

# Using Sharable Ontology to Retrieve Historical Images

Von-Wun Soo, Chen-Yu Lee, and  
Chao-Chun Yeh

Department of Computer Science  
National Tsing Hua University  
HsinChu, 30043, Taiwan  
soo@cs.nthu.edu.tw

Ching-chieh Chen

Graduate School of Library and Information Science  
Simmons College  
Boston, MA 02115, USA  
chen@simmons.edu

## ABSTRACT

We present a framework of utilizing sharable domain ontology and thesaurus to help the retrieval of historical images of the First Emperor of China's terracotta warriors and horses. Incorporating the sharable domain ontology in RDF and RDF schemas of semantic web and a thesaurus, we implement methods to allow easily annotating images into RDF instances and parsing natural language like queries into the query schema in XML format. We also implement a partial structural matching algorithm to match the query schema with images at the level of semantic schemas. Therefore the historical images can be retrieved by naïve users of domain specific history in terms of natural language like queries.

## Keywords

Sharable ontology, historical image retrieval, multi-agent systems, semantic web, XML, and RDF schemas.

## INTRODUCTION

Traditional full-text information retrieval systems rely on keyword matching techniques and do not really utilize the domain ontology to aid retrieval. The proposal of the semantic web [1] using resource description framework (RDF) and schemas (RDFS) [2] allows domain specific knowledge to be represented in an interoperable and sharable manner. A typical challenge in retrieving historical images in terms of text queries is that the users might not have complete historical knowledge and often cannot specify appropriate keywords for a valid query. For example, a user might not know the exact name of a chancellor in Qin dynasty in ancient Chinese history and it will be very difficult for one to retrieve his image in terms of a query of indirect keywords under traditional keyword matching search engines. However, with the aid of historical ontology, it might become feasible.

## METHODS and PROCEDURES

We design a framework that facilitates the retrieval of annotated images by intelligent software agents using the domain specific ontology and thesaurus [5]. Our approaches, illustrated in Figure 1, include the following steps: 1) design tools to annotate historical images using domain specific thesaurus and sharable ontology, [A snapshot of the domain ontology is shown in Figure 2 and a sample of annotated image is illustrated in Figure 3.] 2) design a query parser to parse natural language queries with domain specific thesaurus and ontology into query schemas, 3) heuristically match the query schemas with the

annotated schemas of images [illustrated in Figure 4] and find out the best fitting image as the result.

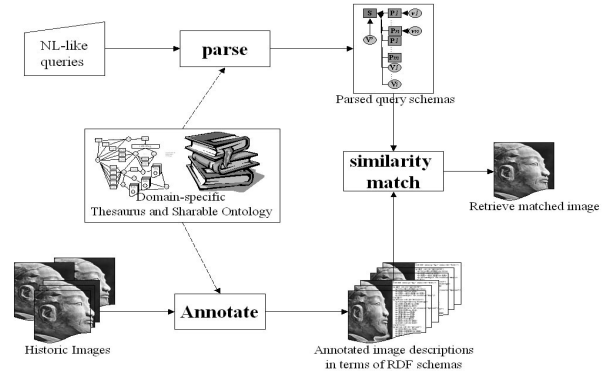


Figure 1. Processes of ontology-based image retrieval

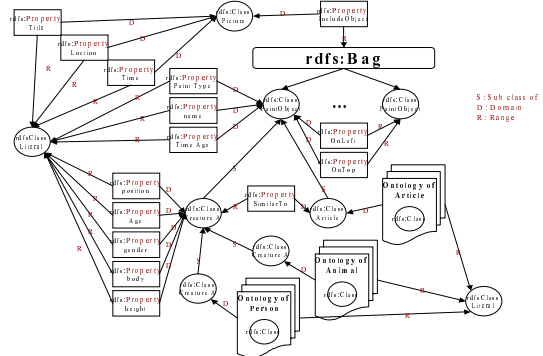


Figure 2. Sharable domain specific ontology for images of the terracotta warriors, horses and related articles

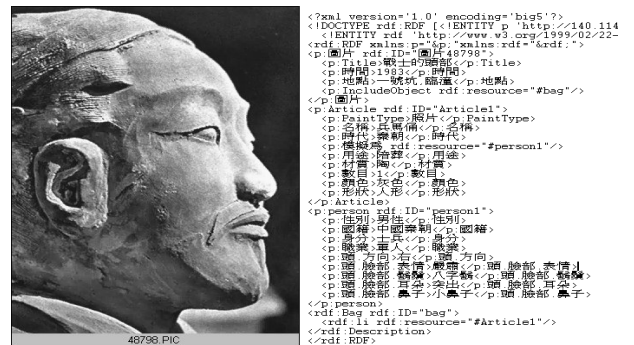


Figure 3. An annotated image of the side view of a Qin terracotta warrior's head

We allow user to specify queries in terms of natural language phrases. For example, a sample query like “秦代

穿著盔甲的將軍” (The general in armor in Qin dynasty) will be parsed and interpreted in terms of the sharable domain specific ontology. The parsed result (after stages of word segmentation, pattern detection, and semantic disambiguation, and etc.) will be shown in the following table that can be easily represented in XML format.

[秦代] (Qin-dynasty)	Value
[穿著] (Wear)	Property
[盔甲] (Armor)	Value
[將軍] (General)	Subject

The XML query schema will then partially match with each image schema in the annotated image database as shown in Figure 4. We designed a heuristic similarity matching algorithm to perform the partial structural match and scoring each partial match. The image that has the best structural fit with the query schema will be retrieved as the result.

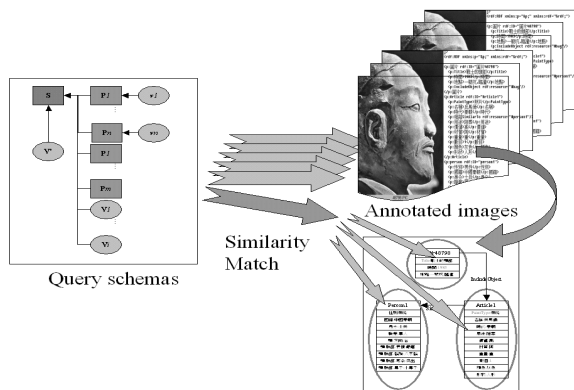


Figure 4. Heuristic structural match of a query schema with annotated images in terms of RDF instances

## DISCUSSIONS

Because we adopt the same sharable domain-specific ontology and thesaurus at the stage of query parsing and reformulation as well as at the stage of annotating the historical images, the problem of historical image retrieval in terms of NL-like text query is transformed into a problem of heuristic semantic and structural match. This allows image retrieval to be conducted at the semantics and domain knowledge level instead of the keyword match level. Unlike [4], the end-users do not have to know about the schemas. This is possible because intelligent information agents conduct the parsing and reformulation of queries by using the sharable domain specific ontology and thesaurus.

## CONCLUSIONS

We designed a query parser that can parse NL-like queries into query schemas in terms of XML format. We implemented tools for annotating the images into RDF instances and a heuristic algorithm to match the XML query with the image RDF instances and find out the best

match images as the result of the retrieval. The sharable domain specific ontology in RDF schemas, we have currently implemented 6 classes and 99 properties among which 4 properties describe the relations between classes. We have also built tools to augment a Chinese thesaurus of general domain with lexical items in historical domain of Qin dynasty. The domain specific lexical items include the names of historical figures, articles, locations, countries, etc. The size of the thesaurus is now more than 70,000 terms that is organized in a semantic hierarchy. We have annotated 50 images selecting from *The First Emperor of China* CD-ROM [3]. We are conducting evaluation experiments on the ontology-based retrieval of 50 annotated images against about 30 sample queries in order to compare with keyword matching retrieval methods. We wish to show empirically that the recall and precision of this approach will outperform traditional keyword matching information retrieval engines in retrieving historical images. It is clear that this research is still at the beginning stage because the Tsinghua research team has used only very small number of Emperor's images resources with limited descriptive information. As the current Chinese Memory Net's (CMNet's) large Emperor image database with much more detailed and rich metadata, keyword, and descriptive information becomes available, we will be able to test more fully and further develop our ontology-based image retrieval techniques.

## ACKNOWLEDGMENTS

This research is supported in part by MOE Program for Promoting Academic Excellence of Universities under grant number 89-E-FA04-1-4, and IDLP project of NSC of Taiwan and is in collaboration with the US NSF/IDLP CMNet project led by Prof. Ching-chih Chen of Simmons College of USA under grant number IIS-9905833.

## REFERENCES

1. <http://www.w3.org/2001/sw/>
2. Steffen Staaba and Michael Erdmann, "An Extensible Approach for Modeling Ontologies in RDF(S)," *ECDL Workshop on the Semantic Web*, 2000.
3. Chen, Ching-chih. *The First Emperor of China*, CD-ROM. Published by the Voyager Co., 1991.
4. Boris Motik and Vlado Glavinic, "Enabling Agent Architecture through an RDF Query and Inference Engine," *10th Mediterranean Electro-technical Conference, MeleCon*, 2000.
5. Von-Wun Soo, Jyi-Shane Liu and Chia-Ning Chiang, "Exploiting Contextual Knowledge for Information Retrieval in a Multiagent-based Digital Library Service System," *NIT 2001*, Beijing. *Global Digital Library Development in the New Millennium: Fertile Ground for Distributed Cross-Disciplinary Collaboration*. ed. By Ching-chih Chen. Beijing: Tsinghua University Press, pp. 299-306, 2001.