Ontology or meta-model for retrieving scientific reasoning in documents: The Arkeotek project

N. Aussenac-Gilles

IRIT, UMR 5505, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex9, FRANCE
http://www.irit.fr/~Nathalie.Aussenac, aussenac@irit.fr, +33 5 61 55 82 93

Abstract :

Electronic publishing and large databases make it possible to store scientific data together with the texts that report scientific studies about these data. The SCD model, inspired from the logicist program, suggests to structure documents according to the role of each paragraph in the overall argumentation. The Arkeotek project promotes the use of the SCD format for scientific writing in archaeology.

To improve information retrieval in the collection of all SCD documents, the Arkeotek project experiments semantic annotation with a domain ontology. This ontology is rather a domain meta-model used to enrich SCD-structured documents. This paper presents the SCD format and the domain model. It illustrates how the target of document annotation influences the selection and definition of concepts in the domain model. It discusses the status of the model with respect to a core ontology like the CIDOC CRM.

Key words: Documents structuring, Logicism, information access to document content, ontologies, semantic annotation.

1 Introduction

Electronic publishing and large databases make it possible to store scientific data together with texts that report scientific studies about these data. The SCD model, inspired from the logicist program, suggests to structure documents according to the role of each paragraph in the overall argumentation. Writing, storing, reading and indexing such documents turns out to refer to new paradigms. Whereas most of current works about digitalized documents propose to develop afterwards some formal models of how documents organize and structure the information they contain, the Arkeotek project promotes the use of the SCD format as a structuring framework to organize their content as writing proceeds. The SCD format aims at making explicit logic inferences and reasoning in scientific productions (papers and monographs) in archaeology. Documents turn out to be models by themselves. Several texts are available now on the Arkeotek web site[1].

To improve information retrieval in such databases, the Arkeotek project experiments semantic annotation with a domain ontology. The ontology is rather a domain meta-model used to enrich structured documents. Thanks to this annotation, domain researchers may express queries to look for some scientific hypotheses or methodological choices in documents.

After an overview of the Arkeotek features, goals and tools (§2), we present the SCD format for structuring scientific documents (§3). Then, we sketch the structure of the domain model and the selected options to define its concepts (§4). We show how they are used to annotate scientific documents, and how this use influences the model. We finally discuss its status with respect to a core ontology like the CIDOC Conceptual Reference Model (CRM) (§5). To illustrate this debate, we comment how the CIDOC-CRM guides a more precise definition of chrono-cultural periods. We conclude with some methodological principles and we particularly underline the necessity of a cross-disciplinary approach to deal with such issues.

2 Overview

2.1 Goals

The Arkeotek project promotes the logicist re-writing of scientific texts in human sciences and their edition under the SCD *(Scientific Construct & Data)* format [13]. This priority originates from the following acknowledgment: it is no longer possible to get through all the publications related to a given field of research. It follows that we do not read anymore: we just browse. However, scientific texts are written not to be browsed, but to be read in their linear form. One of the trends of the logicist program aims at producing texts whose structuring helps to browse scientific constructs as well as to assess their valid foundation [11]. In other words, editing scientific texts under the SCD format aims at increasing (a) the number of publications that a reader can assimilate, (b) the legibility of their foun-

[1] http://www.arkeotek.com/

dation, (c) at last, the possibility of building bases of structured corpus. In the future, it should prompt a better dynamic for research. So the Arkeotek project follows a three-fold goal:

1. To better retrieve in documents the knowledge and reasoning that scientists mobilize to establish their results.
2. To study or edit the way a scientist makes use of sources to produce a result (each research step is considered here as an epistemological data that contributes to measure the result validity).
3. To share scientific sources and documents so that it stimulates the dynamics of research in the field.

These goals are partly achieved after publishing several PhD Theses in the Référentiel collection by MSH-Editions Epistèmes (like [5] and [16]) and the on-line Arkeotek Journal[2].

2.2 Competency questions

The project anticipates two use cases, the intended users including domain experts and scientists:

A) Researchers want to read selected independent documents (available in CDs like [14]): browsing documents is guided by their strong logical structure (the SCD format);

B) Researchers want to retrieve some precise information within the large Arkeotek collection (papers and monographs will all be available on the Arkeotek web site). For instance, they may need to cover the state of the art about a particular topic, for instance about some particular manufacturing technique or production organisation. They may also look for methodological issues (how did an archaeologist produce a scientific result?) They will express a query or browse a domain model (ontology) to get to the relevant parts of structured documents.

A multimedia query and browsing interface will enable the two use modes.

The main purpose for building the ontology is the second use case [2]. The model is built to provide some meta-data to annotate theses and scientific papers related to archaeology. The concepts used as meta-data should capture the semantics of documents, especially scientific reasoning and inferences. Examples of requests could be:

[questions about scientific results in the field]

- *What are the steps of the processing chain for beads?*
- *Which techniques have been used in India for manufacturing beads?*
- *How have pottery production techniques been transmitted over the Senegal Valley?*

[questions about scientific methods or techniques]]

- *Which technique can be used to identify the manufacturing period of potteries?*
- *Which are the possible methods to study human skills?*

2.3 Available Tools and Principles to feed the Models

The SCD documents and the domain ontology form two complementary models that are built up and maintained by domain authors and experts with the support of three tools: an authoring toolbox, a semantic annotation environment and the user interface [3]. The life-cycles of the two models are intertwined. They include the use of the first two tools:

(i) Authors re-write their texts in the SCD format; after a manual decomposition of their scientific production into propositions (in the logic meaning), the Epistemes authoring toolbox[3] guides them to edit a hypertext representation according to the SCD format.

(ii) The ontology is engineered by a domain expert and a knowledge engineer after the analysis of these texts, according to the TERMINAE method [1].

(iii) The ontology is used for document annotation by the authors and the database manager.

(iv) The document collection is maintained together with the ontology maintenance to keep these two models consistent. When annotating a new document, some concepts, relations or terms required for annotation may be missing in the ontology. In such a case, the ontology is up-dated in keeping with the content of the document collection.

The last three stages take place in the same software environment. This makes it possible to import results from a term extraction tool and to use them simultaneously for two purposes: both to define concepts and terms in the ontology and to annotate documents (paragraph by paragraph) with these concepts.

3 SCD Documents as Models

3.1 The logicist program

The "logicist" program is the name given more than 20 years ago to researches aiming at clarifying the mechanisms and foundations of the reasoning underlying scientific constructs [11] [10]. It gave rise to the “schematization” of these reasonings in the sense given by the J.-B. Grize, a logician who defined this term as “models generated by a discourse in natural language” (1974).

The SCD *(Scientific Construct & Data)* format enables authors to capture texts that have been re-written according to logicist principles [15]. Each text is fragmented into several *propositions*, either *initial* or *interpretative*. The resulting structured documents form a model of how facts,

[2] http://www.thearkeotekjournal.com

[3] Epistemes toolbox is a product of Editions Epistemes publishing company.

hypotheses, objects or data are interpreted to produce new scientific facts and hypotheses.

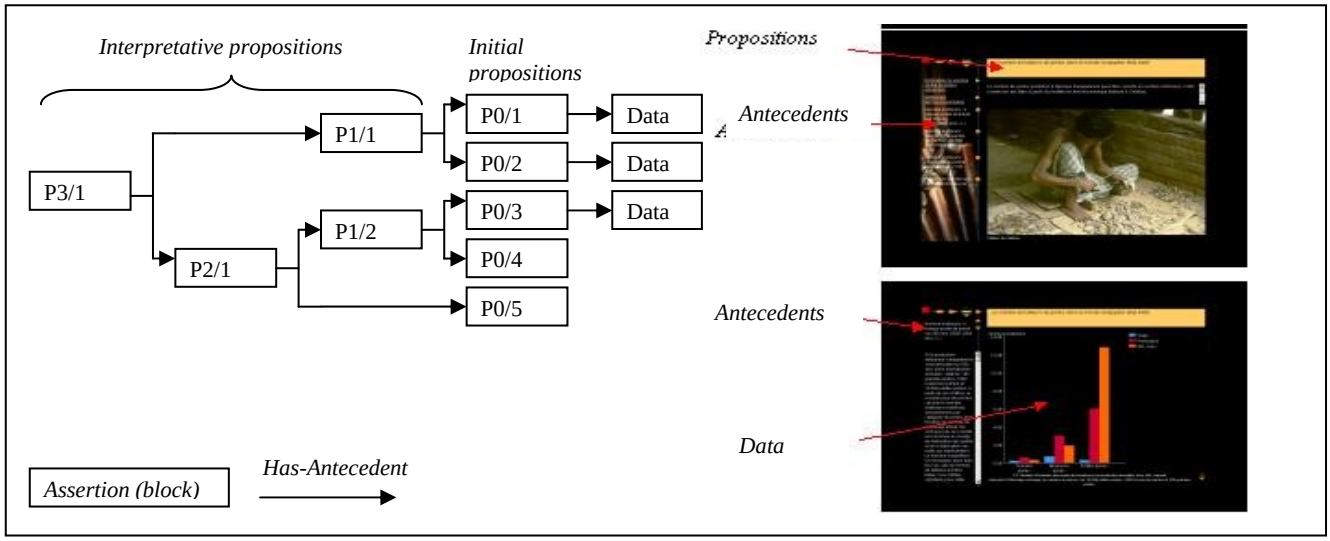
3.2 The SCD format

For each document, a diagram is drawn that forms a synoptic presentation of the chaining of reasoning. This diagram usually is a tree (fig.1). Each node represents a *proposition* that contains a title together with a short text (comments).

In *initial propositions*, comments describe some works, results or measures considered as reference data. They are enriched with multimedia illustrations (pictures on fig.1) that act as a factual database and as cognitive tools for grabbing more efficiently the semantics of the propositions. Initial propositions are the leaves of the tree.

In *interpretative propositions*, comments put to light the logical operations carried out at this step. Interpretative propositions have *antecedent* propositions (their son nodes in the tree) that must be listed and made explicit. All the *antecedents* of a proposition are the statements required to demonstrate its validity (i.e. in fig.1, if P0/1 and P0/2 are true, P1/1 can be inferred). Interpretative propositions are numbered according to their level in the tree diagram (examples of documents structured according to the SCD format are available on the Arkeotek journal web-site http://www.thearkeotekjournal.org).

FIG. 1 – *Document structure according to the SCD format and corresponding displays (from [14]).*



3.3 Document semantic annotation

We decided to use an ontology to characterise the documents content and to support information search in these documents. Document annotation means to associate representative domain concepts to each paragraph (each "proposition" in the SCD terminology). This characterisation should facilitate information retrieval by users, as it has been experimented by numerous projects in the scope of the semantic web [4], [8], [14]. The main hypothesis is that concepts will be more powerful than terms. First, concepts are associated to several terms that allow various terminological formulations of the same idea. Then, relations between concepts make it possible to restrict or to enlarge the focus of a query.

Queries within the annotated document base should lead to some scientific results in propositions. Their grounding will be reachable via the relations between propositions in the SCD format. The reader will be able to track a result from its statement back to the intermediate statements and to the data and facts that justify it.

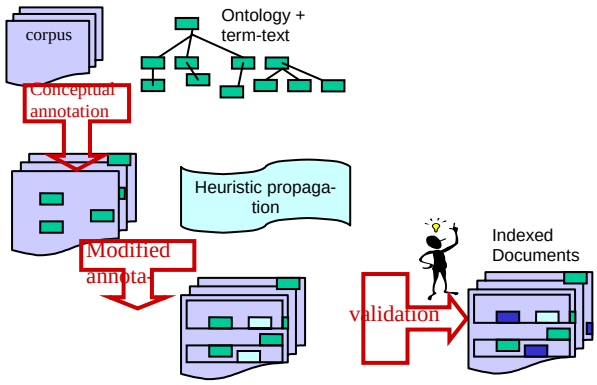


FIG. 2 – *The annotation process*

SCD document annotation is under progress. It is currently based on a naïve process and relies on human validation. Given the domain model and the terms identified in a paragraph, the system suggests that all the concepts associ-

ated to these terms could annotate the paragraph. Moreover, some propagation rules suggest additional concepts, selected among those indexing paragraphs related to the current one according to the SCD format. Then domain experts validate the suggested concepts if they consider them as good descriptions of the knowledge in this paragraph. In the future, this annotation will be combined with classical indexing features, relying on terms and numeral weights that reflect the importance of the concept in the document and in the document collection.

4 Ontology or domain model?

The general question raised by semantic annotation or indexing is "which kind of conceptual resource better meets the requirements of semantic annotation?" Ontologies have been presented as THE solution in most of the studies about the semantic web, but each project uses a particular kind of model, ranging from thesauri to formal ontologies [4], [15].

A first originality of the Arkeotek context is that the documents strong structure bears its own semantics, which is related to reasoning. So semantic annotations should bring an additional value and be consistent with the semantics of SCD format. A second specificity is that we assume that documents and concepts will form together the knowledge base. For these reasons, the model needs neither to fully describe all the information available in the documents, nor to bear all the domain knowledge as if it represented the domain theory. This is why we decided to start with a light formalization and a simple representation (OWL light type). Concepts are defined according to the use of terms, their linguistic validation being considered rather than their formal definition.

4.1 Methodology

We followed the TERMINAE ontology engineering method to rapidly identify most of the terms and associated domain concepts [1]. This method promotes to use natural language processing analysis of domain texts. We selected the whole collection already available in SCD format as a corpus. The SYNTEX syntactic analyser [6] and its ontology editor TermOnto guided the identification of the main domain concepts. Domain experts contributed with their own knowledge to add concepts and structuring relations that were not available in texts. A third knowledge source is the CIDOC-CRM ontology [7]. It provided high level concepts to describe historical periods and their connection with geographical areas.

This model will be checked and updated - if needed - every time new SCD documents will be added to the Arkeotek collection.

4.2 Options concerning the ontology

Our domain model is dedicated to document annotation: concepts and relations are those required to both define the main domain concept and to query and annotate available documents [2]. Therefore the **model needs a rich terminological (or lexical) component** in order to easily identify concepts from term occurrences. Each concept is connected with all the terms that may be used to refer to it either in texts or by domain specialists. In the debate reminded in [15], we decided in favour of a linguistically grounded model, where concepts reflect the practical use of language rather than a formal semantics. As we think that we will not need them, the ontology contains neither axioms nor assertions.

Although texts are scientific productions, **the domain model does not pretend to be the ontology, the set of conceptual primitives of this science**. This model is not even a theory of the archaeology of techniques. Instead, the resource reflects a model made accessible through the way terms and language are used in the texts to be annotated and by domain experts. Texts reflect the point of view of their authors, so concept definitions will reflect these points of view.

This model intends to reflect the conceptual categories that can be differentiated through the use of language. The concepts are those required to describe domain data (archaeological items) and scientific hypotheses in this domain (concepts about the experiments carried out, the techniques under study and their geographical and temporal context) as well as the way results are obtained.

Concepts in the model are those supposed to be necessary and useful to characterize text content and to retrieve information from these texts. The Arkeotek team assumes that users will browse these documents in order to look either for results (interpretations) or scientific methodology. So, for instance, domain data are not annotated in detail because users are not supposed to look for particular archaeological pieces, their picture or their description. Instead, the way a study is carried out is considered as important information to be precisely annotated with various concepts. This choice has a first impact on the **degree of detail of the ontology**. The lower level concepts in the ontology are very domain specific, but they are not always as precise as possible. As a second impact, the ontology contains some **domain specific concepts as well as methodological concepts** that refer to the scientific approaches carried out in this domain (fig.2).

We will illustrate each of these points with examples in the next section. We will report the questions raised and the alternatives that have been investigated. We will also comments why and how the CRM from CIDOC can provide some help and its limitations for our scope.

4.3 Questions raised by concept definitions when annotating initial propositions

The following examples are extracted from a corpus about bead manufacturing in India [16]. Some of the initial propositions describe the corpus of archaeological objects, and others describe experimental and methodological reference techniques or their use in the project.

4.3.1 Initial propositions describing archaeological objects

How precisely should the ontology classify archaeological items?

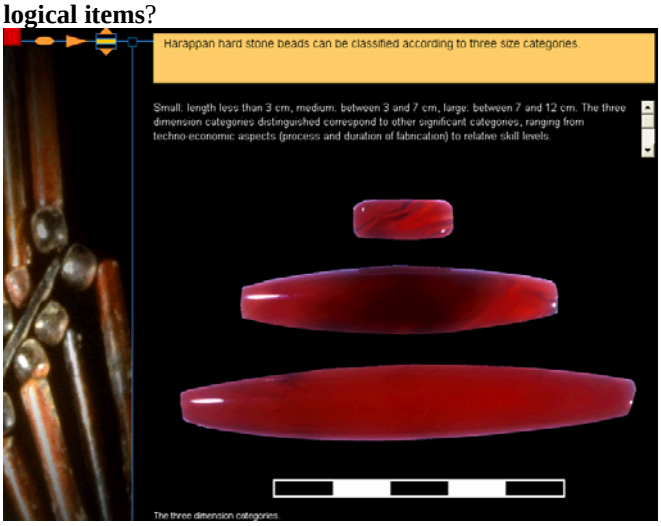


FIG. 3 - *Display of a primitive proposition*

The screen-shot on fig.3 illustrates a primitive proposition: the title (yellow box) summarizes the information detailed below (white letters above black background) and illustrated by a picture. For the sake of readability, we report here the title and comments:

Harappan hard stone beads can be classified according to three size categories:

Small length less than 3 cm, medium between 3 and 2,76 in, large between 7 and 12 cm. The three dimension categories distinguished correspond to other significant categories, ranging from techno-economic aspects (process and duration of fabrication) to relative skill levels).

We have highlighted the terms that domain experts considered as useful annotations for this proposition. Light grey coloured terms refer to concepts describing the data studied and described here (beads) whereas dark grey refers to how these beads are manufactured: the techniques, process used, the time and the skill level required for their manufacturing, and so on.

What does the selected meta-data reveal?

The selection of meta-data tells that what is important here is not to describe a collection (which would be at the instance layer in the ontology). The focus bears on the concepts to be considered by the scientific study: the *objects under study* are *beads*, they are *classified* into *categories* (dimension categories), that will later contribute to understand the *techniques and economic organisation* related to their *manufacturing*.

In the ontology on fig.4, BEADS is a sub-class of NATURE OF THE CORPUS OBJECTS, beads have a SIZE (size and dimension are synonym terms), the term process refers to PROCESSING CHAIN, HARD STONE is a sub-class of MATERIAL, DURATION and SKILL LEVEL are related to a PROCESSING CHAIN. TECHNICO-ECOMIC STUDY is a sub-class of SCOPE OF THE STUDY.

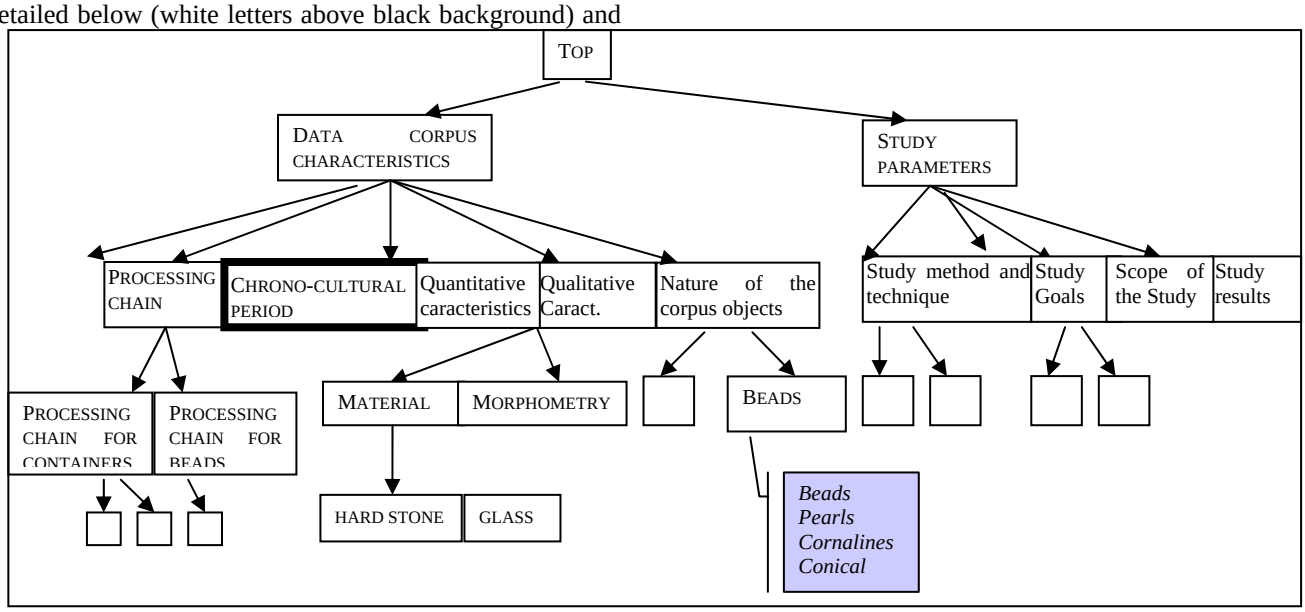


FIG. 4 – *Main concepts of the ontology. Each concept is associated several terms.*

Should we distinguish between precise object classes? No! For instance, hard stone beads, pearls, cornelians (which refer to beads made of cornelian), conical beads, classical harappan (which both refer to the origin, the form and the size), glass beads, could be sub-classes of BEADS. These distinctions were considered as useless. First, these categories refer to very specific classifications depending on the material, the form and the origin of the beads. Then, users are not supposed to browse an archaelogical database collecting descriptions, but scientific analyses made about a collection. Moreover, debates exist between experts about the relevance, frontiers and nature of these classifications. As a consequence, neither BEADS is decomposed into sub-classes, nor SIZE nor FORM is detailed.

This choice has been followed as a guideline for any concept decomposition. When subclass definitions are not motivated, all the terms referring to subclasses are associated to the main concept (for instance, cornelians or conicals are considered as terms referring to BEADS concept, whereas cornelian is a particular kind of hard stone).

4.3.2 Initial propositions describing the scientific approach

Initial propositions about the methodology followed by the researcher contain concepts of a different nature of and raise new issues. For instance, in fig.5, the proposition "analysis of the course of action" is three-fold:

The picture illustrates the proposition (the 6 stages of the knapping method in Cambay).

Title : The course of action is structured by a method.

A method is defined as an ordered set of knapping gestures. At Cambay, the knapping method allowing to create a pre-form from a rough-out includes 6 stages.

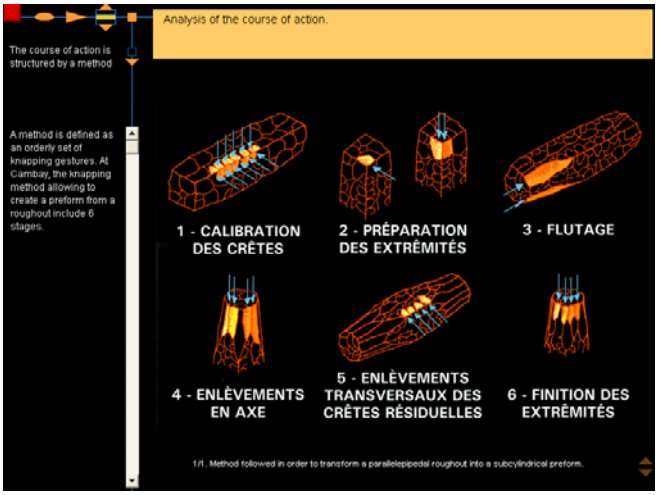


FIG. 5 – *Display of an initial proposition about the methodology followed in the study*

Here most of the terms lead to the definition of meta-data. "Method" is used to refer to the concept KNAPPING METHOD, which is defined in this proposition. KNAPPING is one of the processes of the PROCESSING CHAIN FOR BEADS. KNAPPING METHOD and KNAPPING GESTURES are to be added in the ontology and related to this concept. This is a first issue concerning knowledge representation in our model: what is the status of COURSE OF ACTIONS, METHOD and GESTURES for a given process? Then another issue is "should we define KNAPPING STAGE as a concept and each of the KNAPPING METHOD STAGES as sub-classes of this concept?" A last question is induced by the relations (between COURSE OF ACTION and METHOD, METHOD and GESTURE) that can be identified in the proposition. Archaeologists wished that the proposition should be identified as containing a DEFINITION of KNAPPING METHOD.

4.4 Questions raised by concept definitions when annotating interpretative propositions

In addition to a title and a comment, interpretative propositions are presented together with the titles of their antecedent propositions. On fig.6, the 4 propositions listed on the left part of the screen are antecedent propositions. These 4 assertions conclude to the current proposition "analysis of the course of action". The first initial proposition mentioned here is presented on fig.6.

Let's consider now the concepts that could be identified in the comment of this proposition.

Analysis of the course of action

The course of knapping sequences was noted in terms of the succession of operations and their temporal distribution. It was described and coded with video films and then treated with the program Kronos. This program, developed by A. Kergelen, permits a temporal analysis of the succession of actions which are retranscribed into sequences in the form of a diagram.

Titles of the antecedent propositions are the following ones:

The course of action is structured by a method.

The course of action is analyzed according to knapping strategies.

The course of action is analyzed according to the knapping sequences.

The course of action is analyzed according to the temporal structures.

These titles obviously are complementary to the comments. They provide additional information that is useful to get a synthetic view of the interpretation made in this proposition. This property gave us the idea to define some "**propagation rules**" for the annotations. Most of the concepts associated to a more precise proposition (antecedent) are candidate annotations for its parent propositions. For instance here, the concept KNAPPING METHOD (coming for the proposition in fig.5 that is the first one mentioned on the left of fig.6) could also be a relevant meta-data for the proposition "analysis of the course of action".

The issues raided by interpretative propositions are numerous:

- like for initial propositions, **should all terms be considered as indexing concepts**? For instance here, is the Kronos system a good meta-data?

- **Should we propagate some of the more relevant concepts of antecedent propositions towards successor propositions**? We have identified 3 rules that we have to evaluate.

- Concerning the nature of the concepts and their situation in the ontology, we can notice again two main kinds of concepts: (1) concepts related to the processing chain (coloured with light grey), like knapping sequences, course of knapping sequences, operations, temporal distribution); (2) concepts related to the techniques used to measure and analyse the course of knapping sequences (in dark grey): program, video film, diagram, temporal analysis. The first of concepts are under PROCESSING CHAIN, whereas the second set is under STUDY PARAMETER.

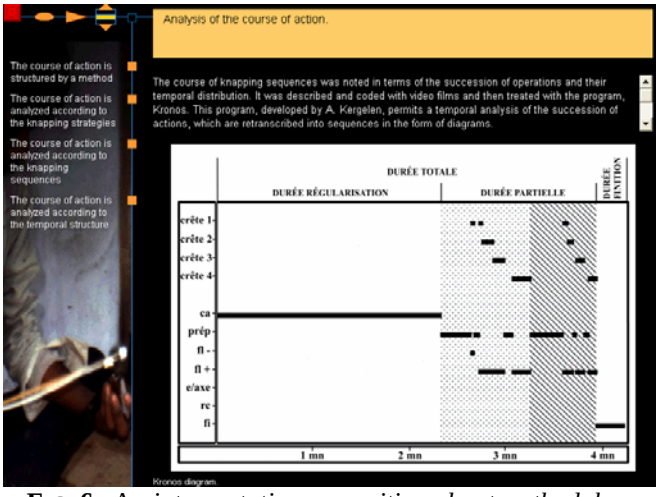


FIG. 6 –A*n interpretative proposition about methodology*

4.5 Current state of the model

The domain model is currently far from being an ontology. It covers 5 sub-domains of the archaeology of techniques with about 200 concepts and 500 terms, mainly about the production of ceramics and beads or about the description of archaeological objects. This model is currently closer to a linguistically grounded structured vocabulary that describes archaeological objects, periods and locations. Surprisingly, very few concepts have been defined to describe the scientific reasoning, the archaeological methods and techniques.

Moreover, concepts are not fully defined and they have not yet been properly differentiated and normalized. As a consequence, many semantic relations, required for concept definitions, are missing. The main concepts are shown on fig.4. Although concepts in this model have a different status from the one they have in the CRM model [7], [9], existing concept definitions in the CRM can be used as a reference to save time in concept definition. In the following, we will show how the CRM model can be used to improve the structure of our domain model. After reporting some of the limitations of reuse, we will focus on the definition of concepts related to time periods.

5 Reusing CIDOC – CRM

5.1 Mapping CRM with our model

CRM has been defined mainly to list and describe cultural objects, their geographical, human and temporal origin. Cultural objects generally are sub-classes of CRM:physical-object. In Arkeotek, the corpus of a study is made of objects which could easily be related to **CRM:physical-object**. But as long as objects are considered in Arkeotek only because a scientific study about this type of object is reported in a document, our model insists more on the type of the objects under study. Another major difference is that CRM does not care about how these objects have been identified, dated, how they were used or manufactured in the past, whereas these features are the core of scientific results in Arkeotek.

Another difficulty comes from all the concepts needed to describe techniques and methods used in archaeology when studying past techniques (right part of the diagram in fig.4). CRM is not concerned with most of these concepts. They do contribute to describe objects but rather the kind of analyses that archaeologists may carry out in order to identify the nature, the material of objects, how they have been manufactured, where and when, by which people, ...

The third difficulty is the large gap that separates the abstraction level of the terms and concepts currently represented in our model compared with the abstraction level of those defined in CRM. A direct connection is not possible and many intermediary nodes are to be identified. For instance, most of the texts refer to beads or potteries or tools. A classification is required under **CRM**:**physical**-object to differentiate relevant object classes and then set concepts like BEADS, knives, potteries, ...

Nevertheless, we consider that the current status of our model is only provisional and that it should evolve towards a better grounded and better structured model. Because some parts of the CIDOC-CRM appear already almost applicable, this ontology is a good candidate to be a core ontology that we would detail and adapt to our needs. Let's consider how it could contribute to improve some concept representation, like the description of historical periods.

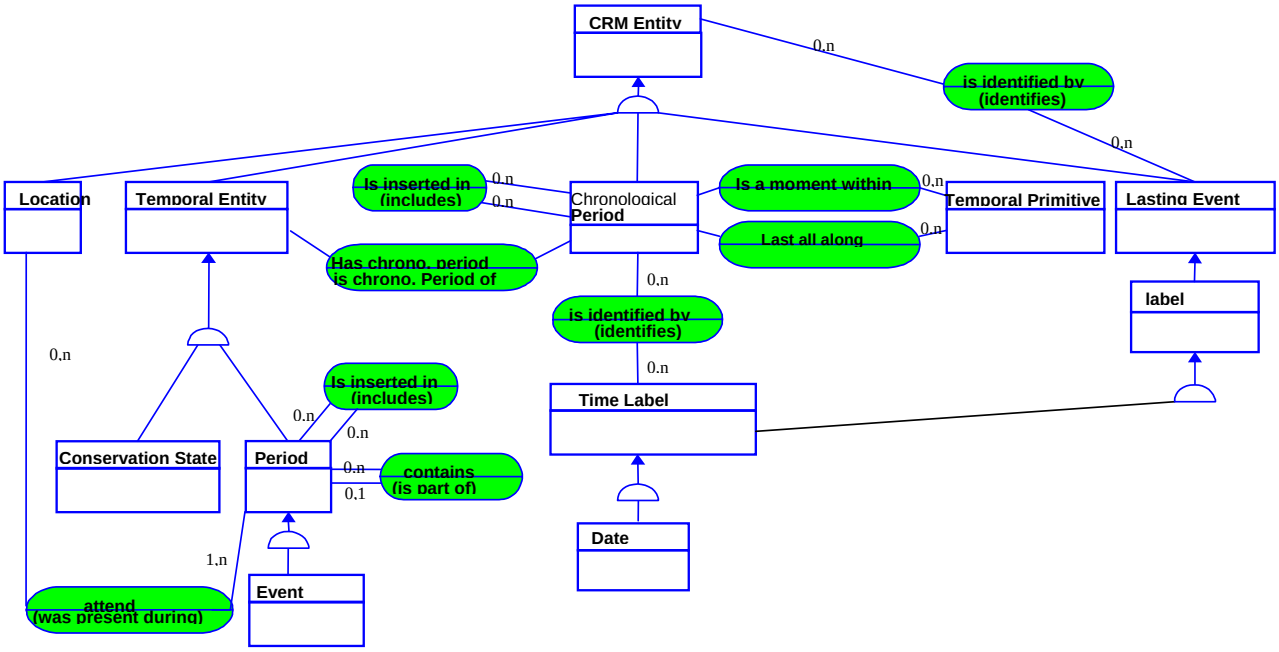


FIG 7 – *Concepts related to the definition of Chronological periods in the CRM 4.3 [7]*

5.2 Reusing the "period" concept

Together with the CRM, its authors propose a method [9] that would facilitate the integration and confrontation of scientific opinions and that would make explicit the relationships existing between multiple period definitions. This position is relevant in a scientific field like archaeology, where diverging opinions may exist according to various analysis scopes.

In the CRM ontology (extract in fig.7), **Period** is a kind of **Temporal Entity** defined by distinct ("defining") criteria based on the archaeological contexts rather than by time and place [10]. The domain experts that built the Arkeotek model also rejected a time and place characterization of periods because it would fall into debates where experts may have diverging views. Experts insisted that periods are agreed upon a research community which shares similar research objects and topics. So the "defining criteria" suggested in CRM are adequate with this remark. These criteria are the types of phenomena or interrelated phenomena that determine the unity and identify the cultural continuity of a period.

Because the Arkeotek model does not pretend to be an ontology, experts thought it was not worth describing each period (which is done with "general characteristics" in CRM). So we will enrich our "period" definition by defining (cultural) criteria, but not with all the general characteristics suggested in CRM for period.

As shown in the excerpt of the CRM entity model on fig.8, events are related to periods because they may happen *during, at the beginning or at the end of* a period. But even starting or terminating events do not define the periods themselves as cultural phenomena.

Another nice idea that we will borrow from the CRM model is the distinction made between a time period and its "extent", described as a temporal evolution over space [10]. A period extent includes temporal and spatial information. Indeed, the Arkeotek domain experts insisted on the close relationship linking a period with a geographical area. As shown on fig.8, the spatiotemporal extent of a period is constrained by spatial, temporal and spatiotemporal relations resulting from scientific observations and evidences [10].

5.3 Beyond reuse: which are the actual goals of the models?

The example of how periods can be represented will help us underline the difference of focus between the CIDOC-CRM and our model.

On the Arkeotek side, some archaeologists have criticized the way scientific papers present results in their domain. They want to show that a rigorous (and costly) rewriting of these papers and theses into the SCD format will help clarify the message, better identify the inferences and efficiently localize some scientific results for further reference and reuse. The goal is ambitious: it is to better capitalize the domain within the scientific community. A part of the domain knowledge is supposed to remain in natural language form in propositions, another part results from the organisation of these propositions into SCD documents thanks to the antecedence relations between propositions. The semantic annotation of these documents with an ontology seems to be an additional layer of "light" formalisation

of the information available in natural language. So this ontology is closer to a structured thesaurus than to a full theory.

On the other hand, CRM results from a sharing effort in the cultural domain, where data-bases already existed with their own data models [8]. CRM integrates most of the schemas of the data bases that should be made compatible. It also reflects an effort from domain experts to reach a consensual view on their concepts, so that search in those data-bases could be unified.

The scope is much different from Arkeotek's goal: most of the domain knowledge remains the domain specialist's know-how and skills about the data-bases. A foundational part lies in a core model, the CRM, which helps at indexing data and documents about cultural heritage. But there is no epistemic ambition: the scientific knowledge is not supposed to be significantly modified. CRM provides a means to reach easily some information (location, origin, manufacturing date or age …) about a specific cultural object or document.

In spite of this divergences of motivation and scope, CRM as a core model could improve the quality of our annotation model. The higher quality the models will have, the better.

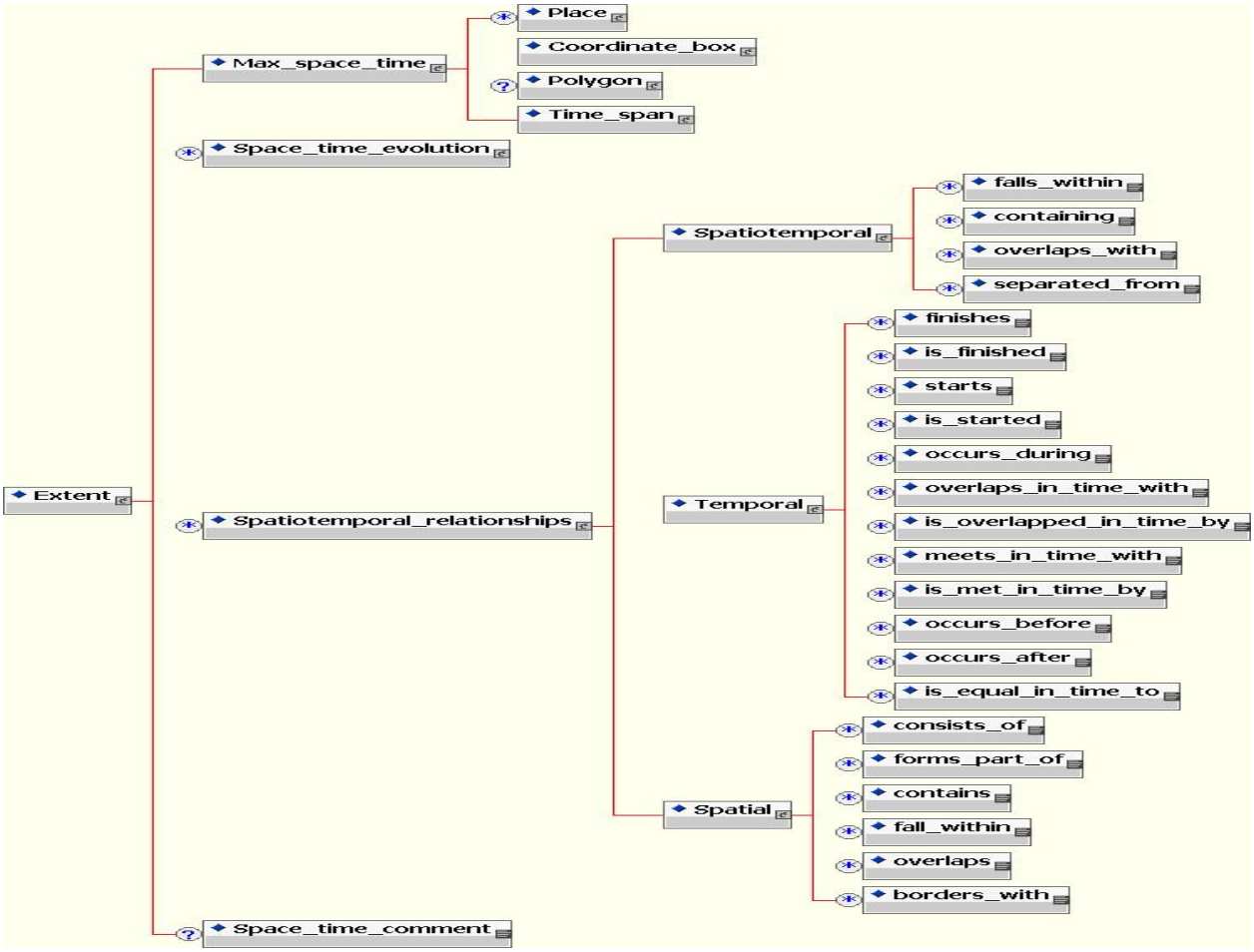


FIG. 8 – *The spatiotemporal relations that "extend" the period concept in the CRM ontology*

6 Conclusion

The Arkeotek project proposes to improve knowledge management in human sciences by structuring and annotating the databases that gather the documents and objects produced by the scientific community. A first layer makes explicit the argumentative structure of documents with the SCD format, and a second layer characterizes their content with some domain knowledge used for semantic annotation. The maintenance of these two models must be coordinated.

The domain model can be considered as a terminological ontology, but it still needs to be normalized to get a more precise definition of concepts. We have shown in this paper that the methods and concepts elaborated by the

CIDOC CRM project can be helpful when modeling knowledge in archaeology like required for Arkeotek. We do not want our model to be a theory of this research field, to state about identities for instance. The model should rather reflect conveniently the concepts that can be identified through the language used in the documents to be annotated. An effort is still required to better structure the model, to normalize it, to get validated concept definitions, and to enrich it with more concepts related to the scientific approaches (techniques, methods and theories) used in archaeology of techniques.

Whatever the quality of the ontology, collaboration between researchers in knowledge engineering and archaeology is here fundamental for developing efficient tools. The overall validation of the approach is quite complex and requires to proceed the cross-disciplinary experiments. A major issue is that properly re-writing documents according to the SCD format is a complex and time-consuming task. Evaluation must involve users (domain scientists), SCD specialists and tool designers but also a human factor analyst that will bring an external look on the approach.

Acknowledgements

We thank Blanche Barthélémy de Saizieu ("Préhistoire et Technologie", MAE, Nanterre) and Patricia Guillermain (UTM, Toulouse) for their contribution as experts to design the domain model. The Arkeotek Project is managed and proposed by V. Roux ("Préhistoire et Technologie", MAE, Nanterre) and J.C. Gardin who proposed the logicist theory and adapted it to guide scientific rewriting in archaeology. P. Blasco defined the Episteme tool-box. The project is financed by the CNRS-MSH, the funding CNRS program about Information Society and the Arkeotek Association.

References

[1] AUSSENAC-GILLES N., BIÉBOW B., SZULMAN N., 2000, Revisiting Ontology Design: a method based on corpus analysis. *Knowledge engineering and knowledge management: methods, models and tools, Proc EKAW 2000*. Juan-Les-Pins (F). LNAI 1937. Berlin: Springer Verlag. 172-188.

[2] AUSSENAC-GILLES N., ROUX V, de SAIZIEU B., BLASCO P., 2005, Ontologies dédiées à la consultation de documents structurés selon un modèle logico-sémantique. In *Actes du colloque de clôture du programme Société de l'Information*. Lyon (F), 19-21 mai 2005. 13-16.

[3] AUSSENAC-GILLES N., ROUX V., BLASCO P., 2006, The Arkeotek project: structuring scientific reasoning and documents to manage scientific knowledge. In Proc. of the workshop on Indexing and Knowledge in Human Sciences, Nantes (F), June 2006. http://www.sdc2006.org/cdrom/contributions/Aussenac_ICSH2006.pdf.

[4] BAZIZ M., BOUGHANEM M., AUSSENAC-GILLES N., 2004, Semantic representation of Documents by Ontology-Document Mapping. In Proceedings of the 2nd ACM SIGIR Workshop *on Semantic Web and Information Retrieval (SWIR 2004)*. Sheffield (UK), July 25-29th 2004.

[5] BOILEAU M.-C., 2005, *Production et distribution des céramiques au IIIe millénaire en Syrie du Nord-Est*, Paris, Collection Référentiels, Editions de la Maison des Sciences de l'Homme - Editions Epistèmes

[6] BOURIGAULT D., 2002, UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de TALN 2002, Nancy (F), 75-84.

[7] CIDOC – CRM web site (ontology in RDFs format): http://cidoc.ics.forth.gr/rdfs/Cidoc_v4.2.rdfs

[8] CIRAVEGNA F., DINGLI A., PETRELLI D., WILKS Y., 2002, User-system cooperation in document annotation based on information extraction. *In proceedings of the 13th International conference in Knowledge Engineering and Knowledge Management, EKAW 2002*. Springer Verlag, LNAI.

[9] DOER M., 2003, The CIDOC CRM, An ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*. Vol. 24 (3).

[10] DOER M., KRITSOTAKI A., STEAD S., 2004, Which period is it? A methodology to create Thesauri of Historical Periods. *CAA 2004*. http://cidoc.ics.forth.gr/docs/parousCAA_4.ppt

[11] GARDIN J.C., 1991, *Le calcul et la raison. Essais sur la formalisation du discours savant*. Paris : Editions de l'EHESS.

[12] GARDIN J.C., 1998, *Prospections archéologiques en Bactriane (1974-1978), description des sites et notes de synthèse*. Paris : Editions Recherche sur les civilisations.

[13] GARDIN J-C., ROUX V., 2004, The Arkeotek project: a European network of knowledge bases in the archaeology of techniques. *Archeologia e Calcolatori*, 15, 25-40.

[14] MIRZAEE V., IVERSON L., HAMIDZADEH B., 2004, Towards Ontological Modelling of Historical Documents. *7th PROTÉGÉ International Conference*. July 2004. Bethesda (Ma, USA).

[15] POESIO M., 2005Domain modelling and NLP : Formal ontologies ? Lexica? Or a bit of both? *Applied Ontology* (1) 1. 27-34.

[16] ROUX V. (sous la dir.) 2000. *Cornaline de l'Inde. Des pratiques techniques de Cambay aux techno-systèmes de l'Indus.* Paris : Editions de la Maison des Sciences de l'Homme - Editions Epistèmes, 545 p. Cédérom bilingue (format SCD).

[17] ROUX V. avec la participation de Ph. Blasco, 2004. Faciliter la consultation de textes scientifiques. Nouvelles pratiques éditoriales. *Hermès*, Critique de la raison numérique, CNRS éditions, 39, 151-159.