Investigating a multi-paradigm system for the management of archaeological data: Corpus Lapidum Burgundiae

Éric Leclercq*, Marinette Savonnet*, Andrés Troya-Galvis*, Stéphane Büttner†
* Université de Bourgogne, Le2I Laboratory, UMR CNRS 6306
9, Av. Alain Savary 21078, Dijon, France
Eric.Leclercq@u-bourgogne.fr, Marinette.Savonnet@u-bourgogne.fr, ac.troya@gmail.com
† Centre d'études médiévales - ARTeHIS Laboratory, UMR CNRS 6298
3 pl. du Coche d'eau 89000 Auxerre, France
stephane.buttner@cem-auxerre.fr

***Abstract*—Scientific Information Systems (SIS) must move beyond data repositories and closed systems, to allow collaborations among different research disciplines, to include new types of data, to control data quality, and to enable semantic interoperability. Archaeological data include textual information, measures, sketches, photographies, 3D models, and a vast amount of links between data and historical information sources. We develop a formal model for ontology-based annotations that conforms to a semi-ring algebraic structure and we define a subset of algebraic operators to query annotations. We show how our approach is instantiated in a collaborative Web platform for the Burgundy Stone project.**

Keywords*—*Scientific Information System, Archaeological Corpus, Ontology-based Annotation, Semantic Wiki.

I. Introduction

SIS must move beyond data repositories and closed systems, to allow collaborations among different research disciplines, to include new types of data, to control the quality of derived data, and to enable semantic interoperability. SIS aim to produce, improve and manage knowledge on a subject through research activities. Unlike enterprise information systems, SIS do not support nor control services or production activities. Moreover, SIS should include collaborative features to allow cooperation among different kinds of users (scientists of different disciplines, domain professionals, etc.).

Scientific data have the following properties: 1) they include various collections of large datasets; 2) they use complex spatio-temporal models; and 3) they enclose both explicit and implicit, hard-to-discover relationships. Moreover, scientific data are heterogeneous as they come from different sources (*e.g.*, observation and reanalysis data in climatology) as well as from different acquisition technologies (*e.g.*, mass spectroscopy in biology, 3D scanner in cultural heritage). A large variability of data models is observed, resulting from the evolution of scientific knowledge (*i.e.*, migration from purely experimental methods to a statistical way of thinking [1]) but also from high performance computing that allows computation at the molecular level as well as at astronomical scales.

Database Management Systems (DBMS) are essential components of SIS, nevertheless Relational-DBMS do not provide enough extensibility. Schema evolutions usually deeply impact applications and imply costly maintenance operations. NoSQL databases such as key-value, column oriented, document oriented or graph databases, have been designed for specific purposes and do not meet the requirements of SIS in terms of constraint checking and query languages. XML and associated Semantic Web technologies provide extensibility but do not scale well for scientific data. Thus, only a multi-paradigm approach can satisfy the properties required by SIS. In this article, we propose to investigate a multi-paradigm approach for data management [2] in the context of archaeological SIS used for building and publishing a corpus. The archaeological data include textual information, quantifiable values and measures, sketches, photographies, 3D models, and lots of links among data and historical information sources.

The application of our proposal aims to provide researchers (archaeologists, historians, geologists) and stone professionals with a platform that can be used to build and to publish a digital corpus of Burgundy stones, including their extraction, and usage in constructions, from antiquity to modern times.

The Burgundy (Bourgogne) Region in France is a territory with many quarries in which some remarkable quality stones were and are still exploited. The research objective of the *Corpus Lapidum Burgundiae* is to determine statements which defined Burgundy as an innovative and influential region in the fields of art, history and architecture through the ages. The circulation and distribution of stones in space and time are analysed. At the same time, spatial analysis tools are used to identify the most important quarries and the areas of distribution in the regional territory and beyond. Indicating the specific qualities of each type of stone (density, hardness, porosity, etc.) and their associated techniques. It will be possible to understand the links between the choice of a specific stone, its technical treatment and its use as a building component. Historical sources and archaeological evidences show the dating of these changes and adjustments, and may help to determine the origin and geographical spread of Burgundy stones. The corpus as well as the analysis results are given to the scientific community, the restoration professionals, the stone sector professionals and the large public.

To meet the requirements of the *Corpus Lapidum Burgun-*

978-1-4799-3170-5/13/$31.00 ©2013 IEEE

diae project, we have developed a Web platform that relies on our SemLab architecture [3]. In SemLab, knowledge takes the form of an ontology used to define annotations: 1) to give a semantics to existing data; 2) to dynamically extend the data schema without modifying the application; and 3) to act as a bridge between data models in order to construct a multi-paradigm data management layer. We use a wiki as user interface as it corresponds to a Web platform with collaborative capabilities for establishing and publishing the digital corpus. The rest of this article is structured as follows, in section II we describe the data management layer of SemLab and we focus on the annotation model and the query operators. In section III we describe an instantiation of SemLab for the *Corpus Lapidum Burgundiae* project, we describe the master data and the ontology as well as the analysis database (one of the specificity of the project) that is dynamically populated and used by GIS analysis tools. Finally, in section IV we summarize our contribution and we discuss our future work.

II. A MULTI-PARADIGM APPROACH

In the following section we give an outline of the architecture, we describe our annotation model and we define the basis for a multi-paradigm query language.

A. SemLab architecture overview

The SemLab architecture is a long term project that aims at defining and developing components for SIS. The first implementation of SemLab with an annotation wizard has been described in [4]. Data model extensibility has been published in [3] and the annotation model in [5].

Master data are strongly structured and they can be identified during the analysis phase, they are recognized by all the application partners and evolve very rarely [6]. In our architecture, data that comply to master data properties are duplicated in a Relational-DBMS, and data which need specific models or data for which it is not possible to set-up a schema are separately stored in specific storage systems.

Ontology-based annotations are used as links between data modelled with different paradigms and the semantics of the domain. Most of the existing annotation models share a common representation written as a triple (s, p, o) where s is the annotated resource; p is the relationship being specified by the annotation; and o is the annotating resource. The ontology is used to constrain the annotation components. Thus, ontology-based annotations are formal annotations that can be understood by a machine and allow making analysis and treatments in an automatic way. Moreover, by using reasoning tools on annotations it is possible to check their consistency and to discover implicit relationships among data.

In our approach, ontology-based annotations allow extensibility at two different levels: at the data schema level, they can be used to add information without modifying the existing applications because an annotation is a very simple and universal structure which allows the construction of generic components; and at the data model level (among different models), they can be used to connect, in a transparent way, data modelled by different paradigms with master data.

Figure 1 summarizes SemLab architecture into three layers. A **data access layer** is in charge of persistence and integrates various data management systems. For example, archaeological domain manipulates essentially documents, that can be stored in NoSQL document-oriented databases such as MongoDB[1]. A **multi-paradigm management layer** includes a specific repository for master data, a persistence service for annotations, and a triple store for the ontology. Annotations can be stored in a Relational-DBMS, in a NoSQL graph database as Neo4j[2] or in a column oriented NoSQL database. This layer also includes reasoning tools such as Pellet and a SPARQL service to query the ontology. An **application layer** includes domain specific applications such as Wikis to publish results or spatial analysis tools to produce maps.

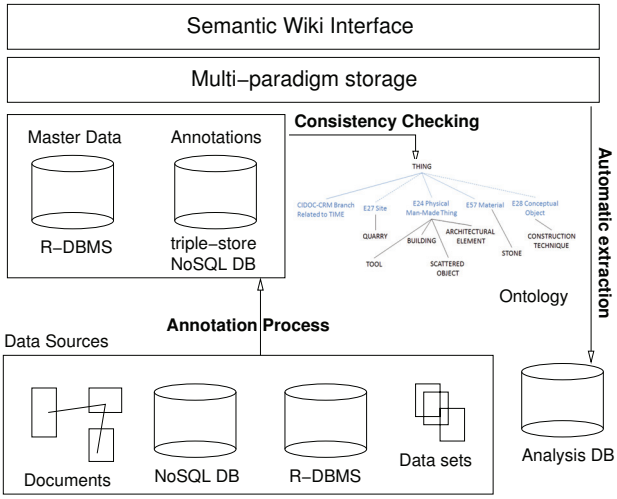

Fig. 1. SemLab Architecture

B. Ontology-based annotation model

Our annotation model defines three basic structures of annotation: simple, complex, and recursive. They share the same basic triple structure (s, p, o), where s, p, and o are constrained by ontology terms [3]. Thus, it is possible to do consistency checking and to develop tools to guide users during the annotation process by using the structure and rules of the ontology. We have defined two constructors to build annotations: 1) complex annotations as a list of annotations on the same subject ($+$ is the set builder, it merges two simple annotations which subjects are the same); and 2) recursive annotations allow adding annotations on a previous one to explain how the object and the predicate are linked with the subject (. is the ordered list builder, it associates two annotations if the subject of one is the object of the other).

For example, for one subject s, by using simple annotations and operators, we can obtain an annotation string which is a finite set of triples. One can easily define an isomorphism between string representation and oriented graph (tree). Figure 2 gives an example of the tree for the following annotation string: $((s, p_1, o_1)((o_1, p_2, o_2), (o_1, p_3, o_3)((o_3, p_4, o_4))), (s, p_5, o_5))$.

[1]MongoDB:http://mongodb.org/
[2]Neo4j:http://neo4j.org/

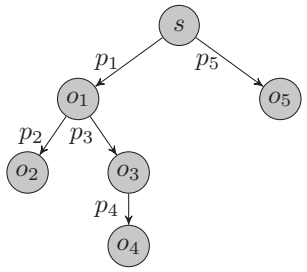

Fig. 2. Annotation tree

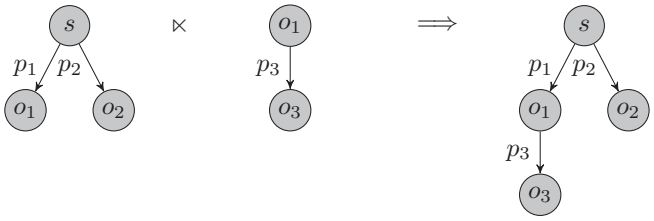

Fig. 3. Example of semi-join operator

C. Formal basis for a query language

Having good theoretical basis is essential to build tools for managing and querying master data and annotations. Formal annotation models such as K-relations have shown their capabilities; by annotating relational data with elements from a particular algebraic structure (usually a commutative semiring) it is possible to compute the corresponding annotations for query results, and also to compute their provenance [7].

To combine the expressiveness and flexibility of annotations with the theoretical formalisms of K-relations, we characterized a semiring structure $\mathcal{K} = (K, +, ., \Box, \Diamond)$ for our annotation model based on (s, p, o) triples. Annotations can be formally represented by a string belonging to a given alphabet $\Sigma = \{\Box, \Diamond, a, ..., z, A, ..., Z, (\ ,\ )\ ,\ ,\}$. Σ^* is the set of words in alphabet Σ and $K \subset \Sigma^*$. Let $(s, p, \Box)$ be the neutral annotation for $+$. It means that the annotation is not complete and should not be used as long as the object is equal to $\Box$. Let $(s, p, \Diamond)$ be the neutral annotation for. it means that the annotation is supposed to be false. In [8] we show that $\mathcal{K}$ conforms to semiring. We characterized the behaviour of a subset of relational algebra operators (union, semi-join, selection, projection). We give an example of an informal definition for the semi-join operator. Two annotation trees $T1$ and $T2$ are compatible for semi-join if the root s of $T2$ is identical to one of the leaves l in $T1$. Then $T1 \ltimes T2$ is the annotation tree $T1$ completed by the concatenation of $T2$ starting from l (figure 3). Note that, the operator . is not commutative so the join operator of relational algebra cannot be defined on annotations.

Therefore, we have a specification of how annotations behave towards relational queries and so it would be possible to define an unified language to query in parallel annotated data as well as their associated annotations.

III. The *Corpus Lapidum Burgundiae* application

In this section, we describe the instantiation and the implementation of specific components of SemLab for a collaborative Web platform for the *Corpus Lapidum Burgundiae*. The development version is available at http://care.u-bourgogne.fr:9874/atlaspierre/, the final stable version will be available at http://lapierre.u-bourgogne.fr in mid-November 2013.

A. Instantiation of SemLab

The objective of the platform, based on Web 2.0 and Semantic Web technologies, is to facilitate the processes of interpretation and analysis of documents and data using the annotation mechanism. From a technical point of view, the user interface of the platform is deployed as a semantic Wiki. Archaeological data manipulated by researchers can be organized into three levels: 1) the raw data or source materials that, are an aggregation of textual and multimedia resources (images, documents, sketches, etc.); 2) the structure that takes the shape of a classic relational database which stores master data; and 3) the meaning of a whole document or of a part of a document. This level describes essential information such a semantic context, provenance, quality, and makes an intensive use of ontology-based annotations. A triple store stores annotations and allows extensibility of the data structure.

B. Knowledge modelling

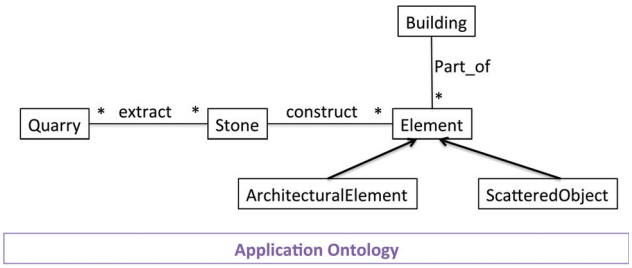

Fig. 4. Extract from the conceptual model of the *Lapidum Burgundiae Corpus*

The first stage is the identification of salient concepts and properties which let us build a conceptual model. In a second stage, we construct an ontology for our application by specializing a domain ontology and by selecting and organizing all the concepts and properties identified in the previous stage. Concepts and properties that can be easily quantified are stored as master data, *i.e.* stones, buildings, and quarries (figure 4).

The master data structure identified from the conceptual model is translated into Wiki templates that allow users to define the structure of a generic article [9] used as starting point for the creation of new articles having the same structure. Semantic Forms[3] developed for MediaWiki allows to define such templates. Other templates are defined to provide users with a synthetic article aggregating some essential information (right part of figure 5). Semantic Forms and Wiki templates are used to generate automatic annotations. Wiki makes a large use of links towards external resources (like other corpus such as Mérimée [4], Palissy[5], CARE [4]). All kind of links in and outside the Wiki can be annotated.

[3]http://www.mediawiki.org/wiki/Extension:Semantic_Forms/fr

[4]Mérimée is a database on the French monumental heritage. http://www.culture.gouv.fr/culture/inventai/patrimoine/

[5]Palissy is a database on the French movable property. http://www.culture.gouv.fr/culture/inventai/patrimoine/

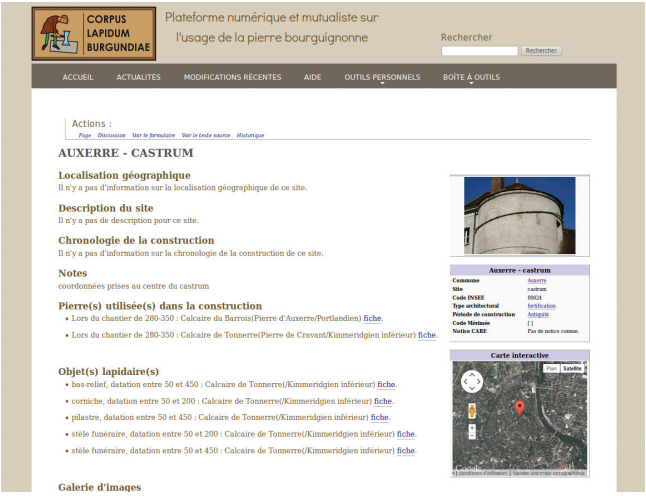

Fig. 5. Wiki interface including master data and documentation

The semantic component consists of annotations made by experts, that are guaranteed by the application ontology.

Within the cultural heritage domain, the CIDOC Conceptual Reference Model (CIDOC)[6] has emerged as a standard domain ontology. CIDOC CRM deals with concepts at a high level of generality. We have developed an application ontology as a CIDOC CRM extension covering the *Corpus Lapidum Burgundiae* concepts. The *Corpus Lapidum Burgundiae* ontology has several parts: a) concepts related to buildings, their spatial relationships and characteristics; b) stones and their characteristics; c) stone cutter tools (chisel, bush hammer, pick, etc.) and techniques of construction; and d) quarries. Figure 6 represents all these concepts (blue concepts with EXX are CIDOC-CRM concepts). Individuals of the ontology are used to populate values in lists for some master data attributes.

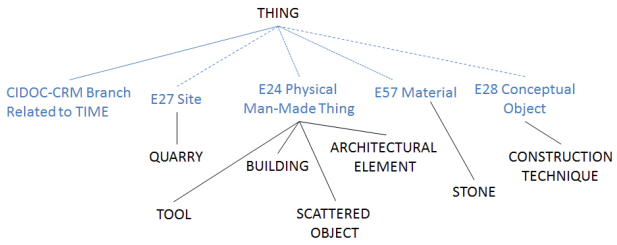

Fig. 6. Structure of Corpus Lapidum Burgundiae ontology

Buildings (e.g. temple, chapel, amphitheatre) are individuals of the concept `Building`, with its decomposition into different `ArchitecturalElement` to describe foundation, pavement, door, column, etc. and `scattered object`. These concepts have been placed under the concept `E24 Physical Man-Made Thing` CIDOC-CRM like `TOOL` concept. Indeed, CIDOC-CRM defines this concept as "all persistent physical items that are purposely created by human activity". `Stone` concept is a specialization of `E57 Material`, `Quarry` is a `E27 Site` and `Construction Techniques` is `E28 Conceptual Object`.

[6]http://www.cidoc-crm.org

By using Mediawiki hooks we have developed a storage mechanism to set up a dedicated database based on master data and annotations that is used by researchers to perform spatio-temporal analysis. Results are then published into the Wiki.

IV. Conclusion

In this article, we give the foundations for a multi-paradigm data management system for SIS. We show how ontology-based annotations can help researchers to establish relationships between data and domain knowledge. The semantics of annotations is given by an ontology which describes accurately the domain knowledge. We have extended previous works on annotation from the database research community to define a formal model for ontology-based annotation based on a semiring algebraic structure. An archaeological application based on a combination of Wiki and Semantic Web technologies is described. This combination preserves the key advantages of both technologies: the simplicity of wiki systems as shared content authoring tool, and the power of Semantic Web technologies w.r.t. structuring and retrieving knowledge. Our future work is directed towards well-founded theoretical models in order to define a meaningful query language. For that purpose, we proposed an analogy between relations and semantic annotations for positive relational algebra operators. This analogy has been used to define operators in an API and will allow us to develop extensions of a query language on annotations for our multi-paradigm data management system.

Acknowledgement

This work is supported by the Burgundy Region (CPER) and the European Union (FEDER).

References

[1] G. Săvoiu, "The scientifiv way of thinking in statistics, statistical physics and quantum mechanics," *Romanian Statistical Review*, vol. 13, no. 11, pp. 13–23, 2008.

[2] D. Ghosh, "Multiparadigm Data Storage for Enterprise Applications," *Software, IEEE*, vol. 27, no. 5, pp. 57–60, 2010.

[3] E. Leclercq and M. Savonnet, "Enhancing scientific information systems with semantic annotations," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. ACM, 2013, pp. 319–324.

[4] P. Chevalier, E. Leclercq, A. Millereux, C. Sapin, and M. Savonnet, "WikiBridge: a Semantic Wiki for Archaeological Applications," in *Proceedings of the 38th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, 2010, pp. 193–196.

[5] E. Leclercq and M. Savonnet, "Structured wiki with annotation for knowledge management: an application to cultural heritage," *International Journal of Digital Information and Wireless Communication (IJDIWC)*, vol. 1, pp. 264–280, 2011.

[6] A. Dreibelbis, E. Hechler, I. Milman, M. Oberhofer, P. van Run, and D. Wolfson, *Enterprise Master Data Management: An SOA Approach to Managing Core Information*, 1st ed. IBM Press, 2008.

[7] T. J. Green, G. Karvounarakis, and V. Tannen, "Provenance semirings," in *PODS*, 2007, pp. 31–40.

[8] A.-C. Troya, E. Leclercq, and M. Savonnet, "Annotated Databases for Scientific Information Systems - an Application to Cultural Heritage," Université de Bourgogne, Tech. Rep., 2013.

[9] A. Haake, S. Lukosch, and T. Schümmer, "Wiki-templates: adding structure support to wikis on demand," in *Int. Sym. Wikis*, 2005, pp. 41–51.