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

Eric Leclercq, Marinette Savonnet, Andrès-Camillo Troya, Stéphane Büttner

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

Éric Leclercq*, Marinette Savonnet*, Andrés-Camilo Troya*, Stéphane Büttner†
* Université de Bourgogne, Le2I Laboratory, UMR CNRS 6306
9, Av. Alain Savary 21078, Dijon, France
firstname.lastname@u-bourgogne.fr
† 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 quality data, and to enable semantic interoperability. Archaeological data include textual information, quantifiable values and measures, sketches, photographies, 3D models, and lots of links among data and historical information sources. DBMS are essential component of SIS nevertheless, existing DBMS including NoSQL DB does not provide enough extensibility and cannot meet all the properties required by a SIS. Our contribution is a multi-paradigm data management system approach that relies on master data and ontology-based annotations. We develop a formal model for ontological-based annotations, we show that this model conforms to a semi-ring algebraic structure and we define a subset of algebraic operator to query annotations. We describe the Burgundy Stone project and we show how our approach is instantiated in a collaborative Web platform that allows researchers to build and publish a corpus.**

Keywords*—*Scientific Information System, Archaeological Corpus, Semantic 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 activities of research and development. Unlike enterprise information systems, SIS do not support activities of production or services. Thus, SIS are strongly collaborative systems involving different kinds of users (scientists of different disciplines, domain professionals, etc.).

Scientific data have the following properties: 1) they include 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 (for example observation and reanalysis data in climatology) and from different acquisition technologies (for example mass spectroscopy in biology, 3D scanner in cultural heritage, thermoluminescence sensors for dating in archaeology). A large variability of data models have been observed in the last decade that come from the evolution of scientific knowledge and methods (migration from purely experimental to statistical way of thinking [11]) and from high performance computing that allows computation at the molecular level as well as at astronomical scales. As database management systems are essential component of SIS they should provide extensibility mechanism.

DBMS are essential component of SIS nevertheless R-DBMS does not provide enough extensibility, schema evolutions usually impact applications and are costly. NoSQL databases such as key-value, column oriented, document oriented or graph have been design for specific purpose and does not meet the requirement of SIS as constraint checking and query languages. XML and associated Semantic Web technologies provide extensibility but does not scale well for scientific data. Only a multi-paradigm approach can satisfy all the properties required by SIS.

We propose to investigate a multi-paradigm approach for data management in the context of archaeological SIS used for building and publishing a corpus [6] about Burgundy Stones. 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 Burgundy (Bourgogne) Region in France has an identity strongly marked by the arts of the construction and the statuary. It's also a territory with many quarries in which some remarkable qualities stones were exploited and are still exploited today. These different aspects are the object of particular and complementary treatments, both in behalf of researchers (archaeologists, historians, geologists) and of stone sector professionals. The objective of the *Corpus Lapidum Burgundiae* project is to determine statements which defined Burgundy as an innovative and influential region in the fields of art history and architecture through the ages. We develop

a digital corpus of stone extractions in the Burgundy region and stone usages in the construction (buildings, sculptors, sarcophagus, etc.) from antiquity period to modern time. The circulation and the broadcasting of stones in the space and in the time is analyzed. At the same time, Geographical Information System (GIS) tools are used to identify the most important quarries and the areas of distribution of their products in the regional territory and beyond. Indicating the specific qualities of each type of stone (density, hardness, porosity, etc.) and associated techniques (modules, tool marks, etc.), it's also possible to understand the possible links between the choice specific of stone, technical treatment and the use as a building component. Textual and archaeological evidences show the dating of these changes and adjustments, and maybe determine the origin and geographical spread of Burgundy stones. Primary data come from existing databases, historical documents and many other types of resources such as photographies, sketches. The results are given to the scientific community, the restoration professionals, the stone sector professionals and the large public, by the development of a collaborative Web platform. For historians and archaeologists working on these aspects, the platform should renew the perception of technical and economic operations for the old building in Burgundy; especially regarding trade flows of stones. Moreover, historians of art and architecture should also find something to think about the propagation of models and issues regarding stylistic affiliation. It's also a way to provide information to restoration professionals of old buildings and to make advertising to the "Pierre de Bourgogne" industry with scientific references its current production with the most distinguished buildings in Burgundy and elsewhere.

To meet the required feature of the *Corpus Lapidum Burgundiae* project, we have developed a Web platform that relies on our framework architecture SemLab [9]. In SemLab, knowledge takes the form of a domain ontology used to define ontology-based annotations which are used: 1) to give a semantics to existing data; 2) to extend dynamically schema without modifying application and; 3) to bridge data model in order to construct a multi-paradigm data management layer. We use a wiki as user interface to meet the requirements of 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 data management layer of SemLab architecture and we focus on the annotation model and query operators. In section III we describe an instantiation of SemLab for the *Corpus Lapidum Burgundiae* application, we describe the master data and the domain 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 **??** we summarize our contribution and we discuss our future work.

II. SemLab: a multi-paradigm approach

We propose a multi-paradigm data management system approach [5] that relies on master data and ontology-based annotations. 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. Architecture outline

The 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 [4]. SemLab uses a hybrid register/repository architecture style for the master data. In this architecture style, the most important data are duplicated in a RDBMS, and data which need a 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 ([10]) share a common representation written as a triple (s, p, o) where s is the subject or the annotated resource; p is a predicate or a relationship being specified by the annotation; and o is the object or the annotating resource. The domain ontology is used to constrain the annotation components. Thus, ontology-based annotations are formal annotations [10] that can be understood by a machine and allow to make analysis and treatments in an automatic way. Moreover, by using association reasoning tools on annotations it is possible to check their consistency and to discover implicit relationship among data.

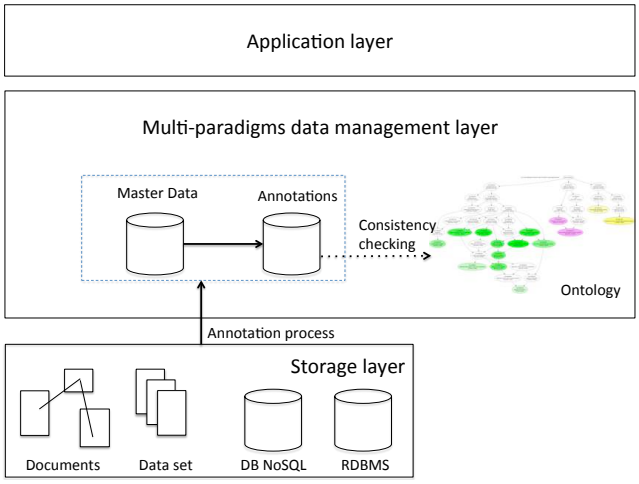

Fig. 1. SemLab Architecture

Figure 1 summarizes SemLab architecture, from bottom to top:

1) A data access layer 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].
2) A multi-paradigm management layer includes a specific repository for master data, a persistency service for

[1]MongoDB:http://mongodb.org/

annotations, and a triple store for the ontology. Master data can be stored in a key value NoSQL database or in a RDBMS depending on their volume. Annotations can be stored in a RDBMS, in a NoSQL graph database as Neo4j[2] or in a column oriented NoSQL database. This layer also include reasoning tools such as Pellet and SPARQL service to query the ontology;

3) An application layer includes domain specific applications such as Wiki to publish results or spatial analysis tools to produce maps.

In our approach ontology-based annotation allows extensibility at two different levels: at the data schema level, they can be used to add information without modifying the existing applications because annotation is a very simple and universal structure which allows to develop generic components; and at the models level (among different models), they can be used to connect, in a transparent way, data modelled by different paradigms with master data. Furthermore, SemLab allows traceability of data and quality control as well as semantic interoperability.

B. Ontology-based annotation model

Annotated databases include annotations as additional information used to allow a better understanding of data. They offer mechanisms to create, store and query annotations linked with tuples. First, they were studied by the database research community, for specific purposes. Annotation models have been developed to deal with uncertainty, trustworthiness, multi-set data and incomplete information [7]. All of this models have simple annotation structure in which terms must conform to a specific semiring. Moreover, annotations cannot be restricted by using constraints and consistency checking is not possible.

Our annotation model defines three basic structures of annotation: simple, complex, and recursive. They share the same basic triple structure (s, p, o), s, p, o are constrained by a domain ontology terms [9] and thus, allows to develop consistency checking and tools to guide users in the annotation process by using the structure and rules of the ontology.

The definitions of the three basic structures are the following:

1) Simple annotation has the structure (s, p, o) where s and p cannot be null. These kind of annotations can be compared to constraints on attribute in the database context;
2) Complex annotation is a list of simple annotations related to the same subject;
3) Recursive annotation is used to explain or to give more details on how the object and the predicate are linked together with the subject by a sub-annotation which is a simple or a complex annotation.

To combine annotations to operators have been defined: $+$ is the set builder and . the ordered list builder for complex and recursive annotations respectively. For one subject s, by using simple annotations and operators, we can obtain an annotation

[2]Neo4j:http://neo4j.org/

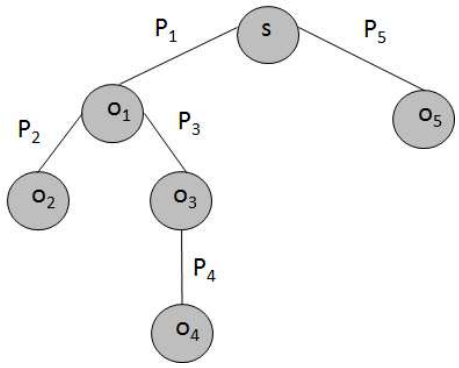

Fig. 2. Annotation tree

string which is a finite set of triples. One can easily find an isomorphism between string representation and oriented graph (tree). Figure 2) give an example of the tree for the following annotation string:
$((s, p_1, o_1)((o_1, p_2, o_2), (o_1, p_3, o_3)((o_3, p_4, o_4)), (s, p_5, o_5))$

Let A be the set of all annotations, $x_1 = (a, b, c) \in A$ and, $x_2 = (d, e, f) \in A$, the two operators are defined as follows (for the sake of brevity, we use a tree representation):

1) Addition: $x_1 + x_2$ is defined by a connection of annotation trees using their subject, it merges two different simple annotations into a complex annotation which subjects are the same ($a = d$). As a result, a is annotated by two objects c and f with respect of the predicates b and e (figure 3);
2) Product: $x_1.x_2$ is defined by the creation of new graph with a path going from a to f (annotations are concatenated), c and dd must be equal. This operator puts two different simple annotations into a recursive annotation where the subject a is annotated by c which is annotated in his turn by f (figure 4). It is used to give details on the previous annotation.

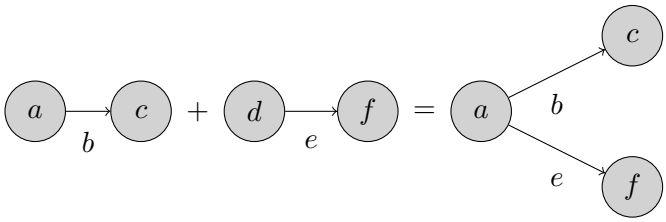

Fig. 3. Addition of two annotations

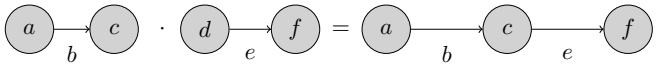

Fig. 4. Product of two annotations

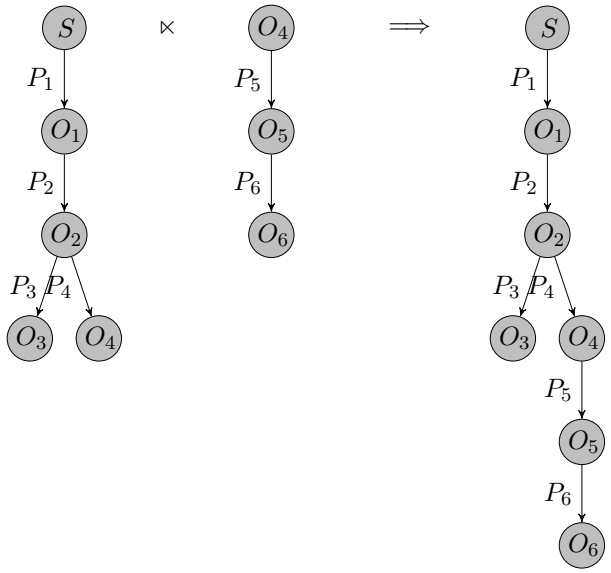

Fig. 5. Semi-join operator

C. Formal basis for a query language

It is essential to have good theoretical basis to manage and query data and annotations. Formal annotation models as K-relations showed their ability; 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 expressiveness and flexibility of annotations with the theoretical formalisms of K-relations, we characterized a structure of semiring $\mathcal{K} = (K, +, ., \Box, \Diamond)$ for our annotation model based on (s, p, o) triples. Annotation 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. 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 it means that the annotation is supposed to be false. We showed that $\mathcal{K}$ is a semiring [12].

We characterized the behaviour of a subset of relational algebra operators (union, semi-join, selection, projection) on annotations[12]. For example, we give an informal definition of union and semi-join:

- two annotation trees $T1$, $T2$ are compatible for union if they have the same subject s as root, then $T1 \cup T2$ is the annotation tree having s as root without duplicated sub-trees;
- two annotation trees $T1$, $T2$ 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 5). Note that, x 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. Description of 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*. We describe the domain ontology and a specificity of the project, i.e. an analysis database.

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 users interface of the platform is deployed a semantic Wiki.

Archaeological data manipulated by researchers can be organized into three levels: 1) the raw data or source material which, in *Corpus Lapidum Burgundiae*, are an aggregation of textual and multimedia resources (images, documents, sketches, etc.); 2) the structure which takes the shape of a classic relational database which stores master data and a triple store which stores annotations and allows extensibility of the data structure; 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. During search or analysis of the corpus all the three levels can be queried using an API.

B. Master data and wiki template

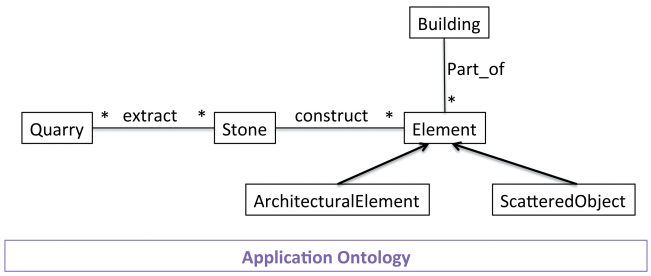

Fig. 6. Extract from the conceptual model of the *Lapidum Burgundiae Corpus*

The first stage is the identification of salient concepts and properties which let us to build a conceptual model. Three groups of elements in the conceptual model have been identified: **quarries** from where are extracted (or supposed to be extracted) **stones** and **buildings** in which stones are used (as an architectural elements, ground, wall or as a scattered object which is somewhere else, in a museum for example). In a second stage, we construct an application ontology for our application by specializing a domain ontology and by

selecting, organizing all the concepts and properties identified in the previous stage. Concepts and properties that can easily quantifiable are stored as master data, i.e. concept that are used to describe stones and their properties, properties of buildings, manufacturing techniques, tools, etc. (figure 6). Individual of the ontology are used to populate values in lists for master data attributes.

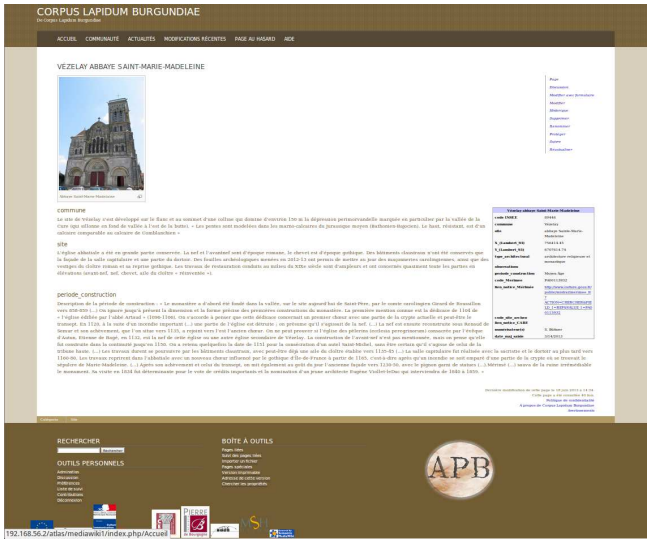

Fig. 7. Wiki interface including master data and documentation

The master data structure from the conceptual model is translated into Wiki templates. Moreover, Wiki template allows users to define the structure of a generic article [8] 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 with automatic annotation capabilities. Other templates are defined to provide users with a synthetic articles aggregating some essential information (right part of figure 7). Semantic Forms and Wiki templates have been extended to automatically store master data in a RDBMS.

Wiki make alarge use of links, links towards pages of the wiki (e.g. link between a quarry and an extracted stone), links towards external resources (like other databases such as Mérimée [4], Palissy[5], CARE [3]) (figure 8). All kind of links in and outside the Wiki can be annotated.

C. Ontology-based annotations

The semantic component consists of annotations made by experts, that are guaranteed by the application ontology. In computer science, ontologies are defined as a formal specification of a shared conceptualization [2] which theoretical basis is the description logic.

[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. 8. Templates and links

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. Typically, application ontologies are a mix of concepts that are taken from domain ontologies and from specific application. 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) stone and their characteristics; c) stone cutter tools (chisel, bush hammer, pick, etc.) and technique of construction; and d) quarry. Figure 9 represents all these concepts (blue concepts with EXX are CIDOC-CRM concepts).

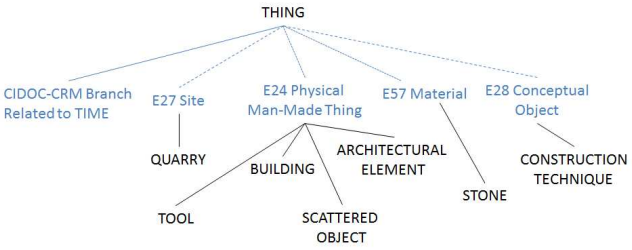

Fig. 9. Structure of Corpus Lapidum Burgundiae ontology

Buildings (e.g. temple, chapel, amphitheatre) are individual of the concept `Building`, with its decomposition into different `ElementArchitectural` to describe foundation, pavement, door, column, etc. and `scattered object`. These concepts have been placed under the concept `E24 Physical`

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

`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 technique` is `E28 Coneptual Object`.

CIDOC-CRM covers specific concepts related to time [1]. The concept `E2Temporal Entity` describes all phenomena which happen over a limited extent in time. Time model extensions are based on following criteria: some absolute benchmarks and a relative chronology based on intervals.

IV. Conclusion

In this article, we have demonstrated that extensibility and semantic data quality required by scientific applications can be achieved by using a multi-paradigm data management system that shares with the Semantic Web the same theoretical foundations. Ontology-based annotations allow researchers to establish relationship between data and domain knowledge. The semantics of annotations is guaranteed by an ontology which describes accurately domain knowledge. We have extended previous works on annotation from the database research community to define a formal model for ontology-based annotation. We have proved that our extension comply with a semiring algebraic structure that guarantee essential properties for query languages. We have instantiated our framework architecture SemLab to meet the requirement of a collaborative web platform.

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. In the near future, we will instantiate SemLab for two other applications one for a corpus of french toponym and another one for a corpus describing medieval sculptures.

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 extension of a query language on annotations for multi-paradigms data management system.

Acknowledgment

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

References

[1] Ceri Binding. Implementing Archaeological Time Periods Using CIDOC CRM and SKOS. In Lora Aroyo, Grigoris Antoniou, Eero Hyvnen, Annette Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *The Semantic Web: Research and Applications*, volume 6088 of *Lecture Notes in Computer Science*, pages 273–287. Springer Berlin Heidelberg, 2010.

[2] Willem Nico Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, Twente University, 1997.

[3] Pascale Chevalier, Eric Leclercq, Arnaud Millereux, Christian Sapin, and Marinette Savonnet. WikiBridge: a Semantic Wiki for Archaeological Applications. In *Proceedings of the 38th Conference on Computer Applications and Quantitative Methods in Archaeology (CAA)*, pages 193–196, 2010.

[4] Allen Dreibelbis, Eberhard Hechler, Ivan Milman, Martin Oberhofer, Paul van Run, and Dan Wolfson. *Enterprise Master Data Management: An SOA Approach to Managing Core Information*. IBM Press, 1re edition, 2008.

[5] Debasish Ghosh. Multiparadigm Data Storage for Enterprise Applications. *Software, IEEE*, 27(5):57–60, 2010.

[6] Jim Gray, David T Liu, Maria Nieto-Santisteban, Alex Szalay, David J DeWitt, and Gerd Heber. Scientific data management in the coming decade. *ACM SIGMOD Record*, 34(4):34–41, 2005.

[7] Todd J. Green, Gregory Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, pages 31–40, 2007.

[8] Anja Haake, Stephan Lukosch, and Till Schümmer. Wiki-templates: adding structure support to wikis on demand. In *Int. Sym. Wikis*, pages 41–51, 2005.

[9] Éric Leclercq and Marinette Savonnet. Enhancing scientific information systems with semantic annotations. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 319–324. ACM, 2013.

[10] Eyal Oren, Knud Möller, Simon Scerri, Siegfried Handschuh, and Michael Sintek. What are semantic annotations. *Relatório técnico. DERI Galway*, 2006.

[11] Gheorghe Săvoiu. The scientifiv way of thinking in statistics, statistical physics and quantum mechanics. *Romanian Statistical Review*, 13(11):13–23, 2008.

[12] Andrès-Camilo Troya, Éric Leclercq, and Marinette Savonnet. Annotated Databases for Scientific Information Systems - an Application to Cultural Heritage. Technical report, Université de Bourgogne, 2013.