Software-Assisted Knowledge Generation in the Archaeological Domain: A Conceptual Framework

Patricia Martín-Rodilla[1]

[1] Institute of Heritage Sciences. Spanish National Research Council. Santiago de Compostela, Spain
patricia.martin-rodilla@incipit.csic.es

Abstract. Knowledge generation processes are traditionally related to the DIKW (data-information-knowledge-wisdom) hierarchy, a layered model for the classification of human understanding. Software components can be situated in one or several of these layers, or assist in the interfaces between two of them.
Most of the knowledge generation processes that occur in the archaeology field involve complex mechanisms of abstraction, relation and interpretation. Is it possible to assist the users in performing these processes?
We have detected problems in the archaeological knowledge generation process that could be improved through software assistance. We propose a conceptual framework based on the structure of the data that is being managed by the user, and on the cognitive processes that the user wishes to perform on the data.
The proposed framework can, arguably, set the foundation for assisted knowledge generation implemented as software systems.

Keywords: archaeology, knowledge generation, software assistance, conceptual modeling, inference.

1 Introduction

Archaeology is a vast domain that produces a large amount of data in different formats: 2D and 3D images, documents, audio, video, etc. and with specific work methodologies to record and manage data. Archaeology, like other disciplines in humanities and social sciences, presents complex needs with regard to temporality and subjectivity[1]. In this context, researchers and archaeology professionals demand technology support to achieve their tasks and new sub-fields emerge, such as Digital Humanities[2], with the necessity of software systems adapted to the domain and it characteristics.

The strongly human-centered nature of the domain – trying to re-build past events through existing data – focuses these needs in assisting to professionals in knowledge generation. This complex process is the central axis and causes the rest of tasks: conservation, preservation, museum activities, etc. However, the existing support for knowledge generation in this domain is limited to databases applications and some ad

hoc approaches to specific data types. There is not an integral solution to assist in knowledge generation processes.

The final goal of the current doctoral research is to provide a conceptual framework that set the foundation for assisted knowledge generation software in archaeological domain, help the users in their cognitive processes: direct observation of the data, complex visualizations, abstraction, etc. This paper summarizes the author's PhD work and project, working for one year and a half, under the supervision of Dr. Cesar Gonzalez-Perez (*Incipit, CSIC*) and Prof. Oscar Pastor Lopez (*Universitat Politècnica de València*). The first six months have been used for studying the domain in a deep way.

2 Problem Description and Research Considerations

In this section, the scope of the research is defined: a formal specification of the problems detected and research question formulated. Also, the research methodology followed is discussed.

2.1 Problem statement and initial hypothesis

Archaeological research builds knowledge based on data about past events. These data are presented in different formats and have been studied for years trying to find special characteristics and improving the recording and management efficiency.

Existing studies identifies specific issues in the archaeological data such temporality or subjectivity and proposes solutions to support them[3] .

However, building archaeological knowledge is a more complex process than supporting these data issues: it is necessary to study what the archaeologist want to do, what questions asked to the data and how archaeologist reach the initial goals.

Therefore, it is necessary to know the archaeologist's processes in the generation knowledge from existing data to valuable archaeological knowledge.

This research tries to discover if it is possible to assist the archaeologists in the knowledge generation processes through software. To carry out it and situate the problem statement, a previous study has been required to detect problems and weak points in the archaeological knowledge generation process.

During five months, archaeologists have been observed in a real work environment and have been interviewed about their work methodologies and software use. In addition, a set of questionnaires have been elaborated in order to study deeply the problems found during the observation period. All surveys have been completed by archaeologists (Incipit staff and non-Incipit staff). These professionals have several personal profiles in terms of age, gender and institutional affiliation (public/private institution, educational/non-educational institution).

The following problems in the knowledge generation process have been detected:

- Intentional use of uncertainty in the intermediate reasoning to generate knowledge. This uncertainty is not supported by existing software tools.

- Use of reasoning based on geographic and temporal data as a start point in the knowledge generation process. However, this initial reasoning is diffused along the rest of processes. This situation generates confusion about the context of the data in each moment of the knowledge generation process.
- Questions asked to the data have not been tracked and specified. This situation involves that non-asked questions are unknown and could be introduce some kind of bias in the knowledge finally generated.
- Previous point involves, in addition, that there are no chances to share the asked and non-asked questions between researchers or users working on the same data sources, complicating teamwork. There is no support for the collective generation of knowledge.
- Lack of priority management of the different questions asked to the data. (Different level of importance has been detected in the user processes).
- Lack of abstract view of the structure of the information management in the knowledge generation process. This problem could form the basis of a low use of the feedback mechanism to build and ask new questions to the data based on the responses obtained in a previous step.
- Homogeneous procedures applied to reasoning derived from direct observation and reasoning derived for more complex mechanisms (relation between data, abstraction, interpretation, etc.). This situation could be including confusion about the level of the DIKW hierarchy where the reasoning is situated and the level of subjectivity and uncertainty that is managed.

The problems detected make up the problematic context of the research and the gap founded that this research tries to fill. Our initial hypothesis is that *we can improve the knowledge generation process in the archaeological domain (minimizing or reducing the problems detected) by building software models that reproduce and incorporate cognitive facets and needs of the user, and allowing the application of visualization techniques and data-pattern recognition specifically adapted to archaeologists' ways of working*.

2.2 Existing solutions

Comparative and empirical studies have carried out to understand the characteristics of the user, the domain and existing software solutions[4]. However, knowledge generation processes in archaeology are not completely supported by software, with existing ad hoc partial solutions in terms of data[5]. Therefore, integral studies of this topic are not found in the existing literature.

However, there are theoretical models about human knowledge generation that defines the corpus of this research. All existing models follow a hierarchical structure based on layers, with other differences:

Cleveland[6] established a model in four layers: Fact & ideas, Information, Knowledge and Wisdom. Cleveland model laid the foundation to a human understanding theory. The intermediate processes between layers are not characterized at this point.

Ackoff[7] added one step more, with five layers: Data, Information, Knowledge, understanding and Wisdom. This model has been used for several years as a reference in psychology and cognitive studies.

Carpenter & Cannady[8], based on other characterizations of the intermediate process between layers[9] incorporated in 2004 feedback flow between layers. They proposed a model with six layers: Environment, Data, Information, Knowledge, Wisdom and Vision. The intermediate steps are tagged with words that suggest cognitive processes between layers, such as rules, goal or values.

All compared models fit in with the cognitive character of the processes that allow us to go up to the next level in the hierarchy. Thus, the characterization of these cognitive processes into the archaeological domain could be an initial point to solve the problems detected. Gardin[10], Stockinger[11] or recently Doerr[12] notes that it is possible a formalization of these processes. However, the level of formalization of these studies is not enough to test the hypothesis initial of this research and include it directly in the specification of an assistance based on software solutions. It is necessary a complete formalization to achieve this goal.

Making a complete review of the literature that supports the current research, we have reviewed the existing methods to assist through software in knowledge generation processes in other disciplines or domains. The main example founds is the biomedical domain, in which context Chen[13] developed a complete model to assist by visualization software the knowledge generation process.

Chen et al.[13] proposed a model based on modules incorporated to the software system that captures the user actions, establishing how the user is generating knowledge. The module incorporates this knowledge to the system: In the next step, the system can adapt it behavior to the user and offer him some helped tools.

Chen references DIKW[7] as a scope of his model, without restrictions in terms of domain. We takes into consideration this model as a basis in the archaeological domain and adapted visualization and pattern suggestions proposed by Chen et al. as possible output of our assisted knowledge generation system.

2.3 Research questions and objectives

Regarding TAR[14] -Technical Action Research - methodologies for software and technical research, there are two categories in terms of research questions:

Category 1: Questions that seek an explanation for a real-world phenomenon, and try to answer what, why or how said phenomenon occurs.

Category 2: Questions that try to build some artifacts from scratch or to improve some existing solution for a problem previously detected.

The main research question of this work emerges from the initial hypothesis. It formulation is: *To what extent is it possible to improve the knowledge generation process in archaeology by assisting the user through software* tools? In order to ask this research question, the research general goal is to test the initial hypothesis through:

- Searching for evidence those support the hypothesis, through user testing and empirical studies.

- Searching for evidence against the initial hypothesis, through empirical studies, user observation, validation test and expert feedback.

In the process of answering these questions, additional collateral questions are emerging that allow us to establish a research context. Currently, the research is focused on the secondary question: What are the major existing problems in the knowledge generation process in archaeology? In order to answer the secondary research question, we have defined to detect these existing problems as a specific goal in the first part of the research.

2.4 Research methodology

Our research follows an established scientific methodology approach, based on an initial hypothesis and a general research question to answer. Secondary questions are emerging along the research process. Each secondary research question involves a set of valuable objectives whose commitment level reflects the level of support of the initial hypothesis. In our case, the research questions and the objectives have been exposed in the previous section.

In addition to the methodology general plan following research methods of validation a specific methodology for the requirements elicitation and the study of knowledge generation methods in this domain is proposed.

The interdisciplinary character of the research and user characteristics has been necessary to apply user-centered methods, in order to understand how users build knowledge from the data and how kind of software assistance would be better to support these processes.

On the other hand, the fact that software systems are involved suggests that we need to formalize user needs into requirements.

In this scenario, a hybrid methodology is proposed. All requirement elicitation processes are extracted following shadowing studies, interviews, surveys and user experiments. This allows us to directly observe the users in a real work environment, together with archaeological sets of data and case studies in their domain. In order to contrast the empirical results in an analytical way, other analytical techniques to extract requirements are applied concurrently, such as discourse analysis and data-mining.

3 Proposal

Our proposal involves a conceptual framework that captures and implements the needs of the users who work in the archaeological domain. The framework is structured in three important parts in the knowledge generation process: data, process and interaction.

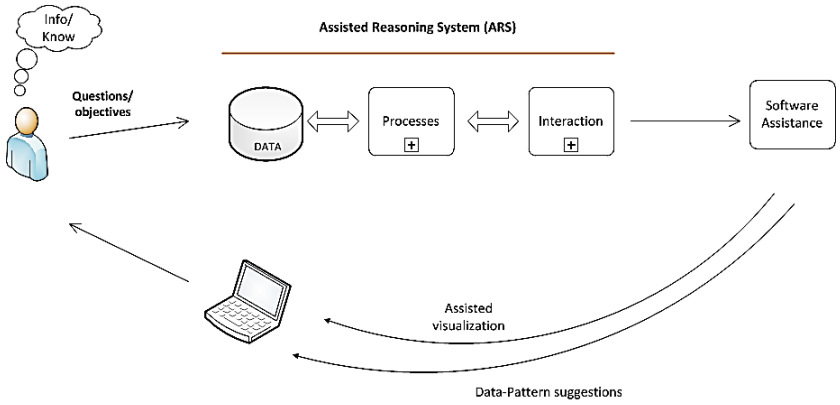

Fig. 1 Conceptual framework proposed and assistance flow.

The characterization and formalization of the three parts of the framework allow the system to offer adapted visualizations following Chen's principles and the designed workflow.

In addition, the system is intended to offer in particular cases pattern recognition and suggestions about patterns previously found in data with a similar structure or content (See Fig.1).

The framework proposed is expected to sustainability reduce the problems detected (see Problem Statement section), adapting to the user in an iterative way and incorporating cognitive information from it.

Testing the initial hypothesis through the implementation of a prototype of the framework is the most challenging part of the research, because there are no existing solutions that assist the user in an integral fashion in the knowledge generation process.

4 Preliminary results

In addition to the problems characterization achieved (see Problem Statement section), the research has been focused on the searching for evidence to support or refute the initial hypothesis, carrying out analytical experiments and empirical studies.

4.1 Analytical results

We have analyzed archaeological document collections from Incipit and other open-data repositories[15] studying the structure of the argumentation narrative, problems in the depicted knowledge generation processes, and evidence of possible software assistance solutions in the field. The analysis was carried by following the discourse analysis method of Hobbs[16], an analytical technique designed to study texts from a linguistic perspective that allows us to characterize argumentation relations between clauses. This method includes ten types identified as coherence relations that

include causal and contrastive argumentations, exemplifications and generalization of arguments, etc.

Following this method, each pair of clauses is tagged by one coherence relation. After, the elements of the relation chosen are characterized in the clause analyzed. Finally, it is necessary to validate the elections chosen in each analysis. In our study the original author of each document form Incipit repositories was contacted and asked to validate the outcome of the analysis

The analysis results indicate that there are types of coherence relations more common in the analyzed texts than others, namely relations based on combining values of several attributes to build the argumentation. This type of inferences is related to combinatory tasks, and some data-mining solutions such as rule-association algorithms[17] could arguably provide software assistance in the knowledge generation process in archaeology.

This analysis allow us to characterize in an analytical classification the argumentation methods detected in the study, based on the objective of the user and the type of inference that was used.

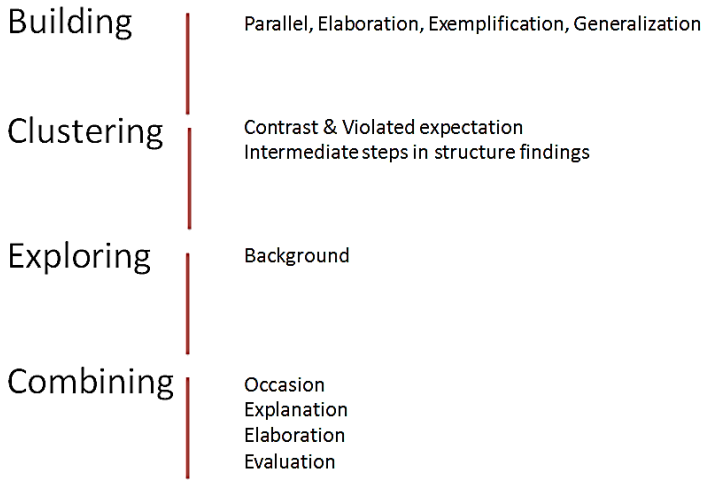

Fig. 2 Archaeological inference classification.

The classification obtained as a result of the analysis is explained in the Figure 2, with the corresponding mappings with Hobbs relations. It is used as basis to formalize the cognitive processes managed by the user in archaeology.

4.2 Empirical results

Two test-sets has been designed and implemented to extract evidence of the assisting possibilities through software in the archaeological domain.

The first test contains multiple choice questions about reasoning modes in archaeology based on real data, tasks related to knowledge generation and initial reasoning

used to process data in large datasets (See Fig.3). There are plans to publish all results and conclusions from this test-set.

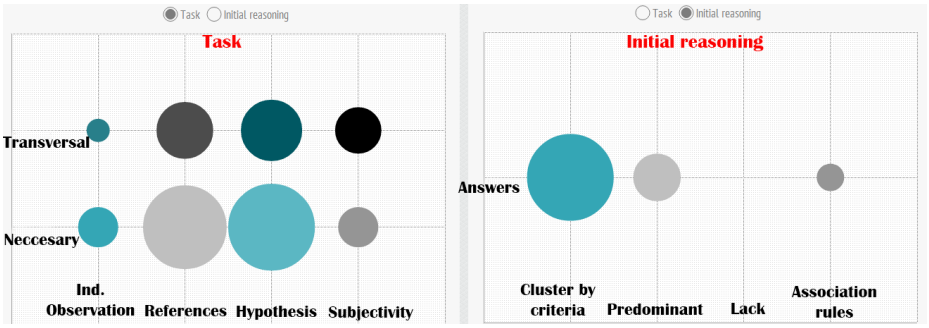

Fig. 3 Some empirical results from test-set 1

The second test-bed includes multiple-choice questions about ten visualizations of real archaeological data sets implemented through well-known visualizations techniques (bars, charts, tree-maps[18], graphs, bubbles…) and different levels of interaction. The users are asked to carry out common task against the visualized data and must answer the questions based on direct observation, attribute value combination, abstraction or interpretation of the data. This second test-bed is being deployed at the time of writing with promising preliminary results.

5 Future plan

The proposed framework will be developed as an integrated metamodel covering the three areas defined above: data, process, and interaction.

Metamodeling the archaeological cognitive processes explained in this paper and extracting applicable interaction primitives are our next goals. In the first case, situational software engineering[19] approaches are chosen to express the behaviors, elections and objectives of the user identified in the Preliminary Results section as an inference classification.

In the second case, the empirical studies being carried out indicate the necessity of different levels of interaction and visualizations, depending on the inference type involved. We continue to work in this direction.

In addition to the integrated metamodel, the implementation of a software prototype is planned, with the objective of testing the solution and systematically analyzing the improvements achieved in the knowledge generation process by assessing the detected problems, their impact and the obstacles encountered**.**

References

1. Gonzalez-Pérez, C. and Parcero-Oubiña, C. A Conceptual Model for Cultural Heritage Definition and Motivation. In CAA'11: 39th Annual Conference on Computer Applications and Quantitative Methods in Archaeology 2011. Beijing, China.
2. Svensson, P. Humanities Computing as Digital Humanities, 2009, Digital Humanities Quarterly.
3. Doerr, M. The cidoc crm, an ontological approach to schema heterogeneity. Semantic interoperability and integration, 2005. 4391.
4. Martín-Rodilla, P. The role of software in cultural heritage issues: Types, user needs and design guidelines based on principles of interaction in RCIS'12: Sixth International Conference on Research Challenges in Information Science. 2012. Valencia, Spain: IEEE XPlore.
5. Gonzalez-Perez, C. et al. Extending an Abstract Reference Model for Transdisciplinary Work in Cultural Heritage, in Metadata and Semantics Research, J. Dodero, M. Palomo-Duarte, and P. Karampiperis, Editors. 2012, Springer Berlin Heidelberg. p. 190-201.
6. Cleveland, H. Information as Resource. The Futurist, 1982: p. 34-39.
7. Ackoff, R.L. From data to wisdom. Journal of Applied Systems Analysis. 16: p. 3-9.
8. Carpenter, S.A. and Cannady, J. Tool for Sharing and Assessing Models of Fusion-Based Space Transportation Systems. in Proceedings of the 40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit. 2004. Fort Lauderdale, Florida.
9. Bellinger, G. Knowledge Management—Emerging Perspectives. 1997; Available from: http://www.systems-thinking.org/kmgmt/kmgmt.htm.
10. Gardin, J.C. Archaeological Discourse, Conceptual Modelling and Digitalisation: an lnterim Report of the Logicist Program. in Computer Applications and Quantitative Methods in Archaeology. 2002. Heraklion, Crete, Greece.: Hellenic Ministry of Culture, 2003.
11. Stockinger, P. On Gardin's logicist analysis, in Interpretation in the Humanities: Perspectives from Artificial Intelligence, R.E.e.J.-C.G. (eds), Editor 1990, British Library Pub.: London. p. 284 - 304.
12. Doerr, M., Kritsotaki, A. and Boutsika, K. Factual argumentation -a core model for assertions making. J. Comput. Cult. Herit., 2011. 3(3): p. 1-34.
13. Chen, M.E. et al., Data, Information, and Knowledge in Visualization. IEEE Computer Graphics and Applications, 2009. 29(1): p. 12-19.
14. Wieringa, R. and Morali, A. Technical Action Research as a Validation Method in Information Systems Design Science, in Design Science Research in Information Systems. Advances in Theory and Practice, K. Peffers, M. Rothenberger, and B. Kuechler, Editors. 2012, Springer Berlin Heidelberg. p. 220-238.
15. Peltenburg, E., Excavations at Kissonerga-Mosphilia 1979-1992 [data-set], 2000, York: Archaeology Data Service.
16. Hobbs, J.R. On the Coherence and Structure of Discourse, in Technical Report1985, Center for the Study of Language and Information (CSLI): Standford, CA.
17. Agrawal, R., Imieli, T. and Swami, A. Mining association rules between sets of items in large databases. SIGMOD Rec., 1993. 22(2): p. 207-216.
18. Johnson, B. and Shneiderman, B. Tree-Maps: a space-filling approach to the visualization of hierarchical information structures, in Proceedings of the 2nd conference on Visualization '911991, IEEE Computer Society Press: San Diego, California. p. 284-291.
19. Gonzalez-Perez, C. and Hug, C. Crafting Archaeological Methodologies: Suggesting Method Engineering for the Humanities and Social Sciences. in Computer Applications and Quantitative Methods in Archaeology (CAA) 2012. Southampton, UK.