Process modeling for Humanities: tracing and analyzing scientific processes

Charlotte Hug, Camille Salinesi, Rebecca Deneckere, Stéphane Lamasse

Dans **Revive the Past : Proceedings of the 39th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Beijing, China, 12-16 April 2011** - *Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA 2011)*, Beijing : Chine (2011) - http://hal-paris1.archives-ouvertes.fr/hal-00662699

Process modelling for Humanities: tracing and analysing scientific processes

Authors:
Dr Charlotte Hug, Prof Camille Salinesi, Dr Rebecca Deneckère
Centre de Recherche en Informatique, University of Paris 1 Panthéon-Sorbonne, France

Dr Stéphane Lamassé
Laboratoire de Médiévistique Occidentale de Paris, University of Paris 1 Panthéon-Sorbonne, France

{Charlotte.Hug, Camille.Salinesi, Rebecca.Deneckere, Stephane.Lamasse}@univ-paris1.fr

Abstract
This paper concerns epistemology and the understanding of research processes in Humanities, such as Archaeology. We believe that to properly understand research processes, it is essential to trace them. The collected traces depend on the process model established, which has to be as accurate as possible to exhaustively record the traces. In this paper, we briefly explain why the existing process models for Humanities are not sufficient to represent traces. We then present different process models from Information Systems Engineering that allow tracing processes according to different perspectives such as activities, decisions or strategies. We assume these process models can be useful to represent research processes in Humanities coherently and thoroughly.

Keywords: process modelling, research processes, trace.

1. Introduction

Humanities researchers use databases to structure, store, retrieve, share, and analyse their data (Doran and Hodson 1975). For instance, better understanding the possible correlation and dependences between variables requires quantitative analysis of large amounts of data. Managing such quantities of data proved much easier when storing them in an adequately structured database. In this context, humanities researchers need to encode the initial data, store them in databases, then, they can run different tests depending on their goals: uncover the organizational structure of data, identify significant trends and patterns, show correlations or dependencies between variables. In fact, experience shows that such databases can be exploited to achieve many different goals, following different ways of working that implement different research strategies and use different statistic means (Doran and Hodson 1975).

This approach is relatively new in humanities, and therefore the primary concern is that of the scientific validity of the processes achieved with databases. This question cannot be solved without modelling the scientific processes employed by the researchers while working using these databases, which in its turn raises two other questions: (a) how to trace those processes, and (b) how to analyse them. Keeping track of scientific processes goes far beyond listing the

sequences of statistical tests used by the scientists. Indeed, what has been done is just a facet of why this was done. The problem of modelling the humanities scientific processes in this context is manifold: what are the good formalisms? How to avoid being too normative (every scientist has her/his own way of working)? How to deal with the cognitive tasks of the scientists (the goals they have in mind) starting from the technical level (we can only record which statistical tools were used, when and on which data)? How to deal with processes in a bottom up fashion (i.e. starting from traces to models and not the other way around)?
There is to our knowledge no definitive answer to these questions today. However, the challenges are important: not only being able to trace and analyse scientific processes will allow to model them, but also will it help to demonstrate the methodological processes used in research projects, compare methodologies, repeat them, and even, as done in other domains of process modelling, improve them (SEI 2010).
This paper focuses on trace modelling that is the conceptual definition of the data structure of the traces of use of databases and data analysis. We know by experience that the modelling of processes is tightly related to the models of the traces. In other words, we build process models starting from process traces. However, it is meaningless to store processes traces in a way that is not consistent with the way we actually want to model the processes. The process modelling literature shows that processes can be modelled in different ways depending on the goal (Rolland 1998). The first goal of process modelling is to help people enact processes and get the expected results in a systemic way. Such guidance requires machine understandable process models and engines to interpret them and control their enactment. It is called the prescriptive aspect of process modelling. Another process modelling goal is to understand what has been done, by who, when, why and what could have been done differently. We call this the descriptive aspect of process modelling (Rolland 1998) and it is comparable to process monitoring. The third goal of process modelling permits to establish a link between the actions taken and the decisions made before these actions. It is the explanatory aspect of process modelling (Rolland 1998).
Our research focuses on the formalisation of the processes followed by humanities researchers. The issue is that humanities research processes cannot be treated as business or scientific workflows, that is to say as models specifying pre-established sequences of tasks that will be applied in a conservative way. Indeed, humanities processes are creative and non-predictable; they have intrinsic variability and emerging features. Our goal consists then in defining an adequate scientific process modelling language that will allow tracing the humanities research processes as completely as possible. Our work focuses on the descriptive and explanatory aspects of humanities research processes; we squarely refuse to describe process models to stipulate how humanities researchers have to work.
Section 2 gives an overview of modelling in humanities and section 3 presents process modelling in information systems engineering and how it can be applied to model humanities research processes. Section 4 concludes this paper.

2. Process modelling in humanities

Process modelling is a topic that has already been tackled in the context of humanities. Ellis and Haugan’s research was to our knowledge the first work focusing on the processes applied by engineers and research scientists to search for information (Ellis and Haugan 1997) (the model they proposed was revised by (Meho and Tibbo 2003)). Later on, Lönnqvist investigated the modelling of information-seeking behaviour of social scientists (Lönnqvist 2007), Hodge studied research process for humanities scholars for digital archiving (Hodge 2000), and Constantopoulos and Dallas defined a process model for digital curation (Constantopoulos and Dallas 2007). All the models developed in these works are based on experts’ interviews and surveys. However, the processes described in these papers are defined at a macroscopic level. Being abstract, it is very difficult to get a deep understanding of the underlying processes starting from the models. Besides, some of these process models are only described textually. This raises a series of problems concerning their understandability (how to navigate through the different parts of the processes) and correctness (how to analyse a process model when it is described in an unstructured form). Figure 1 shows the process model developed in (Meho and Tibbo 2003) to describe the behaviour of information-seeking social scientists. This process is divided in four phases: Searching, Accessing, Processing and Ending, which in turn are composed of multiple stages described textually. For example, the processing phase covers extracting activities, differentiating activities, verification activities, etc.

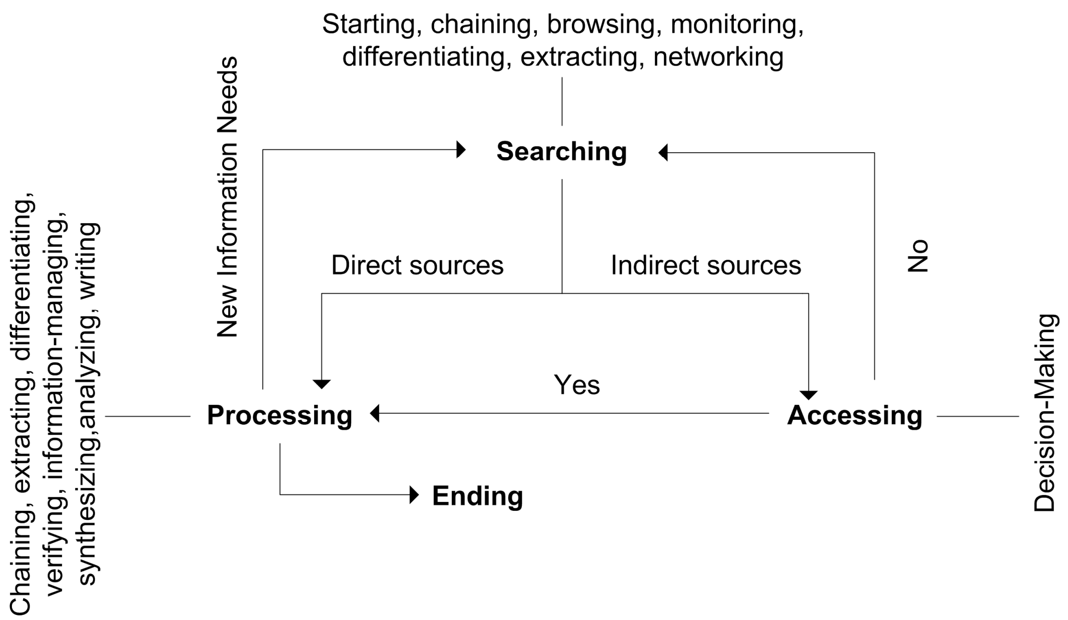

Figure 1: Model of the information-seeking behaviour of academic social scientists (Meho and Tibbo 2003)

The absence of a formal notation is obvious. For instance, the model does not distinguish between phases, tasks, stages or activities. The same representation (arrow) is used to denote both control flows (e.g.: “yes”, “no”) and information flows (e.g.: “new information needs”). The model represents an ending point (“ending”), but no starting point. Not only these concerns raise questions on the usability of the model (how do you interpret a process when you do not know where it starts?) but also in terms of validity (can I trust a process is consistent if the notation used to model it is not?). Using a metamodel that specifies (a) the concepts employed to model processes and (b) their representation helps avoiding such issues.

(Terras 2005) describes how she carried out linguistic analysis on experts’ speeches to compute an agent based system to help papyrologists deciphering ink and stylus texts. Figure 2 presents the process model defined in (Terras 2005) to describe the different activities achieved by the experts while deciphering ancient documents. This process was defined to assist the experts through computer means.

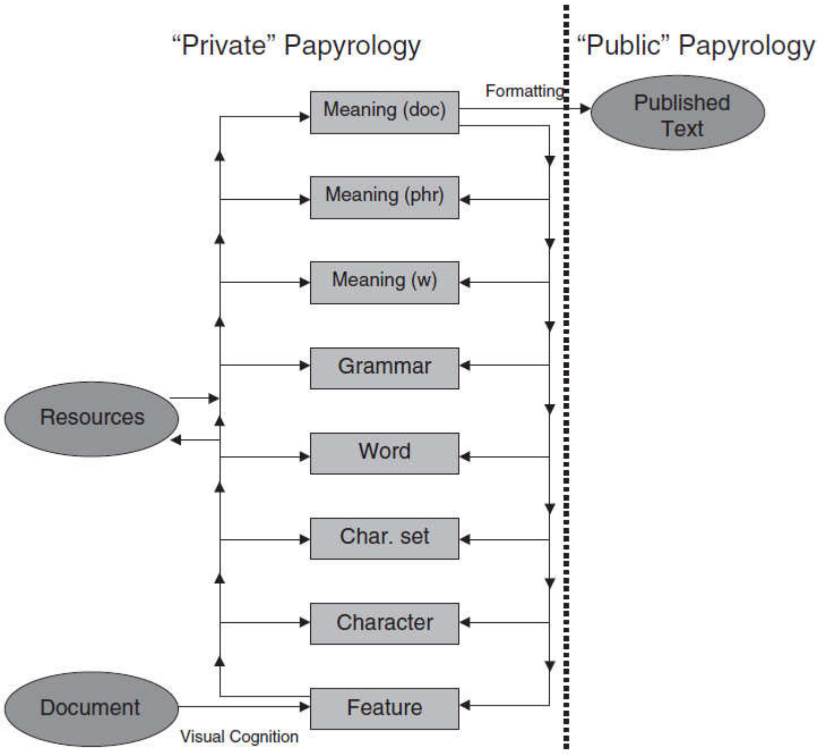

Figure 2: Process model describing of how experts read ancient texts (Terras 2005).

Contrary to the process model shown in Figure 1, this model was specified using a formal notation. In fact, the process is computer understandable, which is necessary to introduce automation in the process enactment. The choice of the metamodel is directly related to the goal of assisting researchers in their work, and therefore mainly driven by its prescriptive aspect. However, to keep track of what researchers do, one needs a process metamodel that has descriptive qualities. For instance, the process model shown in Figure 2 does not include any explanation about the process rationale. The explanatory aspect of the process, that is to say the association between actions traced back and the reasoning beneath them would be more useful information than the sequence of actions.

The European project DARIAH (DARIAH 2011) aims at developing a digital infrastructure to support information and communication technologies based research practices in Arts and Humanities. This project includes a work package on scientific process modelling. It is still an ongoing project and there is no concrete proposal but we believe that the fact that this work package exists confirms that process modelling is a strong necessity in Humanities and that satisfying solutions are still needed.

The existing process models are useful as they give an overall view of research processes in humanities. They were defined for a specific purpose, but at a high level of abstraction that does not allow a deep understanding of the processes. Such process models are comparable to lifecycles. Moreover, the process models do not use a specific and precise modelling language, which can lead to introduce imprecision in the models. In addition, the studied models do not consider the explanatory aspect of the processes, which is necessary if we want to understand the whole researcher's reasoning process, the decisions as well as the actions.

At last, our goal is not to compute process models to automate the work of humanities researchers. We want to provide a full and exhaustive view of the path taken by the researchers to better understand their way of thinking and doing.

A lot of research has been achieved on the process modelling topic in the context of the information systems engineering domain. We believe the results obtained in this area can highly contribute to resolve the issues raised in the context of humanities.

3. Process modelling in information systems engineering
In the information systems engineering area, a process is defined as “a set of correlated or interactive activities that transforms inputs into outputs” (ISO 2008). A process model describes common characteristics of a set of processes. Different types of process models allow representing different perspectives of the same process: activity-oriented, product-oriented, decision-oriented, context-oriented and strategy-oriented. In this section we present the different types of process models using the example of data analysis.

3.1. Activity-oriented process models
Activity-oriented process models focus on interrelated set of activities conducted for the specific purpose of defining information systems engineering artefacts called “products” (Rolland 1998).
Figure 3 (a) presents such an activity-oriented process model. The process model is represented as a graph with nodes that represent activities, and transitions that represent the sequence of activities. The start and end of the process are respectively represented by the black circle and by the encircled black circle, which are two specific kinds of nodes in the model.

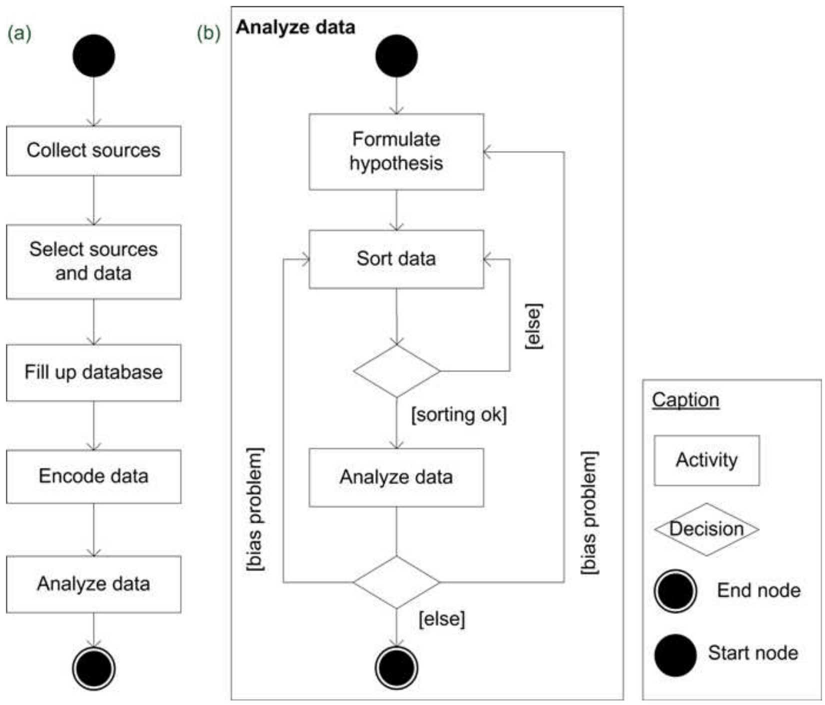

Figure 3: Activity-oriented process model describing the analysis of data.

The model shows that the process is composed of five activities from collecting sources to analyzing data. As this view of the process is macroscopic, a more detailed view can be provided using refinement mechanisms. The “Analyze data” activity is decomposed in sub activities shown in Figure 3 (b). Analysing data involves several activities such as “Formulate hypothesis”, “Sort data”, etc. that can be decomposed into several different sequences formalized by the branches represented by the diamond nodes. The activities are defined linearly but as the complex control structure shows it, it is possible to go back to previous activities depending on the conditions.
Activity-oriented process models focus on the “what”, that is to say on the sequence of the actions carried out (or to carry out). Decisions can be introduced during the process and lead

to different actions. The process models can also include the actors who carry out the actions (not represented here).
This type of process models has well known limitations. In particular, one drawback is that the description of the processes is linear. This is particularly adapted to business activities where the actions are defined precisely and have to be executed in a given order to get the needed result. In fact, it may even suit scientific processes that implement well-defined protocols that need be controlled. However, research in Humanities cannot be only represented as a pre-defined sequence of actions. This approach would be too reductive to fully describe the complexity and the nature of these research processes. We believe that activity-oriented process models are not sufficient to exhaustively describe research processes in Humanities.

3.2. Product-oriented process models
The purpose of product-oriented process models is to represent processes through the evolution of the products that they consume, process and produce (Rolland 1993).
The process model shown in Figure 4 describes the process of “analyzing data” but in the perspective of the “Data” product. The process is modelled as a graph, which nodes represent the different states of the data throughout the process, and arrows, transitions between these states. We can call this model “state transitions” diagram. Like in the activity-oriented process models, we can define backward and forward transitions between the states depending on the condition.

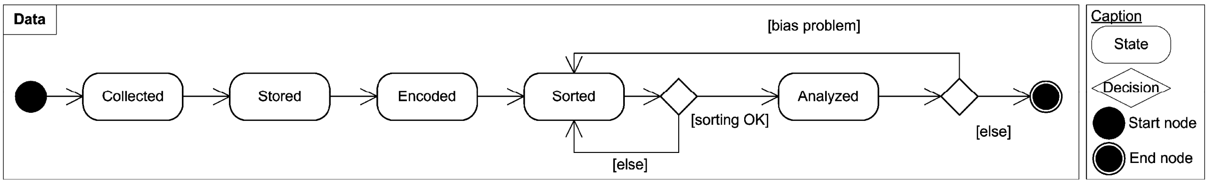

Figure 4: Product-oriented process model describing the analysis of data.

Product-oriented process models present another perspective of the process, complementary to the activity-oriented process models. However, like activity-oriented process models, product-oriented process models are linear, which poses a problem when dealing with humanities processes. Besides, they do not tackle the explanatory aspect of the process, which is a first-class goal in the epistemological context.

3.3. Decision-oriented process models
Decision-oriented process models allow representing the series of decisions that lead to product transformations (Rolland 1998).
Figure 5 presents the different possible alternatives to answer the issue “Which statistical model should I use?” The model is a graph, which nodes are decision-related concepts, in particular issues that call for decisions, alternative answers, arguments supporting these alternatives, and steps to make the decisions when facing the issues.

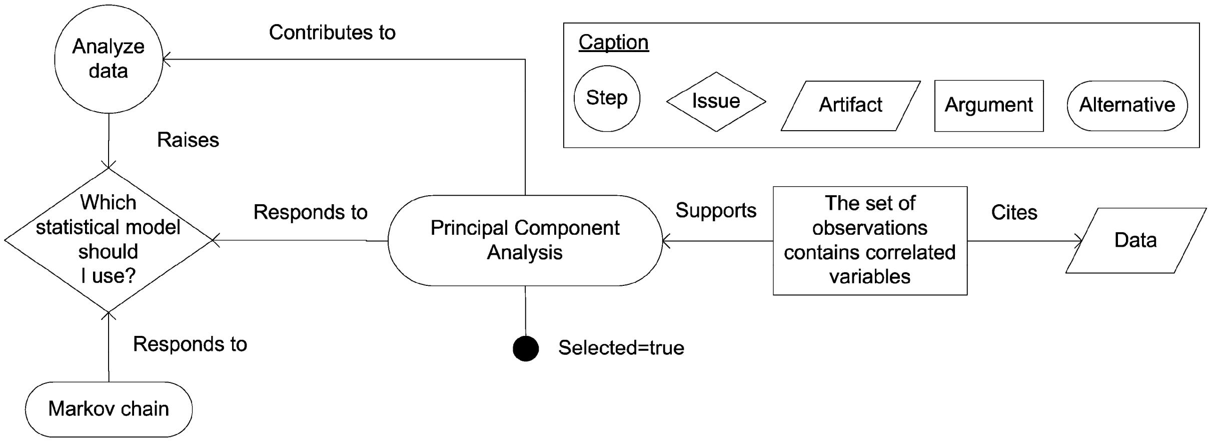

Figure 5: Decision-oriented process model describing the analysis of data.

Different alternatives are proposed to solve the issue, such as "Markov Chain" or "Principal Component Analysis". Arguments can support or object to alternatives, as "The set of observations contains…" based on artefacts, the data in our example. When an argument is selected, it contributes to the progress of the step (black circle in Figure 5).
Decision-oriented process models allow representing the whole process of decision according to the different steps of the process, that is to say the activities. Moreover, decision-oriented process models help understanding why decisions are made, what were the other options available, and how the decisions impacted the continuation of the process.

3.4. Context-oriented process models

Context-oriented process models allow representing the situation and the intention of an actor at a given moment of the project (Rolland 1998). Context-oriented process models "look upon each process as being in a subjectively perceived situation upon which is looked upon with some specific intention" (Rolland 1998). The NATURE project (Rolland 1994) (Jarke et al. 1999), defined a language and a formalism to specify context-oriented process models. The formalism was inspired by artificial intelligence in which expert systems start with goals to reason about problems. The actual reasoning, achieved using rules, depends on the context. The NATURE process metamodel allows representing decision contexts, which are defined as the coupling of a situation and an intention. The situation specifies when the decision can be made; the intention represents why it is made. The metamodel proposes different types of decision contexts, which allow tackling different kinds of decisions depending on whether they are choices to be made between alternatives, or plans to be drawn to organise other decisions (Plihon 1996).
Figure 6 presents the context-oriented process model describing the analysis of data as a tree of contexts and sub contexts. Each node in the tree is a decision context. Two kinds of nodes are represented: fork and rake, which respectively provide details about choice and plan decision contexts.

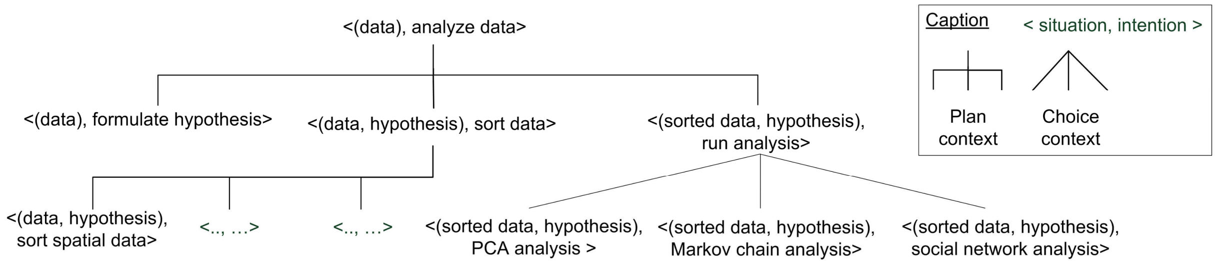

Figure 6: Context-oriented process model describing the analysis of data.

The root of the tree that represents the data analysis process models the context where the researcher has the data at her/his disposal (situation) and she/he wants to analyze this data (intention). To meet this intention, the researcher has to set up a work plan that may involve formulating hypotheses, sorting data and running the analysis. The plan can be very complex and involve retro-actions, loops, repetitions, and of course decisions on how each of the action behind each decision is achieved in its turn. The plan has three nodes that represent decision contexts that can be decomposed in their turn. There are different kinds of context: the execution of a plan context will lead to the execution of the sequence of all the sub-contexts (< (data, hypothesis), sort data > in Figure 6), the execution of a choice context will lead to the selection of one of the defined sub-contexts (< (sorted data, hypothesis), run analysis >). Contexts that are directly executed are called executable context, such as < (data), formulate hypothesis >.

3.5. Strategy-oriented process models

Strategy-oriented process models allow representing, in a single representation, multi-processes, i.e. processes that have a unique goal, but can be achieved in many different ways (Rolland et al. 1999). An intention captures the notion of a task that the application engineer intends to perform whereas the strategy is the manner in which the intention can be achieved (Rolland et al. 1999).

Figure 7 presents the strategy-oriented process model describing the analysis of data. A process map is an oriented graph in which nodes are the intentions underlying the process, and edges (so-called "strategies") indicate how the intentions can be achieved. When an intention can be achieved in different ways, the corresponding strategies are specified in the graph with the intention as the target node. The map has a goal, begins with the intention "Start" and ends with the intention "Stop".

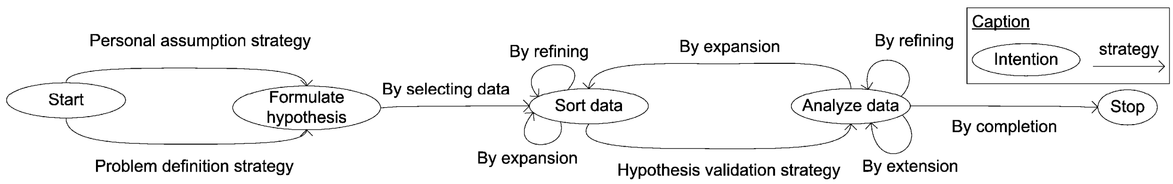

Figure 7: Strategy-oriented process model describing the analysis of data.

This example shows that a researcher that performs data analysis may have three intentions in mind: formulate analysis, sort data and analyze data, knowing that each intention can be achieved using different strategies in different orders. The first intention of the researcher is the hypothesis formulation. This intention can be achieved using two different strategies: by personal assumption or by problem definition. Once a hypothesis has been formulated, the researcher can proceed by sorting data, starting with the hypothesis, or proceed by keeping with hypothesis formulation e.g. using the strategy she/he has not yet used. Similarly, any data that has been sorted can be used to start analysis, but it can also be further sorted by refinement and expansion strategies (and then analyzed in a more detailed or complete way). Several combinations of intentions and strategies can be chosen from the model, hence reflecting the multiple nature of the process underneath.

Context and strategy-oriented process models allow creating rich and complex process models using abstraction mechanisms. This proved effective to "represent multi-facetted processes and plan different ways to elaborate products based on the notion of intention" (Rolland et al. 1999). By focusing on the why rather than what or when, these approaches that were initially designed to guide and trace engineering processes, proved flexible enough to adapt to other contexts such as Business Process Modelling (Salinesi et al. 2002), the design of decision making systems (Gam et al. 2006) or Business Analysis (Thevenet et al. 2006).

3.6. Synthesis

The information systems engineering literature provides many approaches to formalize process models. It is clear from this literature review that each metamodel covers a different perspective of the process (activity, product, decision, context and strategy) and that they are complementary in the sense that there is no formalism that allows dealing with all the facets of a process at the same time. Processes can then be seen under the perspective of different points of view. The specification of any process can therefore be very exhaustive, if necessary, by increasing the number of models depending on their relevance to the context of modelling and to the use of the model.

The different types of process models allow defining the descriptive aspect of process as well as the explanatory aspect, by using decision, context or strategy-oriented types of process models.

In this section we have only presented one example of each type of process model. However each type can be represented as a metamodel. A metamodel is a structured abstract model that allows defining infinity of more specific models complying with the corresponding metamodel. Each type of process model is associated with at least a given specific formalism that is strictly defined. So, defining an adequate process metamodel and an associated formalism allows specifying consistent process models.

We think process modelling for information systems engineering can be applied to humanities research process modelling. First, it allows describing the whole process, including different perspectives. These perspectives allow taking into account the descriptive aspect and the explanatory aspect of humanities research processes, that is to say what has been done and why. Second, each perspective of the process can be described in a precise manner, using a specific metamodel and formalism. The humanities research processes can then be suitably and exhaustively described.

Our approach is not part of Processual Archaeology. We do not want archaeologists, to follow a standard process that would fit every archaeological project. We believe each project, each team, each person, has its own ways of working. By modelling the ways of working using an adequate language and structure (metamodel) can help understanding how archaeologists work. The map process models can actually help representing the richness of a process including the many different ways to achieve a goal, as conducting a literature review or a lithic analysis for example. The process model can then be understood by the archaeologist him/herself or by any other person of the team or the project. Modelling the processes can also help the communication within a project.

4. Conclusion and future works

We need to keep track of humanities research processes to allow researchers to better understand their ways of working. Achieving this goal highly depends on the quality of the process traces modelling. However, existing humanities process models are not satisfying to record tracks in an exhaustive way. The major challenge in tracing processes is then to find an adequate modelling language that covers all the aspects needed when analysing the traces stored.

In this paper, we have presented five types of process models from information systems engineering domain to use them to represent humanities research processes. We think such types of process models can be a strong basis to achieve our goal.

We now need to thoroughly test the different process models with humanities researchers such as archaeologists. We will probably need to design new types of models to specify humanities research processes, based on the existing types, to describe the processes as precisely and completely as possible. This research will include the creation of a new process metamodel for humanities research process and a formalism to represent the models.

References

Constantopoulos, Panos, and Dallas Costis. 2008. “Aspects of a digital curation agenda for cultural heritage.” Paper presented at the IEEE International Conference on Distributed Human Machine Systems, Athens, Greece, March 9-12.

DARIAH. 2011. Digital Research Infrastructure for the Arts and Humanities, “Work package 8 – Technical work: Conceptual modelling”. Accessed May 18. http://www.dariah.eu/index.php?option=com_content&view=article&id=31&Itemid=35.

Doran, James E., and Frank R Hodson. 1975. Mathematics and Computers in Archaeology, Cambridge: Edinburgh: Edinburgh University Press.

Dowson, Mark. 1987. “Iteration in the Software Process: Review of the 3rd International Software Process Workshop.” In *Proceedings of the 9th international conference on Software Engineering (ICSE '87)*, 36-41, Los Alamitos, CA: IEEE Computer Society Press.

Ellis, David and Haugan Merete. 1997. “Modelling the information seeking patterns of engineers and research scientists in an industrial environment”, *Journal of Documentation*, 53, 4: 384 – 403. doi : 10.1108/EUM0000000007204

Gam, Ines, and Camille Salinesi. 2006. “A Requirement-driven Approach for Designing Data Warehouses.” Paper presented in Requirements Engineering: Foundation for Software Quality, Grand-Duchy of Luxembourg, Luxembourg, June 5-6.

Hodge, Gail M. 2000. “Best practices for digital archiving: an information life cycle approach”, *D-Lib Magazine*, 6, 1. doi: 10.1045/january2000-hodge.

ISO (International Organization for Standardization). 2008. “Introduction and support package: Guidance on the concept and use of the process approach for management systems”, SO/TC 176/SC 2/N544R3.

Jarke, Matthias, Colette Rolland, Alistair Sutcliffe, and Ralf Domges. 1999. *The NATURE of Requirements Engineering*. Aachen: Shaker Verlag.

Lönnqvist, Harriet. 2007. “The Research Processes of Humanities Scholars”, in *Advances in Library Administration and Organization*, 25, edited by Edward D. Garten, Delmus E. Williams, James M. Nyce and Sanna Talja, 175-202. Emerald Group Publishing Limited. doi: 10.1016/S0732-0671(07)25009-6.

Meho, Lokman I., and Helen R. Tibbo. 2003. “Modeling the information-seeking behavior of social scientists: Ellis's study revisited”, *Journal of the American Society for Information Science and Technology*, 54: 570 – 587. doi: 10.1002/asi.10244.

Plihon, Véronique. 1996. “Un environnement pour l'ingénierie des méthodes” PhD diss., Université Paris 1 Panthéon-Sorbonne.

Rolland, Colette, Naveen Prakash, and Adolphe Benjamen. 1999. “A multi-model view of process modelling”, *Requirements Engineering*, 4, 4: 169-187. doi: 10.1007/s007660050018.

Rolland, Colette. 1993. “Modeling the Requirements Engineering Process.” Paper presented at the Third European-Japanese Seminar on Information Modelling and Knowledge Bases, Budapest, Hungary.

Rolland, Colette. 1994. “A Contextual Approach to modeling the Requirements Engineering Process.” Paper presented at the Sixth International Conference on Software Engineering and Knowledge Engineering, Vilnius, Lithuania.

Rolland, Colette. 1998. “A Comprehensive View of Process Engineering.” in *Proceedings of the 10th international Conference on Advanced information Systems Engineering (CAiSE '98)*, edited by Barbara Pernici and Costantino Thanos, 1-24. London: Springer-Verlag.
Salinesi, Camille, and Jaana Wäyrynen. 2002. “A Methodological Framework for Understanding IS Adaptation through Enterprise Change.” In *Object Oriented Information Systems*, edited by Zohra Bellahsène, Dilip Patel and Colette Rolland, LNCS, 2425, 211-222. doi: 10.1007/3-540-46102-7_26
SEI (Software Engineering Institute). 2010. *CMMI for Development, Version 1.3*. Pittsburgh: Carnegie Mellon University.
Terras, Melissa. 2005. “Reading the Readers: Modelling Complex Humanities Processes to Build Cognitive Systems”, *Literary and Linguistic Computing*, 20, 1: 41 – 59. doi: 10.1093/llc/fqh042.
Thevenet, Laure-Hélène, and Camille Salinesi. 2007. “Aligning IS to organization's strategy: the INSTAL method.” In *Proceedings of the 19th international Conference on Advanced information Systems Engineering (CAiSE'07)*, edited by John Krogstie, Guttorm Sindre, and Andreas Opdahl, 203-217. Berlin: Springer-Verlag.