ArchaeoKM: Realizing Knowledge of the Archaeologists

Ashish Karmacharya[1,2], Christophe Cruz[2], Frank Boochs[1], Franck Marzani[2]

[1] Institut i3mainz, am Fachbereich 1 - Geoinformatik und Vermessung
Fachhochschule Mainz, Holzstrasse 36, 55116 Mainz
{ashish, boochs}@geoinform.fh-mainz.de

[2] Laboratoire Le2i, UFR Sciences et Techniques, Université de Bourgogne
B.P. 47870, 21078 Dijon Cedex, France
{christophe.cruz, franck.marzani}@u-bourgogne.fr

Abstract. The potentiality of ontology within the faculty of archaeology has recently been felt. However, the use of ontology is limited either within the data interoperability for data sharing within various heterogeneous platforms or data integration of heterogeneous dataset. Thus the full potentiality of ontology is still to be realized within the community of archaeology. We are developing a system "*ArchaeoKM*" which uses ontology beyond data integration. It uses the strength of ontology to reason the knowledge presented within. Additionally, *ArchaeoKM* involves archaeologists to define their knowledge of an excavation site through domain rules which they define through the descriptions and observations of the findings from the site. These domain rules are core of *ArchaeoKM* as they represent knowledge of the archaeologists. The advancement in semantic web technology has been the backbone principles behind *ArchaeoKM*. It relies heavily on semantic annotations to provide semantic view of the data set. The rules and the rule markup languages are major design issues of any semantic web application. *ArchaeoKM* realizes these through the rules defined by the archaeologists and transforms them to machine readable format to generate the necessary results. Moving on, the application integrates spatial operations and functions within semantic web to add value for the knowledge interpretation. The inclusion of spatial operations and functions provided by the commercial database systems enhance the capability of *ArchaeoKM* as archaeologists can combine both semantic and spatial rule to define the knowledge.

Keywords: Industrial archaeology, knowledge management, information system, ontology, spatial data

1 INTRODUCTION

With the advancement in data acquisition technology, the method of data collection has seen a tremendous leap forward. Now it is possible to collect data with very high accuracy. This has provided innumerous advantages in data manipulation but also provides challenges in managing them simply due to the size of the data. In an Industrial Archaeological project where the area for excavation is available for very small duration, this problem gets even more exaggerated. Hence, there is lots of research going on the topics of data indexation and information retrieval so that a next level could be reached where knowledge could be used to manage the findings. This level consists in identifying knowledge and managing this knowledge on data provided by archeological activities. Data are collected according to the requirement of the archaeologists and they are managed by themselves. *ArchaeoKM* facilitates them to manage them through the knowledge generated by identifying the objects excavated from the site and recording it as it is. It provides the functionality of relating the object to another in a dynamic manner so that new relationships could be created at any point of time. Actually, only archeologists are able to perform these tasks through their knowledge of the excavation sites and the objects excavated.

Industrial archeology generates huge amount of data in a very short duration, the collected data are stored in a repository without any relevant structure. Once data are stored, the process of identification of industrial findings with the help of the data repository is carried out. Three major issues have to be underlined here; first most appropriate storing structure which provides easy access to the repository consisting complex and heterogeneous data like 3D point clouds, pictures, images, videos, notes and others. Second – the most feasible process to allow archeologists to annotate, index, search, and retrieve data and documents in order to ease the identification of common archeological findings. Third – the rules to define the object and its relationship with other objects. The rules are very important as they are the one to provide a proper knowledge base for knowledge management. These rules could be based on semantic as well as spatial relationships of the object and *ArchaeoKM* supports both.

Shifting from conventional methods, *ArchaeoKM* is a web platform based on semantic web technologies and knowledge management. It is used to store data during the excavation process and to generate knowledge during the identification process and manage the knowledge generated through the rules formulated by the archaeologists. The platform facilitates the collaborative process between archeologists to generate knowledge from the data set. In general two distinct functionalities could be observed in *ArchaeoKM* – Knowledge Generation and Knowledge Management. The descriptions and observations of the archaeologists are managed through the domain ontology which is basically the representation of the site. The ontology gets populated by the identification process making it a knowledge base for knowledge handling.

2 PREVIOUS WORKS

This section presents the management of spatial data in previous and its limitation for knowledge management. This section includes also an introduction to knowledge management through the Web Semantic technologies.

2.1 SPATIAL DATA

We consider all the geometric data that can create 3D objects models and the 3D environment of those objects as spatial data. Huge amount of works have been carried out in the field of 3D object modeling and virtual reconstruction.

The approach taken in 3D MURALE (Cosmas, et al. 2001)is one of the most comprehensive information systems in the field of archaeology. The main aim of this information system is to measure, reconstruct and visualize archaeological ruins in virtual reality the ancient city of Sagalassos in Turkey. The system composed of a recording component, a reconstruction component, a visualization component and database components. The findings are managed through a database management system which takes into account various data types. Research work like DILAS (Wüst, S. und Landolt 2004)is also a good example of comprehensive 3D object modeling project which is used to model cultural heritage sites. DILAS is a generic, fully object oriented model for 3D geo-objects. The 3D geometry model is based on a topologically boundary representation and supports most basic geometry types. It also incorporates the concept of multiple Levels of Detail (LOD) as well as texture information. Different multi-resolution strategies were developed for spatial object. 3D objects are represented in 3D bounding boxes with 2D object boundaries and the actual 3D geometry. This helps in efficient query operations and automatically derived from the main 3D representation.

As all fully oriented geometry management system, the main issue of those projects is the lack of semantic information in orders to able the management of knowledge on geometrical objects. Interesting concepts on how to represent an object through the semantic information has been discussed through the applications in virtual reality (Cruz, Nicolle und Neveu 2004). The use of spatial and orientation relation of objects with respect to others can represent the objects in the proper way with respect to its surrounding. So this idea concerning the semantic relation of each object with others is an improvement for our objectives.

2.2 KNOWLEDGE MANAGEMENT

Knowledge about documents has traditionally been managed through the use of metadata. The semantic Web proposes annotating document content using semantic information from domain ontologies [(Berners-Lee, Hendler und Lassila 2001). The result is a set of Web pages machine interpretable mark-up that provide the source material. The goal is to create annotations (manually or automatically) with well-defined semantics. In the Semantic Web context, the content of a document can be described and annotated using knowledge such as RDF, and OWL. Resource Description Framework (RDF) (W3C 2004) is a formalism of knowledge representation from the semantic networks field. It is mainly used to describe resources, such as an electronic web document, by a set of metadata (author, data, source, etc.) and a set of descriptors. This metadata is composed of triplet: (objet 1, relationship, objet 2) or (resource, property, value), according to the type of description required. Web Ontology Language (OWL) (McGuinness und Harmelen 2004) is used to specify ontology or more generally some ontological and terminological resources by defining concept used to represent a domain of knowledge. Each concept is described by a set of properties, relations and constraints. The OWL formalism is derived from the description logic fields and has the capability to infer new knowledge from existing knowledge.

Semantic Web annotation brings benefits of two kinds over these systems, enhanced information retrieval and improved interoperability. Information retrieval is improved by the ability to perform searches, which exploit the ontology to make inferences about data from heterogeneous resources (Welty und Ide 1999). (Semantic annotation for knowledge management: Requirements and a survey of the state of the art 2006)

Semantic Web standards for annotation tend to assume that the documents being annotated are in Web-native formats such as HTML and XML. Annotea (Kahan, et al. 2001) is a W3C project which specifies infrastructure for annotation of Web documents. The Annotea framework has been instanced in a number of tools including Vannotea (Schroeter, Hunter und Kosovic 2003). The CREAM (Handshuh und Staab 2002) (Russell, Murphy und Freeman 2008) framework specifies components required by an annotation system including the annotation interface. Those approaches will have limited usefulness for knowledge management in this project. Actually, documents will be in many formats as clouds of points which are not XML based [semantic annotation]. The project LabelMe is a Web-based annotation tool for images that provides a drawing interface. Based on ontology, the user labels images by clicking on it and by adding a key word he enriches the ontology. The user is free to label as many objects depicted in the image as they choose. The knowledge managed by this framework is a terminological definition of graphical objects. Then, it is no possible to define an object which can be found in several documents.

In order to improve this limitation, our platform aims at not only managed the concepts used to annotate documents, but also the instances of concept with its own property values. In this manner, an object found in a cloud of points can be linked, with the help of an instance of ontology, to others documents which contain the same object. From this point, the ontology and its instances are used as an index to retrieve information and documents. The second aim of our platform is to give the possibility to archaeologists to manage index cards on findings. Those index cards represent the knowledge formalized by archaeologists and are managed through a 3D scene where 3D objects are linked index cards.

3 DATA COLLECTION AND PATTERNS

Industrial Archaeology is perhaps best suited field in archaeology to carry out our research as Industrial Archaeological Sites (IASs) are available for very short duration of time. It makes time availability very short to store them which is one of the concerns we want to address here. Additionally, the amount of data that is collected in this short span is very large and diverse. *ArchaeoKM* uses the site of Krupp factory in Essen, Germany. The 200 hectares area was used for steel production during early 19th century and was destroyed in Second World War. Most of the area has never been rebuilt and thus provides an ideal site for industrial archaeological excavation. The area will be used as a park of the ThyssenKrupp main building in 2010. Data are stored in repository as they are collected. The first challenge is to create a proper data structure which helps in retrieving those data efficiently. And it should be also understood that the amount of data that are collected is huge so the structure of database should also be able to handle huge dataset. The next one is facilitate the archaeologists to determine the rules within the data so that the knowledge could be generated from the database. Actually, we are running out of time to collect data. The first challenge consists in creating a relevant data structure which helps in retrieving those data efficiently. In addition, the data which have to be collected are huge so the system should be able to handle a huge data set.

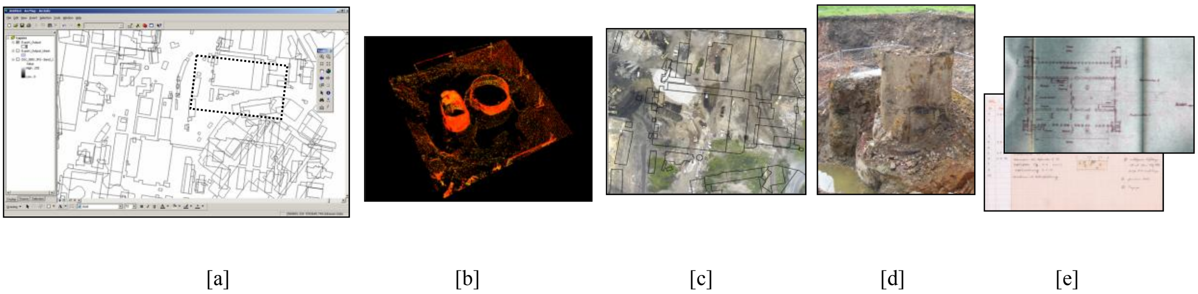


Fig 1: Heterogeneity nature of data [a] Site Plan layed out as GIS data in ArcGIS (highlighted the area of Oven) [b] Orthophoto from aerial image overlayed with the Site Plan (Oven area). [c] Point Cloud of Oven [d] Image of the Oven. [e] (top) Floor Plan (down) Archaeological notes

The nature of the dataset generated during the project is heterogeneous. It could be seen in figure 1. As could be seen the acquired data ranges from scanned point cloud from terrestrial laser scanners to the floor plans of old archive. The primary source of geometric information is provided through the point cloud. The point clouds have resolutions of 0.036 degree and are in Gauss Krüger coordinate system (GK II). It is the main data set used for the 3D object modeling. Beside point clouds, huge amount of images are also collected during the excavation. Most of the images are taken with non calibrated digital camera so do not contain any information about the referencing system. Even though they do not contain any referencing information they posses vital semantic information and could be used for the formulation of knowledge. However, there were photogrammetric flights to acquire aerial images of the area. The aerial images were processed to generate a digital orthophoto with a resolution of 10 cm. The digital orthophoto is again in Gauss Krüger referencing system (GK II). To add on this, huge archive data have been collected. Those data contains floor plans, old pictures and other semantic information. Likewise, the notes taken by archaeologists are also important to acquire semantic information of the findings. ArcGIS databases are also available depending on the site and its nature. These databases are in the GK II reference system. For our example, this database gives an overview of the site and can be overalyed with the orthophoto in order to identify the interesting locations easily as can be seen in figure 1 (b).

4 ARCHAEOKM – THE PRINCIPLE AND THE PROCESS

The main principle behind *ArchaeoKM* is the use of semantic web and knowledge management to facilitate archaeologist for handling their data. However, it does not completely bypass the conventional database system. It still uses the spatial functionalities of existing database system for its spatial rules. Details on how they are managed could be found in papers like (Cruz, et al. 2010), (Karmacharya, et al. 2009). It is collaborative Web platform based on semantic web technologies RDF (Group 2004), OWL (Bechhofer, et al. 2004), SPARQL (Prud'hommeaux und Seaborne 2008) and SWRL (Horrocks, et al. 2004) and knowledge management in order to handle the information provided by several archaeologists and technicians.

4.1 THE ARCHITECTURE

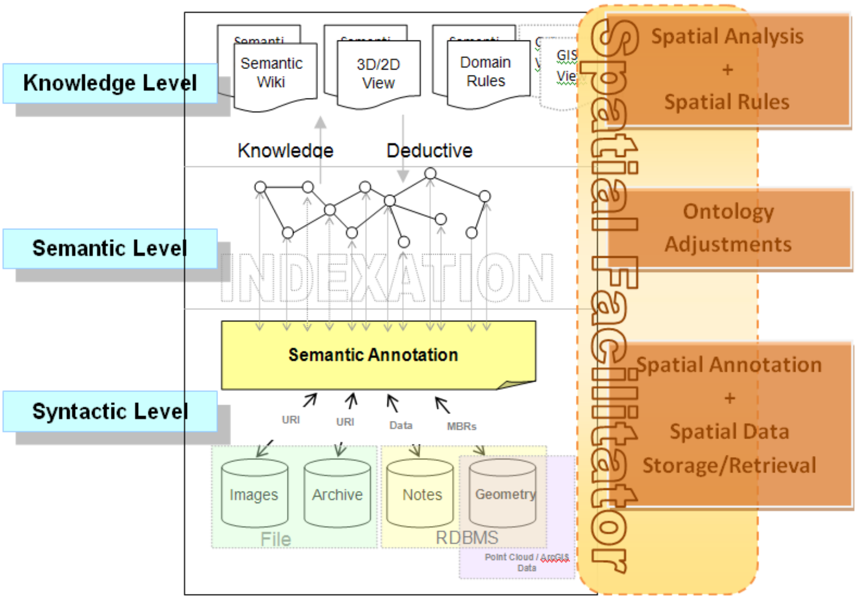


Fig 2: System Architecture

ArchaeoKM is a web based system and functions under three major levels. Each level has its own distinct functionality and is interdependent to each other. Figure 2 shows the system architecture of the system.

The bottom level is the Syntactic level. This level has all the information excavated from the site stored. As discussed earlier, they are either stored in the file formats like images or archive data or stored in the Relational Database Management System like archaeological notes or scanned/GIS data. Today, almost all the database systems have incorporated Spatial Extension included and this has made the storage and retrieval of geometric data very convenient. Additionally, they provide spatial operations and functions which allow us to analyze the geometric data spatially. The geometric information acquired through the terrestrial laser scanners is stored in the database system as spatial data types. Basically, these geometric data are the set of point clouds with 3 dimensional coordinates. They are the major data in context of the ArchaeoKM as they provide visual representations of the findings during excavation. With the help of spatial operations we can derive the bounding boxes of the object during the storing or after during the retrieval of these data. Additionally, the site plan of the area which is digitized and stored as “shp” files in ArcGIS will also be stored in the RDBMS. With the advancement in database technology, today it is possible to store the point cloud as Binary Large Object (BLOB) data type as in Oracle 11g with spatial extension or Extended Well Known Text [EWKT] as in PostGIS 1.3, the spatial extension of PostgreSQL 8.3. The ArcGIS data can be exported to the above mentioned database system either through the tools developed by ESRI or tool within the database systems themselves. An example of such tools will be the loader (shp2pgsql) and dumper (pgsql2shp) tools within PostGIS which allow converting “shp” file to spatial data of PostgreSQL and vice versa. The ArchaeoKM intends to use PostgreSQL to store its data because of its flexibility and cost efficiency over other database systems.

One of the major functionalities within the syntactic level is to define annotations on data. The data needs to be annotated against the objects indexed in the Orthophoto for the proper identification. Through annotating the data semantically, knowledge is generated. As could be seen in the figure 3, the annotations are done through different methods according to the data pattern but basically they are done through three distinct methods: Common Identifier in the spatial data set, Uniform Resource Identifier (URI) to images and the set of data to the Archaeological notes. All those annotations are done through RDF technology. The technology also allows linking these annotations to the components of ontology in the semantic level.

The next level is the semantic level. Through this level the knowledge generated is managed. It is achieved through the ontological structure setup through the rules defined by the archaeologists. Within this level the domain ontology evolves through each valid rules defined. Archaeologists are involved actively in this phase as they are the one best suited to provide entities and their relationships needed to build up the domain ontology. In order to maintain a common standard among the archaeologists to define the terms used in the ontology, existing standards like standards from CIDOC or other Archaeological standards will be used and extended. However, it should be understood that defining a new standards for archaeology or modifying existing standards are beyond the scope of this project.

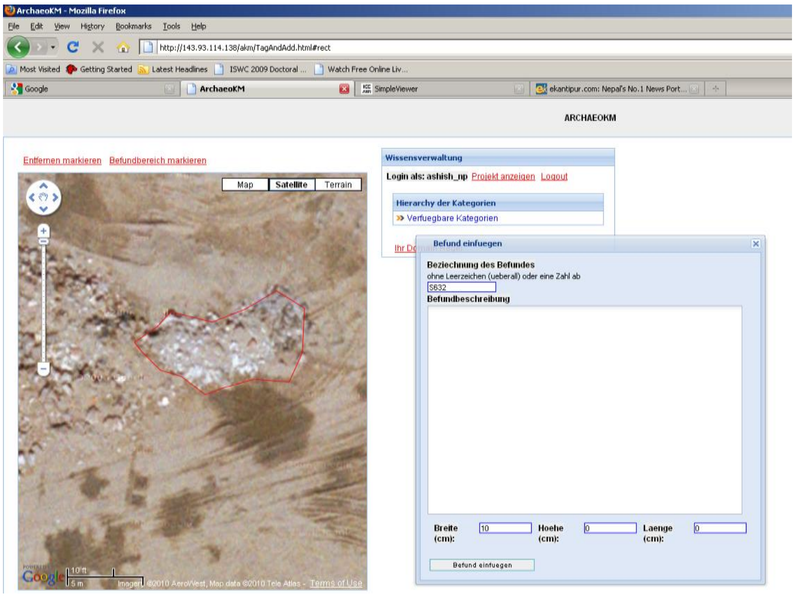
The semantic annotations from the Syntactic level will be indexed semantically to the entities of the domain ontology in this level. This semantic index is the building block of the domain ontology and through semantic annotations provides semantic view of the data. It also provides global schema between various data source making the data integration possible at certain level. This level represents a bridge between interpretative semantics in which users interpret terms and operational semantics in which computers handle symbols (Guarino 1994).

The top most is the most concrete one which represents the organization of the knowledge on the semantic map. This level provides the user interface in form of web pages to display the knowledge generated through semantic map. As could be seen in figure 2, this level has different web pages representing the knowledge. The pages are interrelated and could be navigated according to their relevance. The stand out representation of the knowledge is however through the semantic wiki (Oren, Breslin und Decker 2006). These wiki pages are not only designed to show the knowledge that are generated and managed through the bottom two levels, they are designed to perform semantic queries to derive new knowledge. This will be possible through the interface within the semantic wiki – the semantic wiki will provide a platform through which user can launch their queries and the results will be displayed through the query languages of RDF like SPARQL (Prud'hommeaux und Seaborne 2008) or SWRL (Prud'hommeaux und Seaborne 2008). In this way they will be different from the existing wiki pages. Thus, ArchaeoKM is close to the semantic extension of Wikipedia, but data handling and managing extends beyond textual data. It also handles 3D or 2D object models of the findings besides the textual and image data. It will guide archeologist to define Wikipedia pages concerning subjects and objects of the site that represent knowledge. This level is called the knowledge level because it represents the specification of the knowledge of archeologists concerning the industrial findings.

Besides, the three levels the system architecture contains a component to facilitate the knowledge validation, upgrade and generation. As could be seen in the figure 2.0, it is Spatial Facilitator. This component is responsible for analyzing the spatial data spatially and provides the result either to update the current ontological structure in the semantic level or to create new spatial data. The newly created data themselves could be used to annotate semantically to generate new knowledge. In addition to creating new knowledge in the syntactic level, the spatial analysis on the data can create new entities in the domain ontology in which those semantically annotated data could indexed thus creating a whole set of new knowledge itself. An interface in the Knowledge Level will provide the visual representation of the analysis and could function closely with other components within the level. An example would be creating a buffer within certain feature in the site. This will generate a new set of data (data from buffering the feature) which will be stored in the syntactic level. This buffer could be annotated to generate new knowledge. Likewise, an entity (e.g. *bufferFeature*) will be added in the ontology in semantic level with relationship with other entities and the semantically annotated data (new data after creating the buffer) will be indexed to the entity so to manage the knowledge. In this way this component acts as a facilitator the knowledge handling.

4.2 THE PROCESS

The initial phase of *ArchaeoKM* primarily involves in designing the domain ontology which is basically a descriptive representation of the site represented in a network graph. The process within *ArchaeoKM* can be divided into two broader parts: Knowledge Generation and Knowledge Management. The first part deals with identifying objects in the excavation site and maps the related data and documents to the object. *ArchaeoKM* provides interfaces to support these tasks. As could be seen in figure 3, the objects are identified and tagged with the polygon on the Google map provided within *ArchaeoKM* with proper names. They are mapped to relevant data and documents through the semantic annotation interfaces. This provides a common element for data integration of different data types. In this way the object is first created and populated within domain ontology.


ArchaeoKM - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://143.93.114.138/akm/TagAndAdd.html#rect
Most Visited Getting Started Latest Headlines ISWC 2009 Doctoral ... Watch Free Online Liv...
Google ArchaeoKM SimpleViewer ekantipur.com: Nepal's No.1 News Port...
ARCHAEOKM
Entfernen markieren Befundbereich markieren
Map Satellite Terrain
10 ft
5 m
Imagery ©2010 AeroWest, Map data ©2010 Tele Atlas - Terms of Use
Wissensverwaltung
Projekt anzeigen Logout
Hierarchy der Kategorien
Verfuegbare Kategorien
Befund einfuegen
Befund einfuegen

Login als: ashish_np

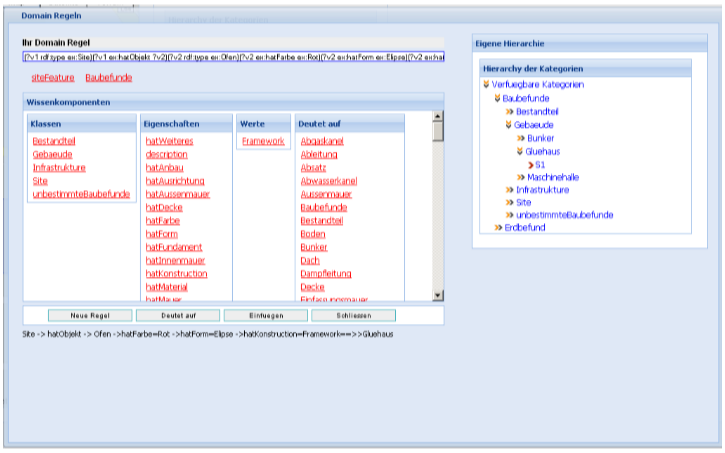
Beziechnung des Befundes
ohne Leerzeichen (ueberall) oder eine Zahl ab
S632

Befundbeschreibung

Breite (cm): 10 Hoehe (cm): 0 Laenge (cm): 0

Fig 3: Identifying and Tagging Object

The second part is basically managing the knowledge generated by identifying the object. *ArchaeoKM* provides two approaches. The first approach is through the interfaces to directly relate the identified objects to the corresponding related objects. This is possible when the archaeologists know exactly how they are related. The second approach is through the domain rules which archaeologists can formulate at any time. An example is provided in figure 4. In this figure, we can observe a rule stating that a site having oven which is red in color and ellipse in shape and have framework as construction type then that site is a *Glüehaus*. This in fact is a very simple and fictitious rule but *ArchaeoKM* can handle more complex and real rules.



(?v1 rdf:type ex:Site)(?v1 ex:hatObjekt ?v2)(?v2 rdf:type ex:Ofen)(?v2 ex:hatFarbe ex:Rot)(?v2 ex:hatForm ex:Elipse)(?v2 ex:ha

Site -> hatObjekt -> Ofen ->hatFarbe=Rot ->hatForm=Elipse ->hatKonstruction=Framework==>>Gluehaus

Fig 4: The Rule interface

5 CONCLUSION

It is apparent that the data of an archeological site can be best managed through the knowledge posses within the archaeologists. *ArchaeoKM* provides an ideal platform to manage this through its knowledge generation, knowledge management and knowledge visualization tools. The prototype is almost finished and will shortly be uploaded for the wider audience. The spatial components are being integrated within the tool and will be available shortly. When complete version of *ArchaeoKM* is uploaded, we believe the tool will be very useful in the archaeological community.

References

Berners-Lee, T, J Hendler, and O Lassila. "The Semantic Web." *Scientific America*, 2001: 34-43.

Cosmas, J, T Itagaki, D Green, M Grabczewski, M Waelkens, and R Degeest. "3D MURALE: A Multimedia System for Archaeology." *Proceeding: Virtual Reality, Archaeology and Cultural Heritage (VAST).* ACM, 2001.

Cruz, C, C Nicolle, and M Neveu. "Use of semantics to manage 3D scenes in web platforms." *Encyclopedia of Multimedia Technology and Networking*, 2004.

Guarino, N. "The Ontological Level." *Philosophy and the congitive sciences*, 1994.

Handshuh, S, and S Staab. "Authoring and annotation of Web pages in CREAM." *11th Int'l World Wide Web Conference.* Hawaii: WWW 2002, 2002.

Horrocks, I., P. f. Schneider, H. Boley, S. Tabelt, B. Grosof, and M. Dean. *SWRL - A Semantic Web Rule Language - Combining OWL and RuleML.* December 21, 2004. http://www.w3.org/Submission/SWRL/ (accessed May 22, 2009).

Kahan, J, M J Koivunen, E Prud'Hommeaux, and R Swick. "Annotea: an open RDF infrastructure for shared web annotations." *International World Wide Web Conference.* Hong Kong: WWW 2001, 2001.

McGuinness, Deborah L., and Frank van Harmelen. *OWL Web Ontology Language.* February 10, 2004. http://www.w3.org/TR/owl-features/ (accessed January 20, 2010).

Oren, E, J Breslin, and S Decker. "Semantic Wikis for Personal Kvnowledge Management." *DEXA Proceeding.* Krakow: Dexa, 2006.

Prud'hommeaux, Eric, and Andy Seaborne. *SPARQL Query Language for RDF.* January 2008, 2008. http://www.w3.org/TR/rdf-sparql-query/ (accessed May 22, 2010).

Russell, B C, A Murphy, and K P Freeman. "LabelMe: A Database and Web-Based Tool for Image Annotation." *International Journal for Computer Vision*, May 2008.

Schroeter, R, J Hunter, and D Kosovic. "Vannotea: A Collaborative Video Indexing, Annotation and Discussion System For Broadband Networks, Knowledge Markup and Semantic Annotation." *K-CAP.* Florida, 2003.

"Semantic annotation for knowledge management: Requirements and a survey of the state of the art." *Journal of the Web Semantics: Sciences and Agents on the World Wide Web 4* (Elsevier) 14 - 16 (2006).

W3C. *Resource Description Framework (RDF).* 02 10, 2004. http://www.w3.org/RDF/ (accessed 06 28, 2010).

Welty, C, and N Ide. "Using the right tools: enhancing retrieval from marked-up documents." *J. Computers and the Humanities*, 1999: 59 - 84.

Wüst, T, Nebiker S., and R. Landolt. "Applying the 3D GIS DILAS to Archaeology and Cultural Heritage Projects - Requirements and First Results." *Applying the 3D GIS DILAS to Archaeology and Cultural Heritage Projects - Requirements and First Results* 34 (2004).