Managing Knowledge for Spatial Data – A Case Study with Industrial Archaeological Findings

Ashish Karmacharya[1,2], Christophe Cruz[2], Frank Boochs[1], Frank 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. Shifting from conventional approaches to an unusual approach in industrial archaeology, we suggest the use of a web platform based on semantic web technologies and knowledge management. This platform is used to store data during the excavation process and to manage knowledge acquired during the identification process of the findings. The principle of our approach consists in using semantic annotations in order to have a semantic view on data sets. The shared ontology that defines an index on the semantic annotations allows us to build a global schema between data sources. This global schema allows annotating, indexing, searching and retrieving data and documents.

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

1 Introduction

Along with rapid growth in state of art technologies, the approach of data acquisition has changed dramatically in the last few years. This has enabled to collect data with very high accuracy increasing the data with the exponential growth. Additionally, with every new device, a new data format is generated. This in one hand helped in analyzing data more accurately but in next has become very problematic to manage, exchange, share and retrieve. The problem is even more visible in archaeological projects where the amount and pattern of data is huge and heterogeneous. 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. Actually, only archeologists are able to identify which objects are important to be recorded.

Today, as different technologies are being used during excavation, different pattern of data are generated. Primary source of data in any excavation site is the set of point clouds obtained through the terrestrial laser scanning process. They are generally used for creating 3D object models. Besides, floor plans, images and other data like archaeological notes are collected during the project. They provide great value in analysis of the archaeological findings in any project.

As Industrial archeology generates huge amount of data in a very short duration the collected data is 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. Two 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 GIS databases. 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.

Shifting from conventional methods, we suggest the use of a web platform based on semantic web technologies and knowledge management. This platform is used to store data during the excavation process and to manage knowledge acquired during the identification process. The platform facilitates the collaborative process between archeologists to generate knowledge from the data set. The main principle of our approach is to use semantic annotation to provide a semantic view on data sets. The shared ontology that defines an index on the semantic annotations allows us to build a global schema between data sources. This global schema allows us to annotate, index, search and retrieve data and documents.

Section 2 presents the previous and existing works in similar areas and presents the notion of knowledge management through the Web Semantic. Section 3 presents data patterns and the collecting process. Section 4 presents the principle and the method of the platform. The last section concludes the paper.

2 Data and Knowledge Management

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

2.1 Previous and Existing Works

Spatial data is any data that represent objects spatially which can be in 0, 1, 2 and 3 dimensional space. There has been huge amount of work in spatial representation of data from excavation of an archaeological project or objects in cultural heritage sites. However, they are mostly driven by 3D object modeling or Geographic Information System and thus limiting an overall complete Information System. Today, there are handful of projects like the project 3D MURALE [1] and DILAS [2] that can be considered as comprehensive Information System in the field of Archaeology and Cultural Heritage. 3D MURALE system is composed of a recording component, a reconstruction component, a visualization component and database components. The findings are managed through a database management system. Once the findings are stored in the database with a proper data structure, the objects are reconstructed through the reconstruction component. This is done by modeling the objects in 3D space. These 3D models are displayed in the visualization component. 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) [3] as well as texture information.

As all fully oriented geometry management systems, the main issue of these projects is the lack of semantic information. Actually, semantic information allows the management of knowledge on geometrical objects. An interesting approach on how to represent an object through the semantic information in a 3D scene has been discussed in [4]. The use of spatial and orientation relationships between objects can represent the objects in an adequate manner with respect to its surrounding. In addition, the idea concerning semantic relationships between objects is a real improvement for our objectives.

2.2 Knowledge management

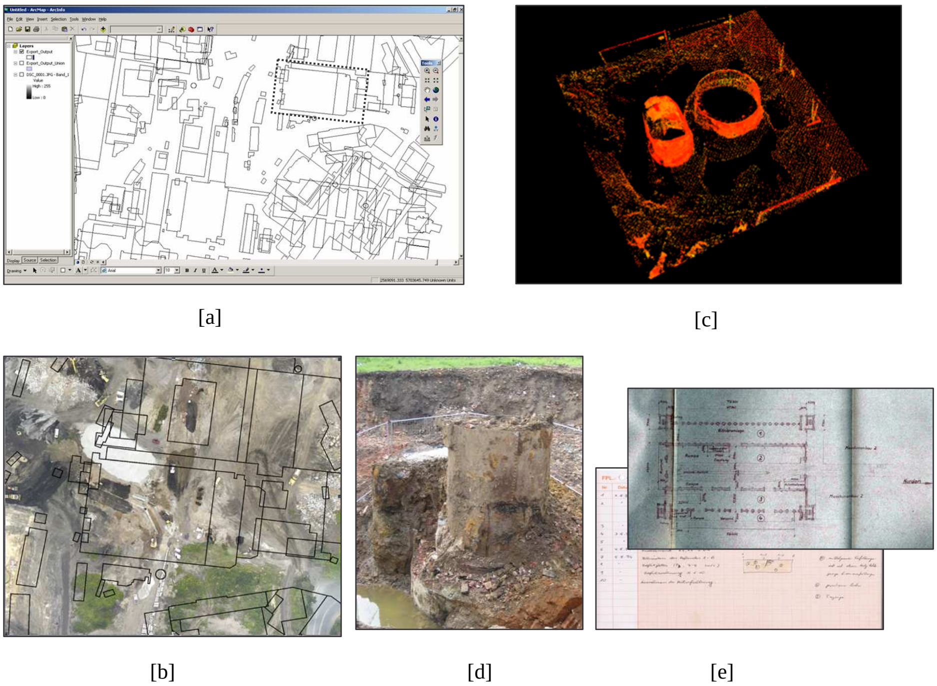
Knowledge about documents has traditionally been managed through the use of metadata. The Web semantic proposes to annotate the document content using semantic information from domain ontologies [5]. The result is a set of Web pages interpretable by machine with the help of mark-ups. The goal is to create annotations (manually or automatically) with well-defined semantics. In the Semantic Web context, the content of a document can be described and annotated using RDF [6] and OWL [7].

Semantic Web annotation brings benefits of two kinds to this platform - enhanced information retrieval and improved interoperability. Information retrieval is improved by the ability to perform searches, which exploit the ontology in order to make inferences about data from heterogeneous resources [8].

Our platform aims at not only managing the concepts defined to annotate documents (which most of the research projects currently focusing on), but also the instances of concepts with their own property values. In this manner, an object found in a point cloud can be linked, with the help of an instance in the ontology to other documents that contain the same object. The second aim of our platform is to give archaeologists the possibility to manage Wikipedia pages on findings. These Wikipedia pages represent the knowledge formalized by archaeologists and are managed through a 3D scene where 3D objects are linked to Wikipedia pages.

3 Data Patterns and Formats

The case study site is the 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.


[a] [c]

[b] [d] [e]

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 Principle and method

This section deals with the methodology that we are proposing. The section begins with the description dealing with the scenario and continues with the details on the system architecture in the second section.

4.1 Case study scenario

The approach to generate and manage knowledge through the semantic annotation of the excavation data can be achieved through the three distinct steps. The first one considers the semantic annotation of the excavation data in order to identify industrial findings in the documents. The second step consists in creating a relationship between documents concerning the same industrial finding. The third step consists in managing semantic objects in order to manage the knowledge with the help of Wikipedia pages.

First of all it is necessary to consider the storing structure of the repository and the services that will be available to store and search data on the various data sets. Geometric and semantic relationships between various objects should be taken into account for efficient management of the objects. The simplest approach would be to store the objects with respect to a 2D map through the bounding boxes. The images of those objects taken from different view points are then related to the respective objects' bounding box by referencing them against the map. Similarly, the points of view of those images are referenced to their respective points in the map. The theory is similar to the scanned point clouds. The geometries of the objects are stored in the database and linked them through the bounding boxes with the 2D map. A similar process is also applied to other datasets. In this way every datasets are transformed in a common referencing system with 2D map implicitly. Thus, all the datasets are linked through a common referencing system and becomes easier to extract information.

The next step should require archaeologists to annotate the documents indexed in 2D map and identify the common archaeological findings in order to create knowledge. It is very important to involve archaeologists in this step as they are the best person to identify the findings. They are the one who should determine the rules through those annotations to generate the knowledge. These rules between the objects will help to enrich the knowledge base and should be incorporated within the ontology. Thus, the ontology will help to create a relationship between the documents. The ontology and the instances of the ontology classes will be defined by archaeologists. In addition, they will also define industrial objects in relation to the documents indexed in the 2D map.

Lastly, the findings during the excavation should be managed properly with the knowledge discovered in the archaeological site. Ontology plays a major part in achieving it. All the findings are referenced against the 2D map through the bounding boxes as semantic objects in the ontology. This could be roughly termed as semantic mapping and is of great value to the archaeological process to determine different behaviors of object in different scenarios. Additionally, those semantic annotations can be interpreted by the machine to be shared, published, queried or used in more general way.

4.2 Architecture

The architecture of the system consists of three major levels responsible for various functionalities. As could be seen in figure 2, these levels are interdependent with each others.

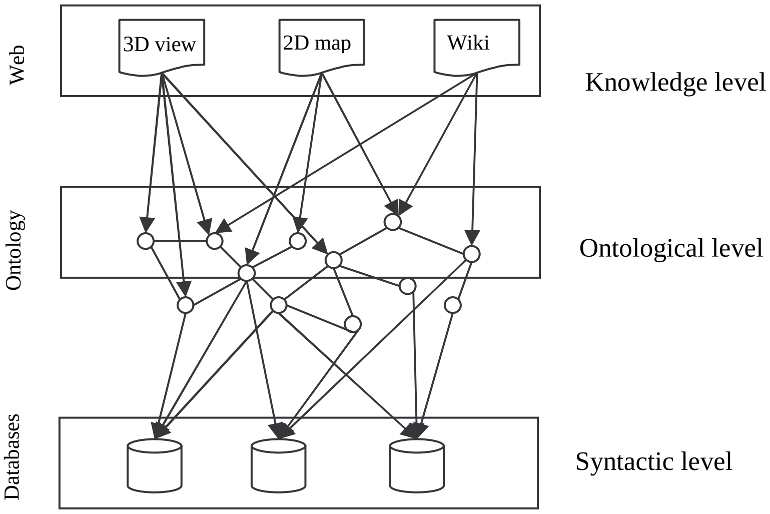


Fig. 2. Data model of the web platform to manage industrial findings.

The bottom level is the Syntactic level which is responsible for storing and retrieving data and documents. This level composes RDBMS (Relational Database Management System). With the inclusion of spatial components in mainstream database systems, it has become very convenient to store geometrical information within these database systems. 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 are stored in the database system with the help of the spatial component. Basically, these geometric data are the set of point clouds with the 3 Dimensional coordinate information. They are the major data in context of the project 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. 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 [9] or Extended Well Known Text [EWKT] as in PostGIS 1.3, the spatial extension of PostgreSQL 8.3 [10]. Similarly, images and documents from achieve are stored in separate or same database within this level. Basically, this level is repository of all the acquirements from the excavation site.

The next level is the Ontological Level. It is probably the most important level with the system architecture because the relationships between different objects are defined within this level. An active involvement of the archeologists is important in this phase as they are the one who can define the relationships of the findings. Through them a set of rules is defined. Through the rules it becomes possible to define the domain ontology for the project. The bridging between this level and the Syntactic level is done through the semantic annotations through which the semantic index is built up. 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 [11].

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 represent different formats of data. The pages are interrelated and could be navigated according to the relevance. In addition, the inter relation of objects within the Ontological level will be shown and could navigate from description of an object to another. Thus, our platform is close to the semantic extension of Wikipedia [12], but data handling and managing extends beyond textual data. The platform also handles 3D or 2D object models of the findings besides the textual and image data. The platform will guide archeologist in order to define Wikipedia pages concerning subjects and objects of the site that represent knowledge added by archeologist. This level is called the knowledge level because it represents the specification of the knowledge of archeologists concerning the industrial findings.

5 Conclusion

We have presented a platform based on knowledge management which is used to handle archaeological data. We are currently prototyping our architecture using JENA on PostgreSQL. The process works on computers in a local network. To implement the framework, we are using JENA (Semantic Web Framework for Java) [13] in order to build and to manage ontologies 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. What was not presented here is the collaborative process between archaeologists. The next issue to resolve is the collaborative work on the ontology which will enable all archaeologists to work on the same Wikipedia page.

References

1. J. Cosmas, T. Itagaki, D Green, E. Grabczewski, M. Waelkens, R. Degeest, et al. "*3D MURALE:A Multimedia System for Archaeology*". Proc. ACM Virtual Reality, Archaeology and Cultural Heritage (VAST 2001). Nov 2001.
2. Wüst T., Nebiker S. Landolt R., "*Applying the 3D GIS DILAS to Archaeology and Cultural Heritage Projects-Requirements and First Results*", Basel University of Applied Sciences, Muttenz, Switzerland.
3. C. Balletti, F. Guerra, A. Adami, "*3D Multiresolution Representations in Archaeological Sites*", CIPA 2005 XX International Symposium,Torino, Italy, 26 September – 01 October, 2005.
4. Cruz, C., Nicolle, C., Neveu, M.: "*Use of semantics to manage 3D scenes in web platforms, Encyclopedia of Multimedia Technology and Networking*", Editor: Margherita Pagani, Idea Group Inc, 2004.
5. Berners-Lee, T., Hendler, J., Lassila, O.: "*The Semantic Web, Scientific American*". (2001), 34-43.
6. Resource Description Framework, Model and Syntax Specification, W3C Recommendation (2004), available at http://www.w3.org/RDF/
7. Web Ontology Language (OWL), http://www.w3.org/2004/OWL/
8. Welty C., Ide N.: "*Using the right tools: enhancing retrieval from marked-up documents*", J. Computers and the Humanities, 33(10)(1999) 59-84, 1999.
9. Oracle, "*Oracle Spatial Developer Guide 11g Release*", Oracle, 2007
10. PostgreSQL, "*PostGIS Manual*", PostgreSQL documentation.
11. Guarino N., The ontological level, in R. Casati B. S. & White G., eds, Philosophy and the cognitive sciences, Hölder-Pichler-Tempsky, 1994.
12. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., and Studer, R.: Semantic Wikipedia. In Proceedings of the 15th international Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM, New York, NY, 585-594.
13. McCarthy, P., 2004. *Introduction to Jena*, [Online] Available: http://www-128.ibm.com/developerworks/java/library/j-jena/