captured alongside the resultant piece of data. This allows us to see why certain conclusions were drawn or decisions made at certain times. Where activities were described as analysis or interpretation during the initial data gathering exercises, these have subsequently been broken down into component parts which can be explicitly described in terms of CRM constructs, e.g. <<E16 Measurement>>, << E14 Condition Assessment>> or, at the very least, <<E13 Attribute Assignment >>, used to assert relationships between CRM Classes. We can therefore model the archaeological process in terms of these component activities which then <<P9 form part of >> larger scale analysis/interpretation events>> such as MAP2 stages.

In addition to the various events which form and investigate the archaeological record, there are static entities described in the model with which the events interact. To use the terminology of the CRM, these are predominantly <<E53 Place(s)>>, <<E18 Physical Stuff>> and <<E28 Conceptual Objects>>, but there are also a number of <<E39 Actor(s)>> involved in particular events as well as a variety of data items such as notes, time-stamps, measurement values etc. These entities relate to physical objects found within the information domain such as finds and samples as well as digital objects such as survey datasets. Also included are the physical manifestations of information objects such as the context sheet carrying the documentary record of a context attribute; the context sheet has a physical presence and needs to be managed as a real-world object.

3.1.1 Phasing and grouping

The CRM has been successfully applied to modelling the main archaeological activities undertaken by the CfA. One of the complex aspects of the archaeological post-excavation work that was tackled in the model was the phasing process, i.e. how the context records created on site, which describe the archaeological features, are processed further to produce a coherent narrative of the site’s history and development. This encapsulates the process of assigning contexts to sub-groups and sub-groups to groups, and further assigning these groups to landscape elements or phases etc., up through the phasing hierarchy.

Once the concept of a context has been established and mapped to the CRM, this can be used to feed into the sub-grouping and grouping procedure to create higher levels of understanding. A context can be modelled as either the physical matter that makes up the context <<E18 Physical Stuff>>, or can be modelled as the place where the context existed at the time it was first registered in a way relevant for archaeological purposes i.e. the recorded location of the context <<E53 Place>>.

To allow the creation of the model, the definitions of the phasing processes were first established and agreed. These follow widely accepted concepts and criteria used for post-excavation site analysis.

A context can be defined as the basic recording unit used on site and is usually site-glossary controlled. A sub-group consists of a number of stratigraphically-linked contexts spanning one phase, and linked together by one of the three accepted processual terms of construction, use or disuse. A sub-group may consist of one or many contexts, and sub-grouping brings together the contexts into more meaningful interpretative blocks. The next level in the site phasing hierarchy is the assignment of sub-groups to groups. Groups are formed by the amalgamation of a number of sub-groups that are brought together into interpretative units, e.g. building, structure, open area. Again the groups can be categorised using a processual term – construction, use or disuse – and may consist of one or many sub-groups. Moving up the hierarchy, groups can then be brought together to a higher level of interpretation and can form landscape elements, area-use groupings, etc, and in turn these can be assigned to sub-periods, periods or phases at the highest level of the stratigraphic hierarchy. The number of intermediate stages required for analysis and interpretation depends on the complexity of the site and of the context record, but the CRM modelling of this part of the archaeological process has taken into account all possibilities.

The concept of sub-grouping, grouping and phasing the context record is modelled as shown in Fig. 3 on the phasing section of the overall archaeological processes model.

Each individual context can be classed in the CRM as a place (<<E53 Place>>), as can sub-group and group. Accordingly, they share Properties due to a shared superclass, but there is no inheritance from contexts to groups. The groups may be defined by some shared property of context (e.g. all contexts associated with the construction of a particular section of wall), but properties are not “inherited” from the lower order to higher order entities. When a sub-group is defined, given a meaningful name/label and has contexts assigned to it, then the contexts inherit the higher order meaning from the sub-group. The spatial extent of the sub-group will not necessarily be the sum of the extents of the component contexts, but their spatial extents will be incorporated into the overall extent of the sub-group e.g. segments of enclosure ditches may be sub-grouped together in an ‘enclosure’ interpretative grouping, but the spatial extent of the enclosure could be different to (and greater than) the extents of the component ditches. The use of the CRM Class <<E53>> to define <<Place>> in this instance is a conceptual interpretation, and the <<Place>> defined is not necessarily just the sum of the component spatial references. Indeed, due to excavation techniques, only part of a feature may be excavated and recorded, in which case the full spatial extent is clearly not known.

The sub-group itself can then be assigned to a group higher up in the hierarchy, and in the same way the group can be classed as a place in the CRM.

Context Place (E53: Place)→ P89 falls within (contains) → Sub-group (E53: Place)→ P89 falls within (contains) → Group (E53: Place) → P89 falls within (contains) → Landscape element (E53: Place) etc.

This is known as recursive grouping, using the repeating pattern identified in the model as the basis for each analytical grouping event moving up the phasing hierarchy. This greatly simplifies the section of the CRM relating to the phasing process.

The stratigraphic sequence of contexts from a site is recorded using the sequence of events (<<E5 event(s)>>) that have lead to the formation of the matter found in the contexts in question rather than the <<E53 Place>> or <<E18 Physical Stuff>> that defines the contexts. The Events can be formation or transformation events. The contexts can be modified in Transformation Events (e.g. chemical change), but as the context is not defined spatially until it is observed upon excavation, this does not imply modification of the spatial extent of the context place. Rather the context matter may have changed its location from a previous extent into the current context place as a result of transformation.

The Events are identified by a time-span (<<E24>> and <<P4>>), and using the time-span and the sequence of time appellations to which the contexts are assigned enables the production of the section of the model that covers the site stratigraphic sequence.

Each of the phasing elements (sub-groups, groups etc) is classified by a Name (<<E17 Type Assignment>>). This action is performed by an <<E39 Actor>> in a variety of roles, the most likely in this case being that of post-excavation analyst or stratigraphic analyst. From this string of the model we can establish how the contexts were classified, who carried out the classification and the role in which they were employed.

In CRM terminology the places forming the sub-groups\groups etc. have witnessed <<E5 Events>> that have <<E52 Time-spans>> relating to the sequence of site development, and the CRM has a series of Properties that allow the relative timing of these events to be modelled. These are known as Allen’s Temporal Operators and these can be used to provide relative dating for the phasing sequence.

The events can be classified by <<E55 Type>>, i.e. glossary-controlled values of phase definitions, and these Events (and consequently the phasing elements) are fully documented. The documents can be <<E73 Information Objects>> or <<E31 Documents>> that refer to places (and therefore sub-groups and groups) and these may include such objects as site plans, phase plans, interpretative land-use diagrams, textual summaries, databases etc. Again each of these information objects will be assigned a descriptive Type that refers to (<<P67 refers to>>) the place that is defined by the sub-group or group.

In summary, this section of the model has taken the contexts themselves – the individual building blocks of the site and their recording documentation – and has modelled the events that allow their subsequent inclusion into the higher levels of interpretation occurring during the post-excavation analytical stage of archaeological projects. A great deal of information has been modelled using relatively few processes. This has shown how the recursive nature of the phasing process functions, and how it has been incorporated into the model so that all aspects relating to the type assignment, dating, documentation etc. of the context, have only to be modelled once, with relationships retained in the higher levels of the phasing model.

3.2 UML modelling results

The CfA processes and data have been analysed with reference to the CRM to ensure that CRM definitions are applicable within the archaeological environment. The concepts inherent in the CRM ontology have been used to model the knowledge gathered from the following work areas within the CfA that have been investigated and analysed. These areas have been considered by reference to the recording forms they employ and the personnel involved

The results of the investigations have been mapped to the CRM in an iterative fashion, i.e. in small portions, each relating to one operational area, e.g. environmental sampling. The models have been developed using the information gathered from the work areas, in conjunction with the concepts declared in the CRM, allowing data to be captured in a consistent way for all the archaeological processes under consideration. Using the CRM ontology to draw on the existing knowledge concepts within the cultural information domain has resulted in a better structure for, and organization of, the domain information. This has given a more comprehensive representation of the data required to answer the complex queries that the ontology has been designed to support. Using the ontology has also revealed repeatable patterns that can be applied in the development of new systems, allowing better representation of knowledge. The models reveal patterns of defined behaviour, which can occur in several places in the domain. Once the behaviour of these patterns is understood, the pattern can be substituted appropriately elsewhere without having to remodel the requirements. This is a toolkit approach to modelling, where pre-defined patterns of behaviour become re-usable components within the Domain. For example, a generic activity component was created using the properties deemed to be relevant to CfA activities. As such, all CfA activities have associated Time-spans and can be seen to involve an Actor or Actors in specified Roles. This activity component is used wherever an activity occurs in the model. Similarly, activities occurring within different process areas can be seen to be closely related (See 3.2.1 Process Areas) such as the way in which geoarchaeologists, geophyicists and environmental scientists conduct survey activities which result in new Information Objects representing spatial datasets. The improvement in data depiction which results from achieving a common understanding of knowledge will have the added benefit of improving communication between people and information systems.

The CRM is an event-driven model, whereas many of the existing paper-based systems in operation at the CfA do not document the events that lead to data being created, gathered etc. An event-driven approach allows assembly of a sequence of events or activities that can be related to specific data which is in someway connected to that event, e.g. gathering or generating new data. Accordingly, the view taken of the system is that an event is simply another entity to which data can be attached. For example, the finding of a coin is an event, a hook to which information regarding the finding of the coin can be attached, information that may not relate directly to the coin itself but to the act of finding the coin. The detailed coin record subsequently produced can then be seen to be the result of the act of finding the coin and the requirement or desire to record it. This adds traceability to the data, a factor that is lacking in many of the current systems. Using the CRM event-driven approach allows data to be linked to an event. As a consequence, the data can then be linked to a timespan (<<E52 Timespan>> in the CRM) via the event, allowing formation of a temporal picture of events and the data associated with specific processes. Using an event-driven approach has meant enhancing the existing systems with Classes that capture events as described within the CRM, allowing us to pose questions such as:

‘Which conservation treatments have been used on leather artefacts from Roman waterlogged sites within the last 10 years.’

The event-driven approach requires extra thought as not all the forms used at the CfA capture the events associated with the data they record. However, utilising the event-driven approach in modelling the system has been beneficial in that it has allowed the building up of a more complete picture of the archaeological process, and has provided the ability to respond to queries that are currently impossible to answer. Using the ontology has given a consistent way of depicting the classes and allowed patterns of behaviour to be modelled in a way that can be applied to other systems within English Heritage.

The various sections that have been analysed are discussed below. The groups do not reflect the organisation of the various work areas, but rather the organisation as reflected in the findings. There are instances, for example Environmental Studies and Finds Processing, where information crosses departmental boundaries, and the diagrams show this. Therefore, they do not show the demarcation of tasks as implied by the CfA organisational structure.

3.2.1 Process Areas

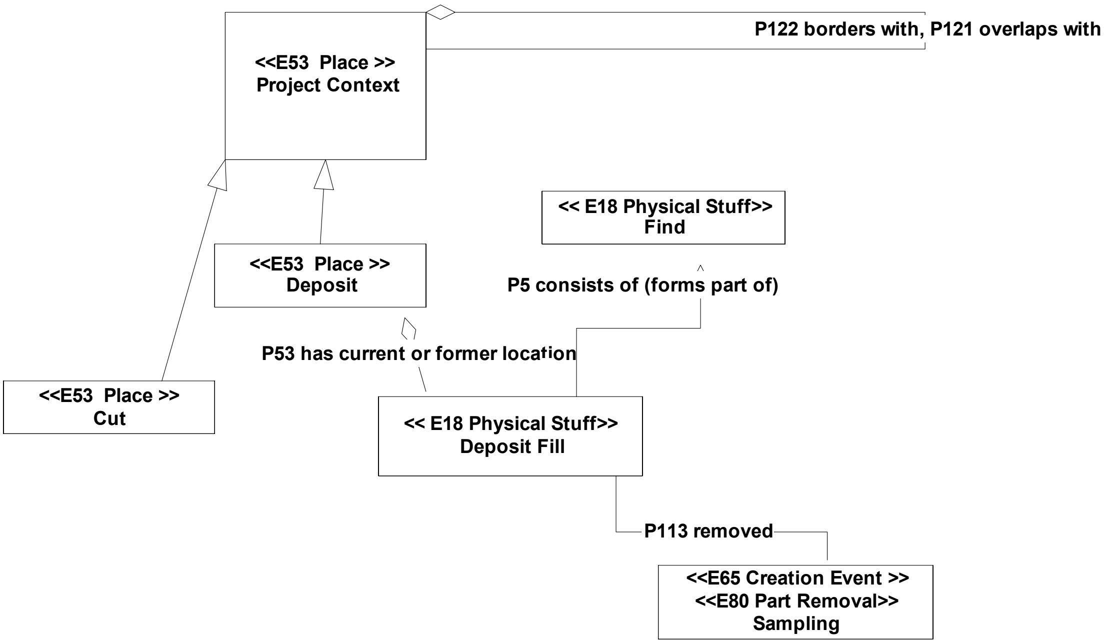
There are a number of existing CfA systems that are used for several different purposes but which do not share information. There is also a problem with the different systems currently in use, in that the output from, for example, Delilah (Context Record) is not easily imported into Labfile (a collections and conservation management system) or the various Access database systems, e.g. those used for archaeobotany and animals bones or dBase systems used for skeleton recording. This limits the sharing of data between groups and is costly in terms of maintenance as resources of various kinds are required to convert data between formats. The modelling exercise has shown the importance of sharing information that is common to a number of processes, in order to avoid the need for re-entry of data, and to prevent errors arising when re-entering.

Recording System

Source: Form “Deposit and Cut”, Archaeology team

This model shows the start of the excavation process, and forms the basis for further analysis of items taken from an excavation, such as finds and samples. It is the basis for capturing all data associated with an archaeological excavation. The Recording model shows the importance of the single-context recording system, as the context recorded in the field is then led into a number of different systems, e.g. Labfile, Bone recording, Delilah. A context is considered as a place that is a container for <<E18 physical stuff>>. Physical Stuff is found in a context, and can be made up of more Physical Stuff within a context, e.g. find, soil, wall, skeleton. The Physical Stuff can be of different types, as shown below, with a deposit fill containing finds, and being subject to sampling. The context is sub-typed into cut and deposit, as on the recording form. The information recorded is different for both. See the detailed model for more information regarding the relationship between cut and deposit.

Recording System



Sampling
Sources: Geoarchaeology, archaeobotany and zooarchaeology specialists

The Sample model represents the information gathered from environmental sources. It is linked to the Recording System by the Class “Context Sample” which is a stereotype stating that all samples behave in the same way as <<E22 Man Made Object>> as declared in the ontology.

The detailed Sample model shows the pattern of assigning an Actor to a particular Event, and has led to the assembly of patterns that can be re-applied as components throughout the system. A sample can have a number of events linked to it that result in it undergoing changes or, for example, lead to the production of analysis reports. Actor assignation is important for tracking who did what to a particular sample, and it allows a sequence of events to be built up telling the life history of a particular sample. This method of representing knowledge is to be preferred to other data-driven approaches, as it allows compatibility with the CRM, and shows where gaps in information are often to be found in the tracking and monitoring process associated with sampling.

Thinking of Physical Stuff as an integral part of the data structure, taking part in events, has aided understanding and capture (i.e. modelling) of the sampling processes. This has helped to highlight gaps where information is not being captured, such as who did what to sample residues, where sample residues are kept etc .The detailed model attempts to resolve this situation, although more work is needed to consolidate and complete it.

Finds
Sources: Forms “Finds”, Archaeology team, Conservation team

The Finds model represents the information gathered from the layout and content of recording forms and interviews with finds specialists and conservators who explained the Labfile system. The Finds model is linked to the Recording model by the Class "Find" which is a stereotype stating that all finds behave in the same way as declared for <<E18 Physical Stuff>> in the ontology.

There are the following similarities between Sampling and Finds processing:

- The condition of Find is assessed multiple times
- A Find changes custody and can be moved, i.e. change location
- A Find can undergo treatment and be altered/ split into its component parts.

Using the CRM has helped to identify the commonality between sampling and finds processing, as the usage of Classes was constrained and, accordingly, channelled thought processes along the same avenue, giving a consistent approach to depicting information within the Domain. This is particularly useful when more than one individual is involved in modelling, as the solution, although produced by several people, will be presented using a common nomenclature.

4. Conceptual Framework

4.1 Archaeological perspective for ontological modelling at CfA,

Within archaeological context recording systems, of the kind used by the CfA, the notion of context is central to recording, with all records being context-based, and all contexts being numbered from a single allocation scheme. This identifier assignation is common to all contexts but beyond this, the nature of what is referred to as context differs greatly. A context can refer to a section of wall, the cut of a ditch, a skeleton, or the secondary fill of a post-hole. A context can even refer to a sample, being used when finds are recovered from a sample or samples.

When faced with this diversity, the approach adopted was to look for commonality between the multitude of context types in order to identify the essence of what is a context. After much discussion and consultation with the CRM consultants, a consensus emerged that the best way of thinking of context is as a place, or in CRM terms <<E53 Place>>: *"This class comprises extents in space, in particular on the surface of the earth, in the pure sense of physics: independent from temporal phenomena and matter."* If we take again the examples of context from the previous paragraph, it is the spatial component which is common to all context types, and the matter component that differs.

Context	Spatial component	Matter component
A section of wall	The context refers to a three-dimensional solid deposit of building and bonding materials,	There is matter in this place, the building material and any

	its extent delimited by the three-dimensional surface of the wall.	bonding material.
The cut of a ditch	The context refers to a three-dimensional plane, defined by the observed cut	There is no matter in this place, the place is a geometric shape (only) representing an interface between two deposits
A skeleton	The context refers to a series of three dimensional solids, the extent delimited by the three dimensional planes that are the surfaces of the bones.	There is matter in this place, the bones themselves, a collection of biological objects.
The secondary fill of a post-hole	The context refers to a three-dimensional solid, its extent defined by the three-dimensional planes delimited by the observed post-hole cut and the upper and lower bounds of the fill.	There is matter in this place, the deposit interpreted as being contiguous.

What we are effectively stating is that the concept of context as used within context recording systems has a duality of meaning. When we discuss context, we are actually talking about a place which is the location of some material (deposits and structures) or defines the extent of other places (cuts and features). These two facets to the concept of context are generally conflated, so we talk about the shape in section of a context, referring to the spatial characteristics of the context, and also the colour of the context, referring to the material nature of the context. Furthermore, for the skeleton context type, the matter in the place we refer to as the context, once excavated, is given a finds identifier and treated as a collection of objects. As such, the distinction between the purely geometric shape and the material bounded by it, if any, serves to make explicit the nature of context and facilitates treating contextual data in an appropriate manner.

One aspect of this solution presented a further problem to the project team, i.e. the nature of context formation. If we are saying that the spatial component to the concept of context is <<E53 Place>> which witnesses the formation of the context, it would appear that we have an object which witnesses its own creation. This is not the case, and relates again to the conflation of the two aspects of the concept. Within the CRM, <<Place>> is ‘independent from temporal material and matter’ with the implication that the places we call contexts have always and will always exist, independent of any context formation processes which may result in material being deposited in the <<Place>>. This may seem unnatural, as it is the process of context formation which appears to define the place that is context, but can be seen as analogous to the use of modern place definitions to describe the past, e.g. the place that is referred to as Oxfordshire can be used when describing Roman Britain, it is used here in its purely geometric sense, as an identifiable spatial unit, despite the fact there was no concept of Oxfordshire in Roman times: Spatial phenomena are independent of time.

The idea of phasing is central to archaeology; it is the process by which the individual components of a site, identified through excavation, are put back together to form a sense of what the archaeological record is telling us. Phasing is achieved by means of a grouping process, whereby the individual components are assembled into meaningful collections which can then be discussed. In terms of the model, groups, on one level, are places in the same way as contexts. We can talk about a building, based on the presence of a number of post-holes interpreted as a building, and this building group can be seen as a place within which the component places are located. In addition, we may wish to talk about the matter in the place, i.e. the structure that is the house. In this case we can say the building itself is the location of the matter which comprises the physical structure, this matter being comprised of the matter at the various places representing the individual contexts. There are, therefore, two ways of grouping material relating again to the spatial vs. matter aspects of context. With respect to the archaeological process, it is likely that the grouping of matter is of less benefit than grouping of places. As all matter has a location associated with it, i.e. the place where the deposit formed and from where the matter was excavated, it is possible to interrogate the properties of the matter by searching through the recursive place groupings. For example, by stating that contexts 001-034 fall within a place we refer to as a *foundation wall* and group 203, and that the foundation wall falls within a place we refer to as a *farm-workers cottage* and group 219, we know that the material found in contexts 001-034 can be found at the place that is farm-workers cottage 219. There may, however, be situations where such grouping is inadequate. For example, a condition assessment of a listed building will result in an overall statement of condition for the building. Places cannot have a condition state, being a purely spatial phenomena; it is the total collection of material in the various places that forms the building which is being assessed. In this case, it is the physical matter which is grouped by stating the building is <<P46 composed of>> the matter which can be found in individual contexts.

The matrix is a particularly interesting component of the archaeological process. Currently at the CfA, matrices are compiled to aid stratigraphic analysis, but they are not integrated into the data management aspect of the recording system. They are seen more as a type of drawing, a product of the observed physical and interpreted stratigraphic relationships identified on site. Having said this, matrices are not a dynamic output of the recorded physical and/or stratigraphic relationships; rather they represent a broader interpretive view of the excavation, incorporating finds and environmental evidence in the iterative process of matrix compilation. In terms of the model, this can be represented in terms of a creation event which an archaeologist uses to create a matrix as an information object based on available information. The real power of the CRM-based model relates to implementation and how we can use this information object; the compiled matrix can be used as a view into related datasets. This makes it possible to examine finds information through a stratigraphic means of presentation, or examine environmental evidence relative to stratigraphic relationships. As such, the current functionality of the matrix as a graphical representation of the observed stratigraphy is retained and new functionality added as the matrix becomes a tool with which to examine other aspects of the total site dataset.

4.2 Relating the model(s) to the sector

The primary aim of the ontological modelling project was to produce a model of the CfA's archaeological processes in order to inform future systems design. Nevertheless it was always a consideration that the main conceptual archaeological processes of the CfA are likely to be closely aligned to other archaeological organisations that carry out archaeological investigations comprising fieldwork recording, analysis and reporting.

Most archaeological organisations which carry out investigations in England use a form of context recording system for recording the individual stratigraphic units of excavation. Most systems relate the recording of objects found within the deposits to the contextual record and need to record information about the condition of the finds and any interpretive assessment of their dating. The process of grouping contexts together to make interpretive analysis of events such as construction, use and disuse and the combination of different elements of the site into distinct phases is also a common conceptual process. Many different activities during excavation and analysis require the taking of samples or sub-samples of different materials and subjecting them to a series of observational and measurement events carried out by specific individuals in various expert roles.

These basic elements form the 'core' of the conceptual model of the CfA's archaeological work as mapped to the CRM. At this conceptual level, it should be possible for other archaeologists to map similar entities within their information systems to this conceptual framework - be they paper-based or digital.

In this way a common conceptual framework for archaeological processes can be developed that will enable cross-searching between data from different archaeological recording systems. This should make it possible to conduct meaningful searches for common conceptual entities across data sets held by different organisations.

4.3 The general use of ontologies

As already outline earlier, an ontology provides a shared and common understanding of data and, in some cases, of services and processes that exist within a domain. This facilitates communication between people and information systems and an enhanced ability to search for information across different knowledge repositories.

As already touched upon in the introduction, an ontology is an explicit formal declaration of how to represent object concepts and other classes assumed to exist in some area of interest and the relationships between them. It is created for the purpose of enabling knowledge to be shared and re-used by and among agents identified within the ontology. This is achieved by making a set of definitions of a formal vocabulary that the ontology commits to use and which allows system builders to specify systems using the specific vocabulary (i.e. ask queries and make assertions) in a way that is consistent (but not necessarily complete) with respect to the theory and concepts specified by the ontology. Practically, this usage is termed making ontological commitments.

Another important notion is that of entity correspondence. Ontologies allow consistent representation of data located across very different information systems and information that resides in many separate domains. To determine correspondence between entities in different systems we need to identify the entities, and then using the ontology we can identify where entities utilise the same concepts (in UML using the stereotype). We can then decide where these stereotypes are essentially the same item. This helps us to utilise only the relevant information in a common data store and enables a much greater degree of searching across different domains that use the same ontology.

4.4 Semantic Web developments, RDF and ontologies

We are presently seeing the advent of the use of languages for ontology, built on reasoning techniques that provide for the development of special purpose reasoning services. In fact, the W3C has created a Web standard for ontology language as part of its effort to define semantic standards for the Web. The Semantic Web is the abstract representation of data on the World Wide Web and is based on the Resource Description Framework standards which provide interoperability between applications that exchange information. RDF uses XML to define a foundation for processing metadata and to provide a standard metadata infrastructure for the Web and organisations. It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners led by Tim Berners-Lee.

In order for the Semantic Web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. This notion is known as knowledge representation. To this end, and in the domain of the World Wide Web, computers will find the meaning of semantic data by following hyperlinks to definitions of key terms and rules for logical reasoning about data. The resulting infrastructure will spur the development of automated Web services such as highly functional agents that can automatically search for data. What is important here is that the work now being driven by the W3C as a way to manage semantics on the Web is applicable, at least at the component level, to the world of application integration, much like XML and Web services.

An example of the W3C contribution to the use of ontologies is the Web Ontology Language (OWL). OWL is a semantic markup language (a bit like HTML) for publishing and sharing ontologies on the World Wide Web. OWL is derived from the DAML+OIL Web Ontology Language and builds upon the RDF. OWL assigns a more specific meaning to certain RDF triples. The future Formal Specification at the W3C specifies exactly which triples are assigned a specific meaning, and offers a definition of the meaning.

Using these Web-based standards as the jumping-off point for ontology and application integration, it is possible to define and automate the use of ontologies in both intra- and inter-organisation application integration domains. Domains made up of thousands of systems, all with their own semantic meanings, are bound together in a common ontology that makes short work of application integration and defines a common semantic meaning of data.

This, indeed, is the goal. Extending from the languages, we have several libraries available for a variety of different domains including financial services and e-business. There are also many knowledge editor packages that now exist to support the creation of ontologies, as well as the use of natural-language processing methodologies. These are available in commercial knowledge mapping and visualization tools using standard notations such as UML.

5. Conclusions and lessons learned

The main conclusion to be drawn from this project is that, despite their assumed/perceived complexity, it is possible to analyse archaeological data and processes and produce coherent models of a range of situations. While more traditional data-modelling techniques, with their rigid sets of rules and procedures, have proved to be too inflexible when approaching archaeological systems, a more descriptive approach using object-oriented techniques facilitates the creation of a model which represents a much closer abstraction from the real world. Rather than treating every situation as unique, an object-oriented approach is focussed on pattern identification and the idea of object Classes and inheritance.

The process of producing a high-level conceptual model of the information domain at the same time as producing detailed class diagrams for the identified Domain Classes was both helpful and at times a hindrance. Undoubtedly, the process of mapping identified Domain Classes to CRM Classes would best be done with a complete and fully developed set of Class Diagrams. But equally, it was the conceptual framework provided by the CRM Class scope notes that provided the basis for the class definitions. In this way, the CRM provides a guide to good practice for our conceptual modelling and the detailed modelling work provides a check on the high-level conceptual modelling. Having said this, greater familiarity with the CRM at the outset would have reduced the need to front-load the CRM modelling activities and would have facilitated a more traditional two-stage modelling then mapping approach.

Indeed, familiarity with the CRM proved to be crucial. It is only through a detailed understanding of the ontology that it can be used effectively, and even then there is still plenty of scope for discussion regarding the way things are to be mapped. Indeed, the initial idea that the Domain Experts would be able to discuss their Domain in CRM terms proved difficult and the approach adopted involved exchanging information with Domain Experts without using any CRM terms, just natural everyday language which everyone could understand, thus hiding the complexity of the ontology.

Another issue here was the use of terms of which people already have an understanding, although not necessarily one which is compatible with that defined by the ontology. An example of this would be <<E27 Site>>, which in CRM terms is *not* a Place and cannot be identified using spatial co-ordinates. The project team in effect acted as translators between the clearly defined concepts expressed in the CRM and the less well defined concepts which archaeologists use every day. Indeed a team comprising archaeologists, archaeological specialists, and IS specialists, all with some training on the CRM proved a good mix of skills for the work in hand. Having team members already familiar with OO techniques eased the process of familiarising the team with the CRM and how it works.

5.1 Gap analysis, enterprise modelling

Modelling archaeological processes is aided by using an ontology as it enables us to identify gaps in the data collected *via* recording forms, especially in terms of discovering why a particular conclusion was drawn or decision made. Commonly, not all the events leading up to a conclusion are recorded and it is difficult to work new evidence into an interpretation if information on previous conclusions is lacking or inadequate.
Enterprise modelling is being used increasingly in commercial environments to understand better the concepts within an organisation, and to look for improvements in efficiency and maintenance. Understanding commonality leads to a tighter organisation in terms of controlling data and access to it. It also leads to more efficient processes as the data is understood and integrated effectively into the processes.

5.2 Ontological software

Associated with the rapid growth in the use of ontologies, a number of applications have emerged designed to facilitate working with ontologies. As part of the initial stages of this project, a short assessment of literature relating to available applications and their functionality was conducted in order to assess the usefulness of such applications for the task in hand.

The notion of using software applications to aid in data modelling and systems design is not new. Such applications have been developed through time, generally associated with specific methodologies, and aid the process by automating data handling and performing various validation routines.

Initial thoughts regarding the use of an ontology development application centred on the prospect of automated indexing and relationship checking, as one would expect with a structured design application, the ontological model on one level simply being a structured collection of entities and relationships. By using an application which treated the ontology as a set of data as opposed to a purely visual representation, the idea was that editing and version control would be facilitated and the dynamic dataset would be capable of being viewed and worked with in different ways. It was also hoped that some level of integrity checking could be undertaken to ensure consistency throughout the model i.e. where concepts re-occur, one description should not negate or contradict another.

The assessment concluded that, for the task in hand, the available applications did not offer significant advantages over a less automated approach; the time it would have taken to identify a suitable application, learn its particular implementation of ontological theory (there are many variations in supported features and terminology), and use it effectively to capture domain knowledge was better invested thinking about the archaeological situation in hand. Furthermore, the available applications primarily support the creations of ontologies from scratch and subsequent population of the knowledge base. This was not the aim of this project: The aim was to produce a conceptual model of the domain, building on an existing ontology. In order to use one of the applications in this way, it would have been necessary to represent the CRM using a common format (e.g. DAML+OIL, RDFS, XML Schema, etc) capable of being loaded into the application, most likely, this would have been the RDFS version available from the CRM website. The application would also have to support all features of the CRM; some features such as polymorphism, are not always supported. Due to these technological issues, the use of a dedicated ontological software application was avoided.

As a result of this, the mapping/modelling exercises made use of more traditional tools where, despite the need for increased manual control, it is possible to produce graphics not restricted by the output from a particular ontology editor. Graphical tools were used to produce diagrammatical representations and word-processors and spreadsheets used to hold tabular representations. Additional complexity was introduced into the process by the conflation of mapping and modelling so the avoidance of cutting-edge software was beneficial to the overall project programme: It may be that the output of this project is used to create a knowledge base using one of the available ontology editors, but such a decision relates to subsequent implementation stages not this conceptual modelling stage.

5.3 Sound data models aid direct mapping to the CRM

A significant proportion of the time invested in the project was used to produce descriptive models of the current situation: Having a clear picture of the current situation is essential. Where good systems documentation exists in the form of structured design diagrams (Entity-Relationship Diagrams, Data Flow Diagrams, UML diagrams) or database descriptions (table definitions, field descriptions, relationship parameters), these are of enormous help. The lack of such documentation forced the project team to effectively reverse engineer systems based the results from the first stage of Revelation and verbal communication with system users. While the interview approach was seen as an important mechanism for extracting domain knowledge from domain experts, the availability of good quality design documentation, is advantageous; the interview process is best used to clarify the meaning of and uses for data items rather than to identify the presence of data items.

Having said this, the process of producing detailed models of the current situation should not be undertaken in isolation, in advance of any application of the CRM. Thinking in terms of CRM concepts, objects and events, is advantageous when describing the domain. As such, while the apparent conflation of the modelling and mapping exercises made the process more complicated, it ensured the detailed models of the domain did not conflict with the conceptual overview being developed in parallel, a danger had a more traditional modelling followed by mapping approach been adopted.

A recommendation from this project would therefore be that any detailed models describing aspects of the domain not currently adequately described, should be undertaken by someone familiar with generic OO principals, as well as the top-level CRM concepts and principals, using a form of description compatible with CRM constructs, such as UML.

5.4 Project management lessons & further business needs

Defining a method for ontological modelling

Because the CRM does not currently come with a simple ‘User Guide’, its application has required some methodological development work by the project team. One practical issue that arose early in the project was how to find a way of producing verifiable models for the domain-experts who may be totally unfamiliar with ontologies or the CRM. In the end a number of complementary diagrammatic and text-based versions of the model were produced.

Difficulties bridging the ‘O’ word gap

There was some initial resistance amongst CfA staff to the use of the term ‘Ontology’, which was not familiar to most archaeologists. This may be partly because the term itself has dual meanings in the different domains of Philosophy and Information Management. Nevertheless, given the requirement to work with the CRM ontology and, after some initial briefings and workshops, it became clear that attempting to use alternative terms would prove equally unsatisfactory when discussing the work with ontology experts. The CfA project also employed a consultant with archaeology and systems design background to help in overcoming some of the communication issues.

Limits of the CRM for project management and admin functions

During early consultations on the development of the model it became clear that certain aspects of the work of the CfA were less likely to be appropriately modelled using the CRM. As the introduction to the CRM states, “Information required solely for the administration and management of cultural institutions, such as information relating to personnel, accounting, and visitor statistics, falls outside the Intended Scope of the CRM” (Crofts 2003). Thus areas of administration and logistics such as systems for recording personnel employed on archaeological projects and some of the more detailed aspects of project management of CfA projects were not covered. It was decided to show on the model where project management systems would fit into the conceptual model, but following discussion with DSU it was felt that further details of project management (e.g. Prince2) might be modelled as part of other CRM projects within EH.

Models for Implementation

CRM modelling is at a conceptual level and as such it does not produce a data model which is directly implementable, but rather defines the conceptual framework for the data which must be incorporated into a successful systems design. The production of a UML model was considered to be the best way of documenting the existing data entities used by the CfA in a way that would enable systems design and implementation to incorporate the conceptual model produced by the CRM modelling. The UML model should also provide a means for assessing the business benefits that would accrue from moving from the existing systems to a new systems design based upon the UML and CRM models.

6. Glossary of all technical terms and acronyms

Term or Acronym	Definition or Description
AOI	Area of Investigation
Allen's Temporal Operators	James F. Allen was among the first people investigating temporal reasoning based on intervals. The thirteen possible relations among two intervals, such as equality, overlapping, inclusion, are commonly known as Allen's operators. For more information see - http://www.cs.rochester.edu/users/faculty/james/
CfA	Centre for Archaeology
CIDOC	International Committee for Documentation
Class	A class is a category of items that share one or more common traits
Commonality	The commonality between classes show their shared attributes
Component	In UML models a component is a short-hand representation of the repeating patterns identified within the model.
Conceptual Objects	Non-material products of peoples' minds, characteristically created, invented or thought, and documented or communicated between people
CRM	Conceptual Reference Model - the CRM is a semantic ontology - a set of rules for describing the possible 'state of affairs' in museums, archives, libraries etc.

DAML + OIL	DARPA Agent Mark-up Language - a web ontology language
Domain	A domain is the class for which a property is formally defined
Event-Driven Model \ Event-Based Model	An event-driven (or event-based) model is a model that is derived from the fact that all things can be described by way of events eg a condition assessment of an object would only be possible as a result of an 'observation event'
FISH	Forum of Information Standards in Heritage
Foreign Key	A foreign key in a database table is the field from another table that is linked to the primary key in the table being used.
GIS	Geographical Information System
Granularity	Resolution or level of detail of the model
ICOM	International Council on Museums
Inheritance	Properties are inherited from superclasses to subclasses - all the properties of a superclass also apply to its subclass.
Instantiation	An instantiaition of a class is an item that matches the criteria of that class
ISO	International Organisation for Standards
MIDAS	Monument Inventory Data Standard
NMR	National Monument Record

Object-Oriented Model (OO model)	Object-oriented models model real-world objects (entities), the constraints on them, and the relationships between them (properties). It is capable of extension without alteration.
Ontology	Ontologies are formalised knowledge, consisting of clearly defined concepts with linking relationships
OWL	Web Ontology Language
Platform-Specific Model	The development of the model relative to the detail of the technology available
Properties	A property defines a relationship between two classes
Range	A range is the class that comprises all potential values of a property
RDF	Resource Description Framework is an evolving metadata framework that provides a degree of semantic interoperability among applications that exchange machine-understandable metadata on the Web
Schemas	Schemas are machine-processable specifications which define the structure and syntax of metadata specifications in a formal schema language. A schema is a diagrammatric representation of a model.
Semantic Web	The Semantic Web is a mesh of information linked up in such a way as to be easily processable by machines, on a global scale. It can be thought of as being a globally linked database or an efficient way of representing data on the World Wide Web.
SMR	Sites and Monuments Record
SSD	Site Sub Division
Stereotypes	A stereotype used in UML means that all the properties of the class in the CRM are inherited by the class when it is used in the UML
Subclass	A subclass is a class that is a specialisation of another class (its superclass)

Superclass	A superclass is a class that is a generalisation of one or more other classes (its subclasses)
System	A system is a process that assembles, stores, manipulates and delivers information
TELOS	Technology for Electronic Library Organisation and the Semantic web
Triples	Triples consist of 2 classes linked by a property, and these form the building blocks of the CRM
UML	Unified Modelling Language
W3C	World Wide Web Consortium. The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential
XML	Extensible mark-up language XML Schemas provide a means for defining the structure, content and semantics of XML documents, including metadata.

7. Acknowledgements

Thanks go to Martin Doerr and Steve Stead for expert advice and guidance on the CRM. Further help, advice and discussion was provided internally by Matthew Stiff, Ed Lee and Emma Turner of the EH Data Services Unit. Thanks also to all CfA staff for their patience, understanding and willingness to engage with difficult and occasionally off-the-wall concepts. Nevertheless the views expressed here are solely the responsibility of the project team; Paul Cripps, Dave Fellows, Anne Greenhalgh, Keith May and David Robinson.

8. Bibliography

Crofts, N, Doerr, M, Gill, T, Stiff, M and Stead, S (eds.) Nov 2003 *Definition of the CIDOC Conceptual Reference Model and Cross Reference Manual. Version 3.4.9. Official release of the CIDOC CRM* http://cidoc.ics.forth.gr/official_release_cidoc.html

Cross, S, et al *Revelation Assessment Report* (CfA 2004 forthcoming)

Cripps, P and May, K *To OO or not to OO? – Revelations from Defining an Ontological Model for an Archaeological Information System*. (CAA 2004 forthcoming)

Denny, M 2002 *Ontology Building: A Survey of Editing Tools* http://www.xml.com/pub/a/2002/11/06/ontologies.html November 06, 2002

Doerr, M 2003 *The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata*, 'AI Magazine',.**24**, 3

Guizzardi, G, Herre, H, Wagner, G 2002 'Towards Ontological Foundations for UML Conceptual Models' in *Proceedings of 1st International Conference on Ontologies, Databases and Applications of Semantics*, Springer LNCS 2519. page no.s????

9. Appendices

A. CRM Model Diagram
B. CRM Class and Properties text description
C. CRM Statements – Excel version
D. UML Modelling
E. Terminologies Listing
F. Training Introduction to the CRM for the CfA

Figure List