C6

ArchaeoKM: Managing Data through Knowledge in Industrial Archaeological Sites

Ashish Karmacharya
Christophe Cruz
Frank Boochs
Franck Marzani

INTRODUCTION

Today when the world is embracing the advancement of technology the art of data acquisition has also changed a lot. Now it is possible to collect data with very high accuracy. This has provided innumerable 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 in 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 archaeological 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 archaeologists are able to perform these tasks through their knowledge of the excavation sites and the objects excavated.

Industrial archaeology 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 of complex and heterogeneous data like 3D point clouds, pictures, images, videos, notes and others. Second – the most feasible process to allow archaeologists to annotate, index, search, and retrieve

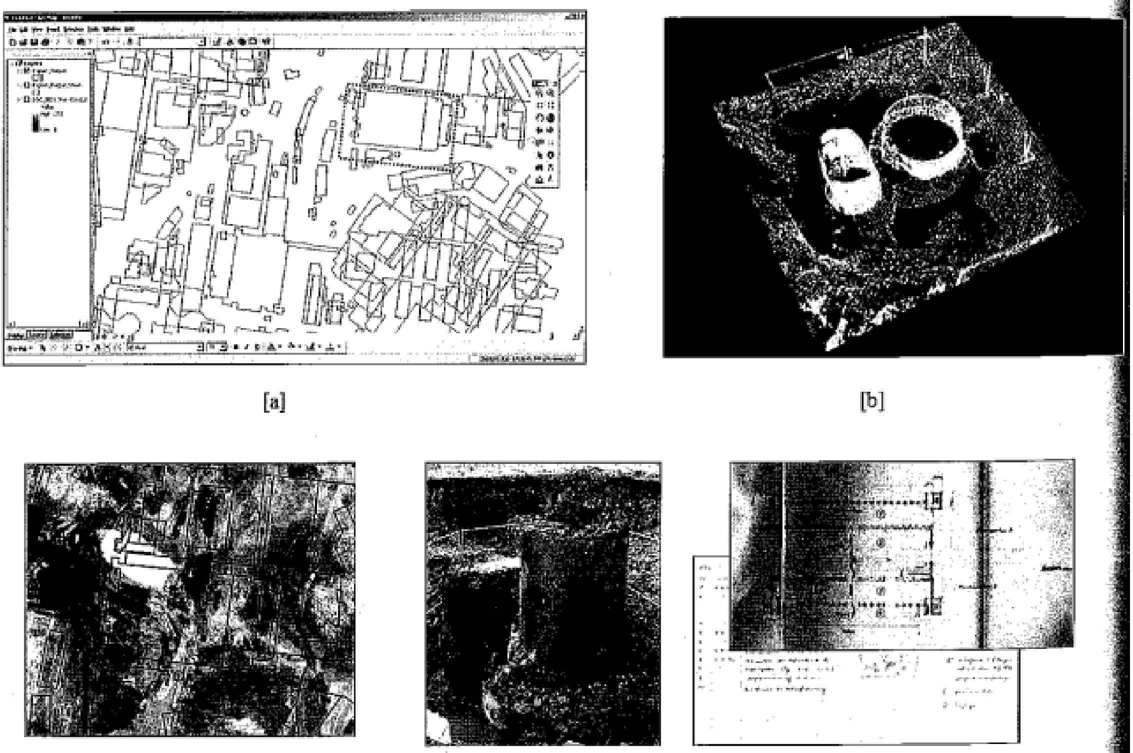

Figure 1: Heterogeneity nature of data [a] Site Plan laid out as GIS data in ArcGIS (highlighted the area of Oven) [b] Orthophoto from aerial image overlaid with the Site Plan (Oven area). [c] Point Cloud of Oven [d] Image of the oven. [e] (top) Floor Plan (down) Archaeological notes (© author)

data and documents in order to ease the identification of common archaeological 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 archaeologists 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.

DATA PATTERNS AND FORMATS

Industrial Archaeology is perhaps the 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 available 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. 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.

The nature of the dataset generated during the project is heterogeneous. It can 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 degrees and are in Gauss Krüger coordinate system (GK II). It is the main data set used for the 3D object modelling. 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 contain 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 overlaid with the orthophoto in order to identify the interesting locations easily as can be seen in Figure 1 (b).

ARCHAEOKM – THE PRINCIPLE AND THE PROCESS

The primary principle within *ArchaeoKM* is the use of semantic web and knowledge management to facilitate archaeologists with 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 can 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.

The Architecture

As can be seen in Figure 2, *ArchaeoKM* is a three layered architecture with the bottom layer being the Syntactic layer. Within this layer all the data and documents collected during excavation is stored in their proprietary format. The middle layer is the Semantic layer. Within this layer, the description of the excavation site is represented

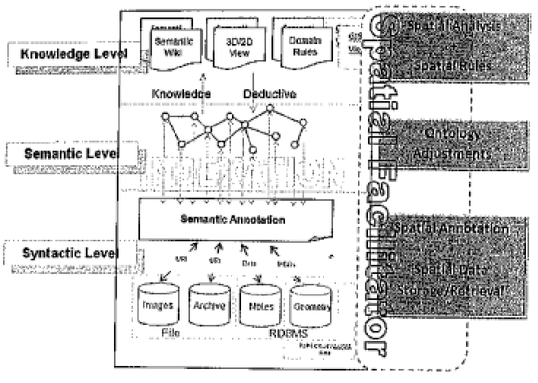

Figure 2: System Architecture (© author)

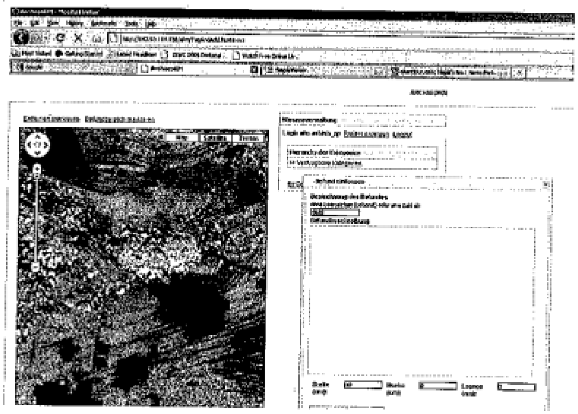

Figure 3: Identifying and Tagging Object (© author)

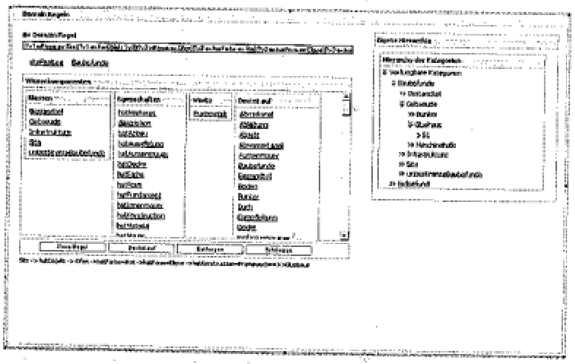

Figure 4: The Rule interface (© author)

in domain ontology. This can be considered as the core of *ArchaeoKM* as it controls all the knowledge generation and management process. The topmost level is the Knowledge level. This level is the face of the application and so consists of different user interfaces to visualize and interact with the knowledge. As can be seen, a parallel facilitator supports all the levels to store spatial data, retrieve and perform spatial analysis and visualize the results within all levels of *ArchaeoKM*. Details are given in A. Karmacharya, et al. 2009, A. Karmacharya, C. Cruz, et al., *ArchaeoKM*: Managing Archaeological data through Archaeological Knowledge 2010.

The Process

The initial phase of *ArchaeoKM* primarily involves 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 can 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.

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 an oven which is red in colour and elliptical in shape and has a framework as construction type then that site is a Glühhaus. This in fact is a very simple and fictitious rule but *ArchaeoKM* can handle more complex and real rules.

CONCLUSION

We have presented *ArchaeoKM* based on knowledge management which uses the knowledge possessed by archaeologists to manage the archaeological data. We are currently testing a prototype with real data and real archaeological knowledge. The architecture was prototyped using JENA on PostgreSQL. To implement the framework, we are using JENA (Semantic Web Framework for Java) in order to build and to manage ontology in JAVA. JENA helps us to handle an OWL database. We use the request language of JENA to retrieve data. Possibilities of integrating the reasoning capability of OWL DL (Web Ontology Language) to generate new knowledge through the existing one are being explored.

REFERENCES

Bechhofer, S., et al (2004): OWL Web Ontology Language. W3C Recommendation. February 10, 2004. URL: http://www.w3.org/TR/owl-ref/ [accessed November 27, 2009].

Cruz, C. et al (2010): Integration of Spatial Technologies and Semantic Web Technologies for Industrial Archaeology. WEBIST. Velencia: WEBIST.

Group RDF Working. Resource Description Framework (RDF). February 10, 2004. URL: http://www.w3.org/2001/sw/wiki/RDF [accessed May 22, 2010].

Horrocks, I. et al (2004): SWRL – A Semantic Web Rule Language – Combining OWL and RuleML. December 21, 2004. URL: http://www.w3.org/Submission/SWRL/ [accessed May 22, 2009].

Karmacharya, A. et al (2009): *ArchaeoKM.* Toward a better archaeological dataset management. CAA. Williamsburg.

Karmacharya, A. et al (2010): *ArchaeoKM.* Managing Archaeological data through Archaeological Knowledge." CAA2010. Granada: Extended Abstract CAA2010. Support of Spatial Analysis through a Knowledgebase – A new concept to exploit information shown for Industrial Archaeology. Interanational Cartography Conference. Santiago: ICC2009. 2009.

Prud'hommeaux, E.; Seaborne, A. (2008): SPARQL Query Language for RDF. January 2008. URL: http://www.w3.org/TR/rdf-sparql-query/ [accessed May 22, 2010].