Access and Annotation of Archaeological Corpus via a Semantic Wiki

Eric Leclercq, Marinette Savonnet

Access and Annotation of Archaelogical Corpus via a Semantic Wiki

Éric Leclercq and Marinette Savonnet

University of Burgundy
Le2I Laboratory - UMR 5158
B.P. 47 870, 21078 Dijon Cedex - France
Firstname.Lastname@u-bourgogne.fr

Abstract. Semantic wikis have shown their ability to allow knowledge management and collaborative authoring. They are particularly appropriate for scientific collaboration. This paper details the main concepts and the architecture of WikiBridge, a semantic wiki, and its application in the archaelogical domain. Archaeologists primarily have a document-centric work. Adding meta-information in the form of annotations has proved useful to enhance search. WikiBridge combines models and ontologies to increase data consistency within the wiki. Moreover, it allows several types of annotations: simple annotations, n-ary relations and recursive annotations. The consistency of these annotations is checked synchronously or asynchronously by using structural or domain constraints.

1 Introduction

Document analysis is crucial to archaeologists when trying to understand the evolution of patrimonial buildings and sites. Documentary sources provide partial evidences from which researchers will infer possible scenarios on how a building may have been transformed through the ages. The aim of the CARE project (Corpus Architecturae Religiosae Europeae - IV–X saec.) is the constitution of integrated corpus of the French Christian buildings dated from the 4th to the beginning of the 11th century. It aims at facilitating work of comparisons, exchanges and discussions with numerous foreign researchers and specialists. The project has been launched in France on January 1st, 2008 after acceptance of the French National Agency for Research and will last 4 years (2008-2011). More than sixty researchers from about twenty universities, diverse research institutions and heritage management institutions are working on. Various categories of staffs are involved: field archaeologists, historians, art historians, draftsmen, topographers, PhD students, etc. They are collecting and analyzing data concerning approximately 2700 monuments. The corpus of multimedia documents (including texts, maps, and photographies) concerning every known building will be gradually published in the form of classic books.

The request of a Web 3.0 application with a collaborative component and the need of document management led us to choose a solution based on a wiki rather than a database. A prototype is available at http://care.u-bourgogne.

fr. Despite the power of wiki (free input, rich user-interface, traceability, bi-directional links between pages, etc.), it is difficult to answer a specific query because of the purely textual information stored. Consequently, an approach which can provide a semantic annotation of the content is necessary. In addition, requirements for interoperability and data exchange must be taken into account since the design phase of the application. The semantic web thereby provides such kind of solutions by increasing the expressiveness of data representation, and by allowing reasoning tools and semantic search.

The computer application part of the project has started in September 2008, a prototype has been held with MediaWiki and Semantic MediaWiki through May to July 2009. After this prototyping phase we notice that some functionalities are missing in Semantic MediaWiki. For example n-ary relations are not fully supported, the scope of a tag is generally a document. As in Semantic MediaWiki, annotation can be enhanced as the knowledge evolves. In most of semantic wiki approaches, subjects of annotation are the whole document, we propose a recursive annotation model to cope with different levels of knowledge granularity as well as extension of domains. In [8], the authors propose an equivalent representation between OWL concepts and Semantic MediaWiki constructs. WikiBridge approach allows to annotate an element with different annotations in several parts of documents. This functionality can be used to highlight a specific object described in a document. In [7], authors provide facilities to ensure the content quality of a wiki, including constraint and auto-epistemic operators. They introduce semantic checking with three kinds of constraints that are mostly structural: 1) domain and range; 2) concept cardinality; and 3) property cardinality. In WikiBridge, structural constraints checking is included in the annotation process while domain dependent constraints are checked asynchronously.

The rest of the paper is organized as follow: Section 2 describes our architecture through the physical and logical structure, the semantic layer, the information access layer. Section 3 concludes the paper.

2 WikiBridge's architecture

Our proposal is to use MediaWiki to develop a numerical corpus by integration of individual contributions. We have extended MediaWiki with some DBMS capabilities and semantic tools: form based acquisition interface, annotations, query engine.

2.1 Document Structuring

The archaeologists' work is focused on documents: documentary sources and documents of excavations are used to analysis of buildings; in result paper forms are produced. Moreover, document exchange, information retrieval and integration are uses of these various documents. The multitude of purposes

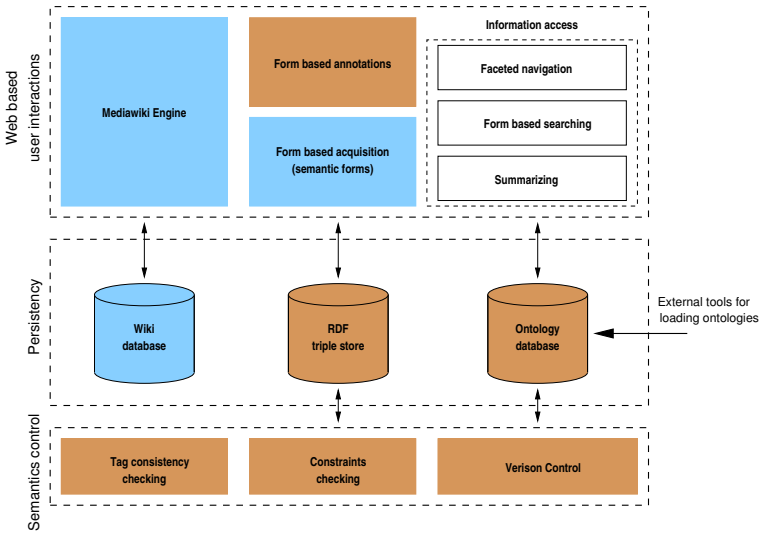

Fig. 1. WikiBridge's Architecture

and the diversity of document content types led to different structuring needs. Standards such as Open Document Architecture and SGML (Standard Generalized Markup Language) consider that document has at least two structures of representation.

The physical structure defines the document presentation. This structure consists of physical elements such as style sheets (CSS, XSLT).

The logical structure defines an organization (relationship of composition, sequence) of information contained in the document. This organization represents the different parts of the document. It is composed of titles, chapters, paragraphs, notes, figures, etc. Organization of a document in the CARE corpus is as follows: topography, documentary sources, a succession of states describing the evolution of architectural building. In each state, plan of building with concepts of space, architectural elements and function are known from elements of relative dating such as construction techniques, building materials, sepulchers etc. This logical structure can not structure the knowledge and therefore does not allow easy information access.
The logical structure of the document could be stored in a database with attributes of type LONG, but a specific tool must be developed to display, to edit the different documents and their structure. Wiki is a suitable tool for representing these two structures.

The semantic structure has been introduced by other authors [3]. It represents the information itself i.e. the meaning of document content. The semantic structure describes information that a user or an agent asks when searching. It is superimposed on the document and allows to manipulate the rules and not chapters or paragraphs.

2.2 Physical and logical structure layers

The physical structure is covered by MediaWiki and the logical structure is managed by Semantic Forms extension[1] for MediaWiki. Corresponding modules are described in light grey in figure 1. Each part of the paper document – a word file– (figure 2.a) is represented by a model (figure 2.b), models can be composed. A model is defined by using a mini-scripting language and forms are created on-the-fly on the basis of models. Two types of acquisition form have been created: a form for entering a record corresponding to atomic building and a form corresponding to a group of buildings. Some specific fields (select lists) and free text based fields are proposed. For instance, they are respectively used for selecting administrative regions of a building and describing liturgical installations in a building. Finally, a non-expert in archeology can easily feed the wiki (figure 2.c), by copying and pasting, from paper forms already made by archaeologists. Results are stored in the wiki database.

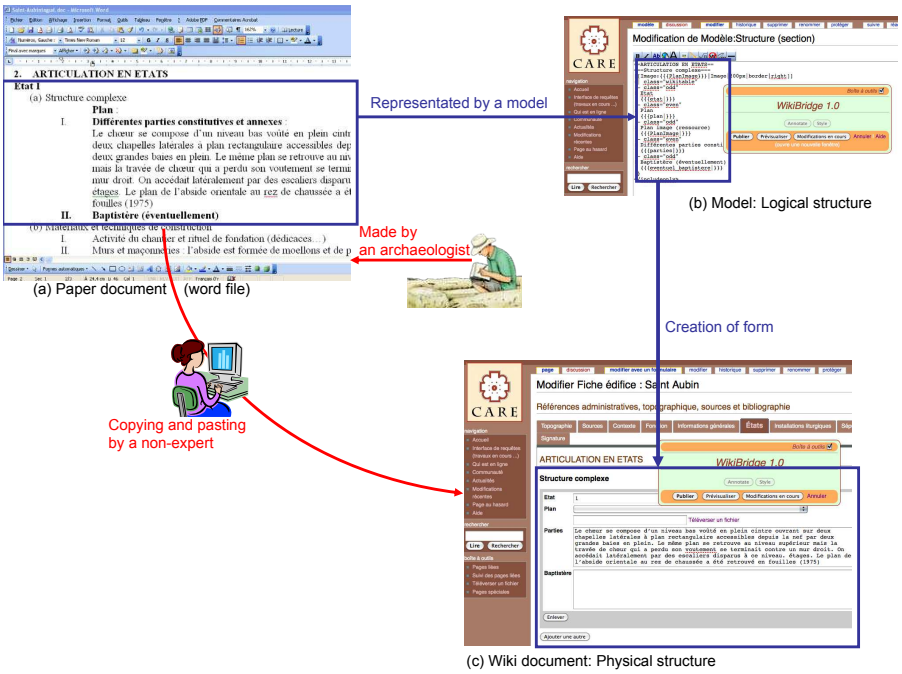

Fig. 2. Acquisition form

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

2.3 Semantic layer

To improve quality of search, we expanded MediaWiki with semantic components (medium grey box in figure 1). Annotations, made by experts, are guaranteed by a domain ontology. Experts directly enter and modify annotations through an extension of the wiki's editing interface (figure 3) which relies on the form based annotation component. We restrict access to ontological knowledge management to a predefined set of Wiki users: we argue that implementing such functionality without adequate process-level support might have uncontrolled consequences on the operation of the overall wiki system. Knowledge engineers interacting with archaeologists create the domain ontology with standard tools like Protégé. The scope of domain ontologies includes concepts and relations of thematic area. Specific extensions of domain ontologies are defined in the context of a distinct usage of the more general knowledge model [4]. CIDOC Conceptual Reference Model [2] [2] is a domain ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. We have made a specific extension of CIDOC ontology for the European Christian buildings. It consists of:

- found objects : type of buildings, architectural elements (e.g. nave), liturgical installations (e.g. altar), wall structures and pavements ...
- religious aspects of these objects: function, consecration;
- spatial aspects: relative position of an object with another,
- architectural evolution of objects: creation, destruction and modification by adding or deleting element.

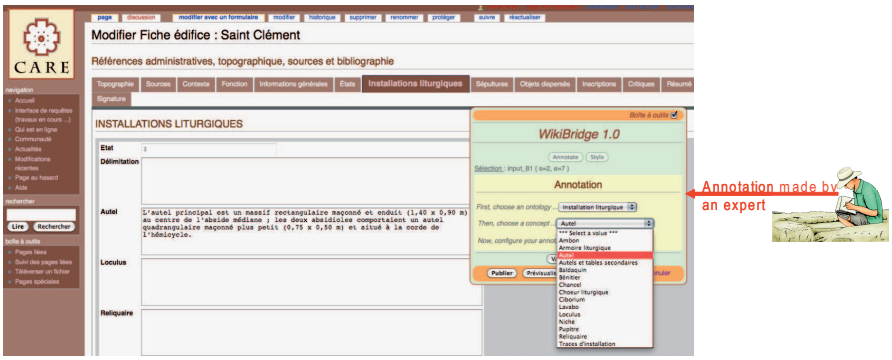

Fig. 3. Annotation Interface

[2] http://cidoc.ics.forth.gr/

Persistency Two persistency levels have been distinguished:

- A persistence level related to knowledge which includes ontology and annotations. We explicitly store in a relational database the conceptual model defining the structure of the domain ontology (Figure 4). Ontology is loaded from Protégé by a specific program. As a result, annotations are stored in RDF data in the RDBMS.
- A persistence level related to document structure is realized by MediaWiki with the Semantic Forms extension.

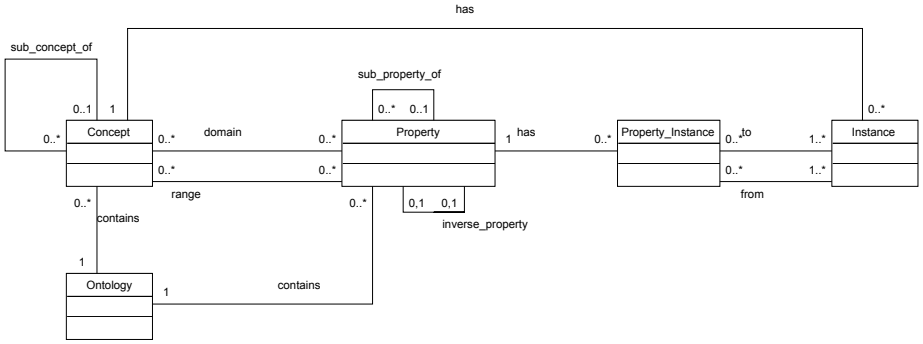

Fig. 4. Schema database for ontology

Annotations and consistency checking Two types of annotations have been identified:

- Simple annotation allows to tag a subject by describing some of its properties by attributes values (literal) couples. These kind of annotations can be compared to a restriction on attribute's domain in the database context. Theses annotations are mostly related to the ABox level.
- Complex annotation references TBox and ABox levels:
  - n-ary relation allows to map a subject with two or more values and references to other elements (subjects). In this case, some values properties reference another subject. For example we can annotate an altar with its dimension, its building material, its location in the nave. The nave is detailed in another part of the document.
  - recursive annotations allows to explain or clarify an attribute by a sub-annotation which is a simple or a complex annotation.

Moreover, annotations related to the same subject can be expressed in different parts of a document or in different documents. We propose a mechanism

to merge annotations and to visualize all the annotations related to one subject in the annotation interface.

In order to implement our annotation mechanism, we choose to use the model of semantic values proposed by Sciore et al. [5] for mediation of relational databases. They define recursively semantic values by the association of a context to a simple value. A context is a set of elements which are assignment of a semantic value to a property. We extend this model by allowing values to be references to other elements (part of documents, subjects). For the aforementioned altar example, the annotations are:

$$1.3(dimension = "width", unit = "m") \tag{1}$$
$$0.95(dimension = "height", unit = "m") \tag{2}$$
$$2.4(dimension = "length", unit = "m") \tag{3}$$
$$marble(buildingMaterial = "stone") \tag{4}$$
$$\#nave143(spatialRelation = "contained"(spatialPosition = "center")) \tag{5}$$

Annotations (1), (2), (3) should be merged but the semantic values model treats them as separated annotations. We can introduce an intermediary annotation such as 3D to allow combination of multiples semantic values. The value is then a specific attribute of the annotation.

$$\begin{aligned} dimension = "3D"(&dimensionY = "width"(unit = "m", value = "1.3"),\\ &dimensionZ = "height"(unit = "m", value = "0.95"),\\ &dimensionX = "length"(unit = "m", value = "2.4")) \end{aligned}$$

Annotation modeling using semantic values allows automatic conversion of units (for example between meters and inches). The same type conversion can be used for dates from centuries to values interval. Conversion can be used in query processing or for multi-lingual support.

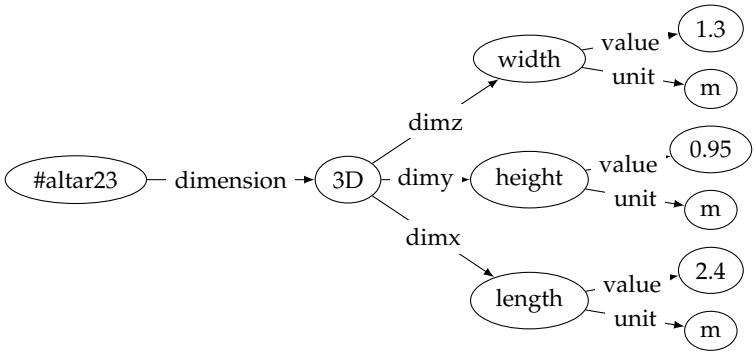

Fig. 5. RDF like transformation of semantic values

In WikiBridge, semantic values are reified in a list of atomic annotations i.e. couples (property, value), related to an object (subject), that are stored as triples in the database. An identifier is given to each atomic annotation allowing recursive semantic values. Annotation tuples can be translated in RDF and displayed as graph (figure 5).

Annotations are defined by users through a wizard that controls two kinds of constraints:

- Domain values of properties using ABox capabilities;
- Structural consistency of properties using TBox capabilities (for instance, a cathedral can have a nave but cannot have an atrium).

This two kinds of constraints can be checked using the ontology structure in OWL format. Nevertheless, some domain dependent constraint cannot be embedded in the structure. For example "a building cannot be dedicated to a saint before is death date" is represented by the following rule:

$$\texttt{isConsecrated(?b,?p)} \leftarrow \texttt{hasConstructionDate(?b,?d1)} \wedge \texttt{hasDateDead(?p,?d2)} \wedge \texttt{d1} \geq \texttt{d2}.$$

OWL is mainly based on description logics [1] (DL). Some features of DL make it difficult to use for validating data annotations through integrity constraints (IC): 1) OWL-DL works in open world assumption; 2) OWL does not use the unique name assumption. Finding inconsistent annotations require to evaluate OWL rules in a closed world assumption to detect violation.

In order to implement constraints two solutions have been tested: 1) translation of constraints in a programming language such as procedural SQL ou PHP and 2) use of a reasoner and a set of constraints stored in a file. The second solution was chosen because it allows to define and to add dynamically new constraints as knowledge evolve. The domain dependent constraints are checked when users validate an annotation while domain values and structural properties are checked when users build the annotation through wizard (figure 3). Three approaches are described in [6] : 1) skolemisation-based semantics, some constraints are tagged as IC; 2) ruled-based semantics based on interaction with logic programming that provides negation as failure under the closed world assumption and 3) query-based semantics that relies on boolean epistemic queries for expressing constraints.

2.4 Information access layer

Information access layer has been built with taking into account some features about users. We have thus identified a usage typology in accordance to 1) kind of usage: reader, investigation, clarification; 2) knowledge degree of the domain: domain specialists like archaeologist researchers and non specialists. On this basis, we can distinguish:

- general public with a general knowledge of the area who wants to find information on the known elements;

– experts understanding meaning of annotations who need access to detailed information;
– researchers who need to make analysis i.e. cross-checking data from multiple articles and make emergence of new knowledge.

To handle these different types of users, we offer three types of queries:

1. faceted browsing allows users to explore by filtering available information through an ontology tree;
2. form based searching provides semantic search by filling in parameters associated with ontology concepts. Two types of interfaces (figure 6) for building semantic queries are developed: a wizard lets users to specify search parameters to engine and users can create query models that are then stored;
3. aggregate view for each article as factbox.

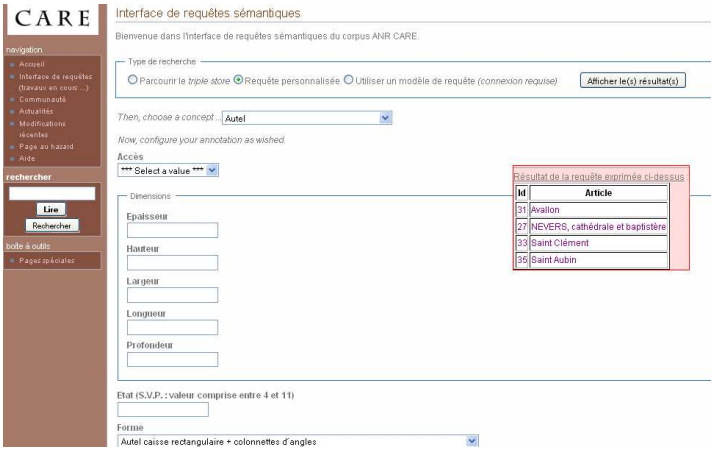

Fig. 6. Query interface

Three kinds of results can be displayed: 1) results can appear in a list containing links to articles, at the right annotation place, so where the information is given; 2) user can then manually navigate through articles interlinked; and 3) users can select annotation to be displayed in the result. From this result, users can obtain the list of the articles in which have the same annotation. This third kind of display is a mix of result list and factbox and allows more sophisticated analysis.

3 Conclusion

A feasible combination of wiki and Semantic Web technologies should preserve 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 this article, we have demonstrated that flexibility and data quality required by scientific applications can be achieved by using wiki with semantic web technologies.

We use annotations to make links between logical layer and semantic layer. The semantics of annotation is guaranteed by an ontology including constraints which allows to describe accurately domain knowledge. Our dual approach allows to cope with evolution of knowledge by modifying the ontology and annotations dynamically without modifying database schema.

Actually, we only verify structural constraints in a synchronous mode when users annotate the document. The next version of WikiBridge will automate verification of integrity constraints by Pellet reasoning engine and annotations will be marked by an ontology version. Remain the problem of inter-ontologies version consistency.

Some geomaticians of the Social Sciences and Humanities Research Institute of Dijon will conduct specific spatial analysis by providing GIS tools from end of 2010. For thorough analysis, specialized tools (GeoMondrian[3] and PostGIS [4]) interconnected by Web Services will be proposed to specifically address the spatio-temporal aspect. For simple spatial analysis, OpenLayers[5] applications are developed. For this, we are developing web services to export data to PostGIS and GeoMondrian, some web services will be used by OpenLayers applications to provide general public users with geo-analysis capabilities.

Acknowledgments This work is supported by the ANR (ANR-07-CORP-011).

References

1. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
2. M. Doerr. The cidoc crm - an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24:2003, 2003.
3. Line Poullet, Jean-Marie Pinon, and Sylvie Calabretto. Semantic Structuring Of Documents. *3rd Basque International Workshop on Information Technology (BIWIT)*, pages 118–124, 1997.
4. Sebastian Schaffert, Andreas Gruber, and Rupert Westenthaler. A semantic wiki for collaborative knowledge formation. In *In Semantics*, 2005.
5. Edward Sciore, Michael Siegel, and Arnon Rosenthal. Using semantic values to facilitate interoperability among heterogeneous information systems. *ACM Trans. Database Syst.*, 19(2):254–290, 1994.
6. Evren Sirin, Michael Smith, and Evan Wallace. Opening, Closing Worlds - On Integrity Constraints. In *OWLED*, 2008.

[3] `http://www.spatialytics.org/projects/geomondrian/`
[4] `http://postgis.refractions.net/`
[5] `http://openlayers.org/`

7. Denny Vrandecic. Towards automatic content quality checks in semantic wikis. In *Social Semantic Web: Where Web 2.0 Meets Web 3.0*, AAAI Spring Symposium 2009, Stanford, CA, march 2009. Springer.
8. Denny Vrandecic and Markus Krötzsch. Reusing ontological background knowledge in semantic wikis. In Max Völkel and Sebastian Schaffert, editors, *Workshop on Semantic Wikis*, volume 206 of *CEUR Workshop Proceedings*, 2006.