A Semantic Based Approach to GIS: The PO-BASyN Project

Glauco MANTEGARI – Alessandro MOSCA – Bernardo RONDELLI

Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Milan (Italy)
{glauco.mantegari; alessandro.mosca; bernardo.rondelli}@disco.unimib.it

Abstract

The paper introduces an overview on the researches we are currently carrying out on the integration of semantic aspects in Geographical Information Systems. Through an overview on the activities in the context of a project concerning the study of Bronze Age settlements in northern Italy, the paper shows how issues concerning semantics and GIS are currently faced to define an innovative and integrated system.

Keywords

GIS, Knowledge Representation, Nonmonotonic Reasoning, Artifact Classification

1. Introduction

Geographic Information Systems (GIS) offer a consistent set of search, retrieval and analysis tools dealing with spatial information, which are mainly based on relational database query techniques and quantitative math-based elaborations.

Recently a number of authors have noted that the conventional vector and raster data models for GIS are inadequate for the representation of many complex geographic phenomena, thus giving origin to a semantic gap in GIS database representation.

A number of authors have considered object-oriented data models to address this semantic gap and other related issues of information representation in GIS databases. Many of these efforts, however, have not focused explicitly on conceptual representation, focusing instead on the representation of complex geometry and geometric relationships as extensions to the vector data model. Another approach to the issue of GIS database representation draws from the field of ontologies (Fonseca *et al.* 2002). Research in ontology, in the framework of GIS, addresses how GIS users' conceptual models of geographic domains may be elicited, formalized, and represented within a GIS context. However, much of this work in ontology-driven GIS has focused on developing ontologies of geographic domains for purposes of data sharing and standardization and it has not been extended within the context of advanced GIS functions such as knowledge discovery (Mennis 2003).

Recent advances in the field of knowledge representation and the growth of web technologies and applications represent a fertile ground for enhancing these capabilities, offering also new perspectives for archaeology.

By means of the exploitation of Data Integration and Semantic Web technologies, it is possible, for example, to support several innovative GIS functionalities, which are related to the semantic access, navigation and querying of heterogeneous multi-layered data and information.

In this framework, it is possible to envisage some new analytical and explanatory capabilities, which are more connected with knowledge representation techniques applied to describe in a deeper way the entities that are situated in the GIS environment.

This paper focuses in particular on this aspect, proposing an innovative approach to artifact representation and classification, as a starting point, to support the material culture analysis and the settlement dynamics study in the context of the Bronze Age in Northern Italy.

The paper is organized as follows: section 2 briefly introduces the scenario of a semantic support in GIS; section 3 describes the main characteristics of the project under development; section 4 focuses on Knowledge Representation techniques applied to the study of material culture; section 5 presents the conclusions about the work carried out until now.

2. Towards a semantic support for Geographical Information Systems

Geographical Information Systems have evolved a lot since their introduction. The geodata management capabilities of these systems, as well as their analytical power, make GIS a set of applications of crucial importance in different contexts. Recent advances in the field of Knowledge Representation and the impact of the related models and techniques are deeply influencing the possibilities of Information Systems, by moving the interest towards the semantic aspects of information. GIS are just at the beginning of this process. In particular, there is an increasing interest in the representation of geographical and thematic (i.e. attribute-related) information by means of semantic-based models. In this way the traditional and well-consolidated entity-relational structures may be enriched, opening for new possibilities in data integration, retrieval, sharing and analysis. In fact, GIS have proved their efficiency in representing and analyzing simple and discrete entities (both from a spatial and temporal perspective), while, in most cases, the reality is much more complex, (i.e. spatially non-discrete and temporally dynamic) and it is impossible to represent with the traditional structures.

A large number of research has concentrated their effort on the study of these aspects, from the investigation of the cognitive concepts of space to the proposals of new models and technologies, such as the temporal GIS (e.g. Christakos *et al.* 2002).

More specifically, in the last ten years there was a clear tendency to represent spatial knowledge into ontological frameworks (e.g. Casati *et al.* 1998). Researches on ontologies and the development of standard languages (e.g. the Web Ontology Language - OWL[1]) offered the possibility to go further in these directions and to experiment a semantic support for GIS. The results are still partial and mainly limited to Web-based systems and services, which increased quickly both in number and complexity originating the new scenario of the "Geospatial Web" (see Sharl and Tochtermann 2007).

It is thus interesting to explore these new models and technologies in the context of a discipline, like archaeology, where the representation of knowledge is particularly complex and the geographical and temporal components play a crucial role.

3. A case study: the PO-BASyN project

The proposed case study is the PO-BASyN (Po Valley and Bronze Age Settlement Dynamics) project. The main objective of the project is to investigate the Bronze Age settlement dynamics in the Po Valley, in northern Italy.

This case study offers a paradigmatic case for geographical and environmental uniformity, data richness, with a large number of excavations conducted in the area during the last decades, old investigations checked and revised, and with a large number of research institutions working on these subjects each year.

The Project involves the Chair of Prehistory and Protohistory at the University of Milan, the Department of Archaeology at the University of Bologna, and the Complex Systems and Artificial Intelligence Research Center at the University of Milan Bicocca. Our contribution to the Project is related to the methodological framework and technical aspects comprising two different inter-linked sub-projects:

a) the support to the integration and semantic retrieval of different types of archaeological information with effective semantic web-based interfaces;
b) the improvement of the classification and interpretation of archeological finds.

With regard to the the first sub-project, it is important to stress that the existence of a large database with information on sites and finds requires the development of a system for their classification and harmonization, which may be helpful both for data retrieval and analytical purposes. The case study offers a rich and qualitatively relevant corpus of heterogeneous data, which were obtained through excavations, surveys and by describing and archiving museum collections. These data are both structured and semi-structured and they are characterized by heterogeneous semantics.

Considering this, the first sub-project aims at providing innovative tools for the sharing and retrieval of information and knowledge to everyone interested in northern Italian prehistory and protohistory.

The system prototype, which is at an advanced stage of development and is fully operative, currently consists of a web portal, named "ArcheoServer", which has an important section dedicated to a Web-based GIS.

The WebGIS system has been developed since 2005 (Mantegari *et al.* 2007). Recent improvements were done in the direction of making the system

[1] http://www.w3.org/TR/owl-ref.

a geospatial web application (Mantegari and De Salvo, this volume). In this scenario and considering the above-mentioned directions in GIS research, efforts are directed at the definition of a semantic support for the system, using the technolgies of the Semantic Web. In particular, we are currently developing a domain ontology, which will be used for the representation and retrieval of both geographical and thematic information. This sub-project is in an early stage of development, even if some tools for the semantic annotation and retrieval have already been developed. For this reason it will not be discussed further in this article.

4. Knowledge representation and material culture analysis

The aspects related to the semantic retrieval of information and to the sharing of information, data and explicit knowledge are of crucial relevance. However, we consider that it is necessary to go further and in particular to take into account the fact the "entities" georeferenced in the GIS have a value which depends on the thing they represent. Each "entity" is defined and interpreted in function of its composing parts and in function of both quantitative and ontological considerations. Thus semantics are of major importance for interpretation and it is not possible to reduce everything to purely statistical and quantitative aspects.

The second subproject is mainly connected with the manner in which we describe, analyze, or classify the phenomena we are studying. Settlement analysis in archaeology (and in particular in prehistory) seeks to build up from the spatial distribution of material culture and anthropogenic modifications visible in the contemporary landscape to an understanding of the dynamic cultural and environmental processes of human settlement systems. This said, in the context of the Po Valley, different specialists have analyzed settlement dynamics through distributive analysis of different types so as to determine clusters and describe population distribution and dynamic as a sequence of snapshots of "clustering of dwellings" which change form and dimension in time. By doing this, each archaeologist implicitly, or more or less explicitly, uses a qualitative reasoning: this dwelling is connected to this other dwelling because they have material cultures that qualify as analogies on the basis of a defined typology. Moreover, we assume that material culture is a knowledge symptom and reification of a knowledge-complex[2]. Hence, knowledge represents the primary analytical unit of our inquiry.

For that reason, our approach has been centered, in a first phase, on the analysis of material culture in terms of knowledge representation and, in particular, on an attempt to support the classification of artefacts.

The artefacts are classified by correlating the heterogeneous information in order to infer initially unknown characteristics (for example, chronology or a functionality starting from the shape and the observation of morphology). Our approach considers it fundamental to increase the quality and quantity of information that an archaeologist uses in order to define parameters of the artefacts to be classified. An enlargement of the number of possible correlations between the information relative to the artefact implies a better articulation of the axiomatic base on which the classification is based and therefore a greater analytical power of the system itself.

This said, we consider the analysis of the archaeological material culture as a starting point for a study of the "subject who *made*, *destroy*, and *use* it". In this sense the material culture can be viewed as an "observable", i.e. a property of the system we are studying that can be observed directly.

We assume, in fact, that the relationships between subject and object (from now on, 'artefact') gives rise to a "2-cycle" (Knappett 2005 - "*Material culture invariably entails a codependency of mind, action and matter*").

The idea is to define the relationships between subjects and artefacts from a twofold perspective: on the one hand, the subject is the artefact-maker (and, in this sense, it provides the artefact with a mereological structure, a morphology, a set of explicit constraints

[2] In the course of the history of human thinking a number of different philosophical perspectives on what the human knowledge is have emerged and, as a consequence of this, a number of different attempts have been carried out to provide a comprehensive definition of it. In the "Plato on Knowledge in the Theaetetus" lemma of the Stanford Encyclopedia of Philosophy (by the Prof. Timothy Chappell), the following abstract characterization is reported: "Nothing is more natural for modern philosophers than to contrast knowledge of objects (knowledge by acquaintance or objectual knowledge; French connaître) with knowledge of how to do things (technique knowledge), and with knowledge of propositions or facts (propositional knowledge; French savoir)".

on its dimensionality, a number of expected functions, and a suitable matter-energy substrate). On the other hand, the artefact, resulting from a productive human activity, brings a number of "organizational rules" (or "social constraints") that have a direct or indirect impact on the social organization in which subjects are living, e.g. in terms of behavioural rules such as the emergence of new communication and transaction acts (Dopfer and Potts 2008).

4.1. A formal knowledge representation problem

The need for representing and manipulating complex knowledge is a topic with a long history in the Artificial Intelligence (AI) research area. In particular, the problem of how to produce formal and explicit representations of knowledge, and how these kinds of representations can be manipulated by reasoning engines, have become the main research object of one of most active and lively sub-field of AI, called Knowledge Representation and Reasoning (KR&R). On the one hand, KR&R can be defined as the discipline regarding how knowledge can be represented symbolically. This general definition does not specify, of course, what knowledge is and what the symbolic representations we may use are; however, for the sake of this paper it is sufficient in order to clarify the background of our research. In (Brachman and Levesque 2004) reasoning is defined as "[...] the formal manipulation of the symbols representing a collection of believed propositions to produce representations of new ones". Therefore, knowledge representation and reasoning are two sides of the same coin and have to be considered as strictly intertwined because "[...] every representation embeds at its core a conception of what constitutes intelligent reasoning.

The knowledge representation language we use is called AnsProlog (or 'A-Prolog', a short form for 'Programming in logic with Answer sets') a language in the framework of declarative programming based on stable model semantics (also known as 'answer set semantics'). Unlike Prolog, ASP has well-defined declarative semantics for non-monotonic features (e.g. "negation as failure"), that is independent of a particular inference mechanism. This means that AnsProlog can be considered as a specification and a program; therefore, AnsProlog representations eliminate the ubiquitous gap between specification and programming. Finally, AnsProlog programs are logic programs where disjunction is allowed in the heads of the rules and negation may occur in the bodies of the rules; such programs have been widely recognized as a valuable tool for knowledge representation and commonsense reasoning. Note that one of the attractions of AnsProlog is its capability of allowing the modeling of incomplete knowledge.

In brief, a program of ASP is a pair $\{\sigma,\Pi\}$ where σ is a signature and Π is a collection of rules. Without entering into the details, an AnsProlog program consists in a collection of rules of the form:

$$L_o \textit{ or } \dots \textit{ or } L_k \text{ :- } L_{k+1}, \dots, L_m, \textbf{not } L_{m+1}, \dots, \textbf{not } L_n.$$

Where, the L_is are literals in the sense of classical logic. The rule can be read as: if $L_{k+1}, \dots, L_m$ are to be true, and if **not** $L_{m+1}, \dots,$ **not** L_n can be safely assumed to be false, then at least one of L_o *or ... or* L_k must be true.

Different sub-classes of AnsProlog have been defined according to the different specifications of their signature and, as a consequence, according to their different expressive power. The rules we will introduce in what follows have been written in the function-free disjunctive logic language (also known as DLV).

AnsProlog has efficient implementations[3] that have been used to program large applications in the fields of database query languages, knowledge representation, reasoning, and planning. We refer to Baral (2003) for a comprehensive overview of the language AnsProlog, its formal semantics, theoretical foundations, and implementations.

Ceramic artifact classification is a problem solving activity exploiting domain dependent heuristics. From an epistemological perspective a classification heuristic can be viewed as the result of correlating and integrating heterogeneous pieces of the knowledge one has about the objects of a given domain. We assume that a classification heuristic, at least in a scientific research context, comes from a relevant correlation of the attributes and features we used to describe the object.

If one accepts these premises, the first step towards the definition of a comprehensive set of classification heuristics in a given domain consists

3 See, for example, the DLV Project http://www.dbai.tuwien.ac.at/proj/dlv/

in providing a set of sufficient (and, possibly, necessary) ontological conditions establishing what a given entity is for a given scholar or research community (i.e. what are the relevant attribute categories that are used to describe the entity, and what are the contingent attributes an entity must have in order to be recognized as a member of a given ontological realm and what are the necessary ones). The definition of the ontological conditions concerning the entities of a given domain is what we call a meta-model of these entities.

Given that, the second step consists in defining classification heuristics as rules whose preconditions are sets of entity correlated attributes and whose post conditions are class and type membership sentences (from now on, "class-classification" and "type-classification rules").

In what follows, we provide an introduction of the meta-model of the artifact we have defined and implemented, and we give an example of the resulting classification heuristics. In the formal syntax of DLV, the ceramic artifact meta-model can be implemented as:

```
artifact(X)   :-   hasStructure(X),
                   hasDimensionality(X),
                   hasMorphology(X),
                   hasMatter(X),
                   hasSpatialLocation(X),
                   not -hasFunctionality(X),
                   not hasTemporalLocation(X).
```

The rule says that the dimensions along which a generic entity can be recognized as a ceramic artifact are the following: (i) Structural (i.e. the entity has a discrete mereological composition); (ii) Material (i.e. the entity is made of specific chemical substances and physical intermolecular forces); (iii) Dimensional (i.e. the entity has a dimensionality qua 'res extensa'); (iv) Morphological (i.e. the entity has a specific shape and profile); (v) Temporal and Spatial (i.e. the entity has temporal and spatial localizations); (vi) Functional (i.e. the entity is characterized by having functions as 'technical' – final use – and organizational – social impact). Note that thanks to the expressive power of DLV, we can represent that a given entity, in order to be recognized as a ceramic artifact (and to be classifiable by the system), does not necessarily have an attached specific functionality and a temporal specification. The idea behind this representational choice is that we plan to enlarge the logic program we are developing with rules explicitly devoted to the identification of these kinds of information.

Moreover, a ceramic artifact can be interpreted as a compound object made of different atomic and aggregate components. The hierarchical structure establishing the relationships among these components has been constrained by introducing a 'direct part of' relation (reflexive and intransitive), holding among entities that live in contiguous levels of the hierarchy, and a 'part of' relation as the transitive closure of the first one. Components are characterized by having specific shapes, dimensions (e.g. height, weight, diameter), profiles (e.g. convex, concave), and the presence/absence of decorative elements. On the basis of this information, we implemented a set of rules that provide similar characterizations for the whole artifact.

For what concerns the class-classification heuristics that are devoted to the automatic deduction of class memberships, consider the following rule as an example:

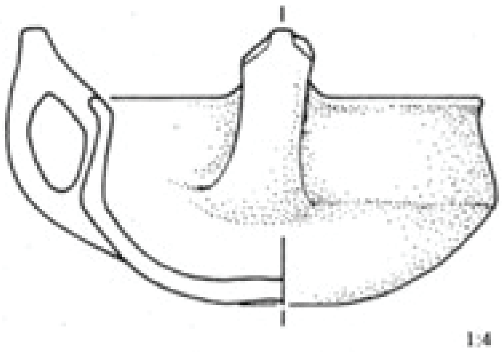

Fig. 1. A common example of a Bronze Age ceramic artifact, that is identified by our classification heuristics as a member of the "cup" class.

```
class(X,cup)  :-   artifact(X),
                   openShape(X),
                   hasProfile(X,articulated),
                   handle(A), partOf(A,X),
                   hasMaxDiameter(X,D), #let(D,"35").
```

The rule correlates different pieces of knowledge in order to characterize the set of ceramic artifacts called "cups". In particular, the rule establishes that a cup is a ceramic artifact (i.e. an entity for which we can provide information according to the meta-model introduced above), with an articulated profile, a handle component, and with a maximal diameter value lower or equal to 35 centimeters. Whenever the system encounters a description of a ceramic entity, coming from the geospatial database, that satisfies all the rule preconditions, the system is able to automatically infer the class of this entity.

Consider that this kind of class-classification rules can be further specialized in order to introduce type-classification rules. The following rules is about the “Tabina Cup” type:

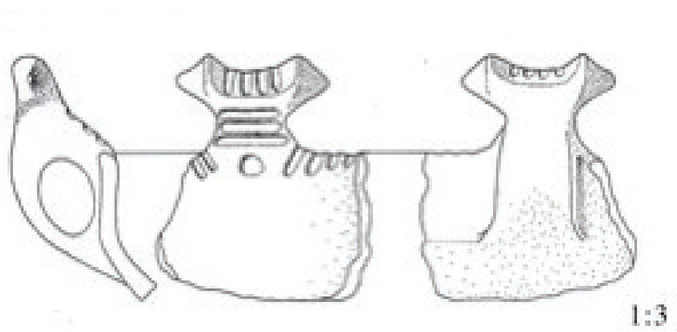

Fig. 2. A drawing of a cup with the distinctive vertical rising horned handles. These elements are recurrent in the area of the so called Terramare culture, and are the symptoms of a Middle Bronze Age production. This type of cups is called “Tabina”, according to the archaeological site of Tabina di Magreta (Modena), where similar objects were discovered for the first time.

```
type(X,tabinaCup)   :-   class(X,cup),
                         carina(Y), partOf(Y,X),
                         hasHProfile(X,verticalRisingHorned).
```

The rule establishes that a cup, characterized by having a carina component, and a handle with a typical profile called “Vertical Rising Horned Handle”, can be recognized as an artifact of the type “Tabina[4]”).

4.2. From DLV to GIS

Recent developments of the DLV Project have provided very useful API (Application Programming Interface) supporting the design and the implementation of JAVA based applications that can incorporate the DLV functionalities. The idea behind the integration of our logic program and the semantic-improved GIS system (see Part-I) is quite simple. The data that are stored in the geodatabase, together with their associated explicit definition, can be extracted by means of the suitable SQL queries. Then, a software module (called ‘wrapper’) written for example in JAVA, takes care of translating the syntax of the retrieved data into that of DLV according to the predicates that have been defined and axiomatized in the program. Therefore, the wrapper provides as output a respective list of new facts that can be included in the logic program, and used to extract new relevant inferences from the knowledge base in terms of classificatory sentences.

Geospatial data, whose representation have been enriched by means of standard Semantic Web techniques and languages, become inputs, in the form of facts, of the logic program (e.g. an acceptable input could be the set of information concerning an artifact, or a specific dwelling structure of an archaeological site). The logic program performs its inferential tasks and produces as output a set of classificatory sentences concerning the input data (e.g. information that have been just implicitly present in the input data, type classification of the artifact of interest) or a warning concerning the logical inconsistency of the data itself. Finally, the inferred information translate, according to the syntax specification of the WebGIS system, into data that enhance the stored geospatial dataset and that can be used in the future for further inquiries.

5. Concluding remarks

The paper briefly discussed the semantic enrichment issue that has affected research into GIS systems in recent years. The explicit and formal representation of geospatial data in GIS systems is presently perceived as a key research challenge in computer applications in archaeology. The basic idea is that the application of the existing Semantic Web formal standards and techniques can improve GIS systems by allowing the design and implementation of new and efficient querying/retrieval, integration, and navigation/ access functionalities. These functionalities aim at supporting scholars in having large repositories of heterogeneous information as the basis of their analytical activities, and to perform significant inquiries on these. As an example of the exploitation of the semantic enriched WebGIS platform, the paper introduced some of the most relevant characteristics of the automatic pottery classification system. The introduced logic based classification system, that is able to exploit a semantic enriched geospatial dataset for analytical purposes, has been integrated

4 Tabina is a place in the North of Italy, near Modena; it identifies an archaeological site that gives a name to a specific chronological phase of the Middle Bronze Age.

into the WebGIS functionalities of the web portal Archeoserver.

Bibliography

Baral, Chitta (2003). *Knowledge Representation, Reasoning and Declarative Problem Solving*, Cambridge University Press.

Brachman, Ronald J. and Hector J. Levesque (2004). *Knowledge Representation and Reasoning*. Elsevier, San Francisco.

Casati, Roberto, Berry Smith and Achille Varzi (1998). Ontological Tools for Geographic Representation. In: Nicola Guarino (ed.) *Formal Ontology in Information Systems*. Amsterdam: IOS Press, 77–85.

Davis, Randall, Howard Shrobe and Peter Szolovits (1993). What is a Knowledge Representation? In: Brachman R. J. and Levesque H. J. (ed.) *AI Magazine* 14 (1), 17–33.

Dopfer, Kurt and Jason Potts (2008). *The General Theory of Economic Evolution*. Routledge.

Fonseca, Frederico, Max Egenhofer, Clodoveu Davis and Gilberto Câmara (2002). Semantic Granularity in Ontology-Driven Geographic Information Systems. *AMAI Annals of Mathematics and Artificial Intelligence - Special Issue on Spatial and Temporal Granularity* 36, 121–151.

Knappett, Carl (2005). *Thinking Through Material Culture: An Interdisciplinary Perspective*. Philadelphia: University of Pennsylvania Press.

Knappett, Carl (2006). Beyond Skin: Layering and Networking in Art and Archaeology. *Cambridge Archaeological Journal* 16(2), 239–251.

Leone, Nicola, Gerald Pfeifer, Wolfgang Faber, Thomas Eiter, Georg Gottlob, Simona Perri and Francesco Scarcello (2006). The DLV System for Knowledge Representation and Reasoning. ACM Transactions on Computational Logic, 499–562.

Mantegari, Glauco, Maurizio Cattani, Raffaele de Marinis and Giuseppe Vizzari (2007). Towards a Web-based Environment for Italian Prehistory and Protohistory. In: Jeffrey Clarke and Emily Hagemeister (eds.) *Digital Discovery. Exploring New Frontiers in Human Heritage. CAA2006. Computer Applications and Quantitative Methods in Archaeology Proceedings*. Budapest: Archaeolingua.

Mennis, Jeremy L. (2003). Derivation and implementation of a semantic GIS data model informed by principles of cognition. *Computers, Environment and Urban Systems* 27, 455–479.

Gerlek, Michael and Matt Fleagle (2007). Imaging on the Geospatial Web Using JPEG 2000. In: Arno Sharl and Klaus Tochtermann (eds.) *The Geospatial Web*. London: Springer, 27–38.