Using Protégé for Automatic Ontology Instantiation

Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal

Wendy Hall, Paul H. Lewis, Nigel Shadbolt

Intelligence, Agents, Multimedia Group
ECS School, University of Southampton, UK
{ha, sk, dem, mjw, wh, phl, nrs}@ecs.soton.ac.uk

ABSTRACT

This paper gives an overview on the use of Protégé in the Artequakt system, which integrated Protégé with a set of natural language tools to automatically extract knowledge about artists from web documents and instantiate a given ontology. Protégé was also linked to structured templates that generate documents from the knowledge fragments it maintains.

INTRODUCTION

The value of most semantic web services will be dependent to a large extent on the richness and consistency of their underlying ontology instantiations. Ontology instantiation refers to the insertion of information into the Knowledge Base (KB), as described by the ontology (i.e. instance creation). One rich source of knowledge to instantiate ontologies is the web. However, automatically locating and retrieving knowledge from the web can be a daunting task, especially with the current scarcity of semantic annotations. A number of approaches have been taken to speed up this process using a variety of techniques, such as Information Extraction (IE) from text, harvesting information off structured documents, gathering knowledge from existing annotations, and accessing online databases and gazetteers.

The Artequakt [1] system is concerned with automating the extraction of knowledge about the life and work of artists from web documents, instantiating a given ontology with this knowledge, and using it to generate tailored biographies. Protégé [8] was chosen as the ontology and knowledge base (KB) component, which is responsible for the maintenance and supply of the gathered knowledge.

ARTEQUAKT

Artequakt's architecture (Fig. 1) comprises of three key areas. The first concerns the *Knowledge Extraction* tools used to extract factual information from documents and passing it to the knowledge server. The second key area is the *Protégé Server*, which is concerned with information management, storage, and supply. The final area is the *Narrative Generation*, which constructs the biographies. These areas will be briefly described in the following sections.

PROTÉGÉ SERVER

The Artequakt ontology was mainly constructed from selected sections in the CIDOC Conceptual Reference Model (CRM[1]). The CRM ontology is designed to represent artefacts, their production, ownership, location, etc. This ontology was modified for Artequakt and enriched with additional classes and relationships to represent personal information about artists, their family relations, relations with other artists, details of their work, etc. The Artequakt ontology was implemented in Protégé, and currently contains 42 classes and 228 slots.

A Java HTTP server was developed to allow external access to the Protégé KB. Around 15 different types of queries were encoded on this server to enable other tools to retrieve and update the knowledge base. Some of these queries are generic (e.g. instance slot value query), while others are more specific to our application (e.g. get a date description, an artist's personal detail).

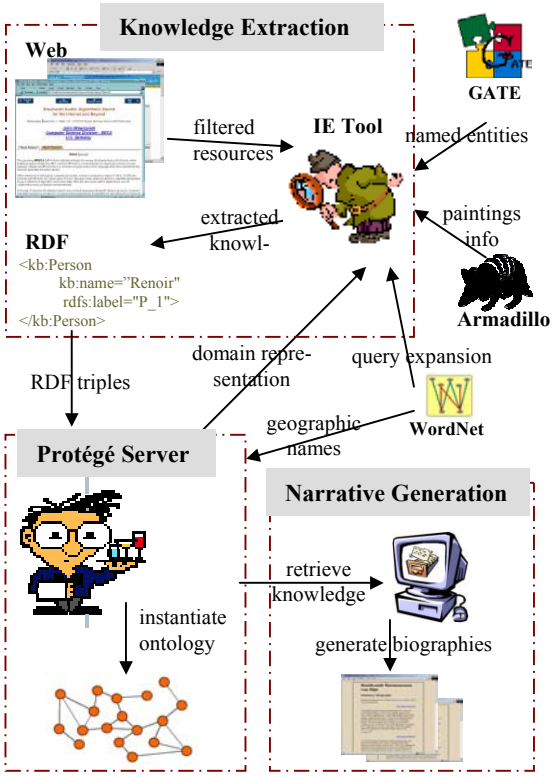


Figure 1. Artequakt's system architecture

[1] http://cidoc.ics.forth.gr/index.html

KNOWLEDGE EXTRACTION

Artequakt's knowledge extraction tool (fully described in [2]) aims to identify and extract knowledge triples from text documents and to provide it as RDF triples for entry into the KB. Artequakt uses an ontology coupled with a general-purpose lexicon (WordNet), an entity-recogniser (GATE, [5]), and a wrapper (Armadillo, [4]) as supporting tools for identifying knowledge fragments.

The extraction process is launched when a user requests a biography for a specific artist that is not in the Protégé KB. The query is passed on to a search engine and the search results are analysed with respect to relevancy to the domain of artists. Selected documents are then analysed syntactically and semantically to identify any relevant knowledge to extract. Below is an example of an extracted sentence:

"Renoir was born in Limoges on February 25, 1841. "

Documents are then passed on to GATE [5] to recognise the type of entities in their text. Annotations provided by GATE for the above sentence highlight that '*Renoir*' is a person's name, '*February 25, 1841*' is a date, and '*Limoges*' is a place. Artequakt will then access the Protégé knowledge server to map these entities to the ontology classes. If a match does not exist, WordNet will be used to expand the given terminology. At the end of this stage, three instances will be created for the Person, Place, and Date classes to represent the three extracted entities above.

After extracting the main entities and inserting them into the KB, the task now is to analyse the sentence further to identify the relations between those entities. Further interaction between the IE tool and Protégé will now be required to name those relations. In the example sentence above, the verb 'born' was expanded in WordNet to 'birth', which matches with two potential relations in the ontology; *date_of_birth* and *place_of_birth*.

By comparing the type of subjects and objects of these relations as expressed in the ontology, with those identified earlier in the sentence, the following knowledge triples will be identified:

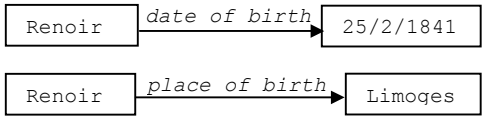


Figure 2. Knowledge triples extracted in Artequakt

Artequakt's IE component is also linked to an Armadillo [4] server (Fig. 1), which was specialised for the retrieval of information about paintings (title, date, dimensions, etc). Armadillo is a wrapper induction tool developed at Sheffield University to discover new linguistic and structural patterns automatically when extracting information from well-structured pages.

Artequakt's IE process terminates by sending all extracted knowledge to the Protégé server in RDF format.

ONTOLOGY INSTANTIATION

When the Protégé server receives a new RDF file, it activates a *feeder* process to parse the given file and add its content to the KB. This will be followed by a consolidation process to find and resolve duplications.

Knowledge Base Consolidation

When acquiring knowledge from multiple sources, it becomes necessary to be able to compare this knowledge and merge any possible duplication to ensure the integrity and consistency of the KB.

Consolidation in Artequakt is concerned with analysing and comparing relational values of instances to identify inconsistencies and duplications (detailed in [3]). This comparison is not always straightforward because relational values are often extracted in different formats and specificity levels (e.g. synonymous place names, different date formats). Artequakt applies a set of heuristics and expansion methods in an attempt to match these values. Consider the following sentences:

1. *Renoir was born in the 19th century in Limoges.*
2. *Renoir was born in 1841 in Limoges, France.*
3. *Renoir was born on Feb 25 1841 in France.*

Matching the above requires some temporal and geographical reasoning. When a new place name is inserted into the KB, the Protégé server will access WordNet (using the JWordNet[2] API) and expand the given place name to its synonyms, subparts, and superparts. This allows the system to relate '*Limoges*' to '*France*'.

Simple temporal reasoning was used to consolidate dates with respect to precision. In our previous example, the third date is used, as it is the most specific.

At the end of the consolidation process, the knowledge extracted from the example sentences above will be merged into the triples given in figure 2.

Figure 3 shows some of the above knowledge, as well as three of Renoir's paintings as found by Armadillo [4].

BIOGRAPHY GENERATION

Once information is extracted, stored and consolidated, the Artequakt system repurposes it by automatically generating biographies of the artists (more detail in [7]). The biographies are based on templates authored in the Fundamental Open Hypermedia Model (FOHM) and stored in the Auld Linky contextual structure server [6].

Each section of the template is instantiated with paragraphs or sentences generated from information held in Protégé. The KB informs the templates of the *theme* of the paragraphs (e.g. influences, family, style) and the generation tool selects the relevant ones and structures them in the desired form and order. Figure 4 is an example output of Artequakt. Very little text generation is used is currently (e.g. 1st and last sentences in fig. 4), but this will be the focus of the next phase.

[2] http://jwn.sourceforge.net/

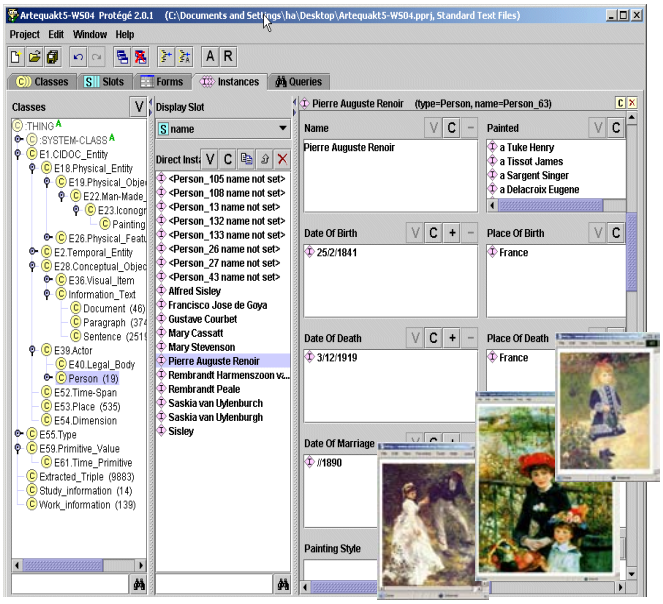
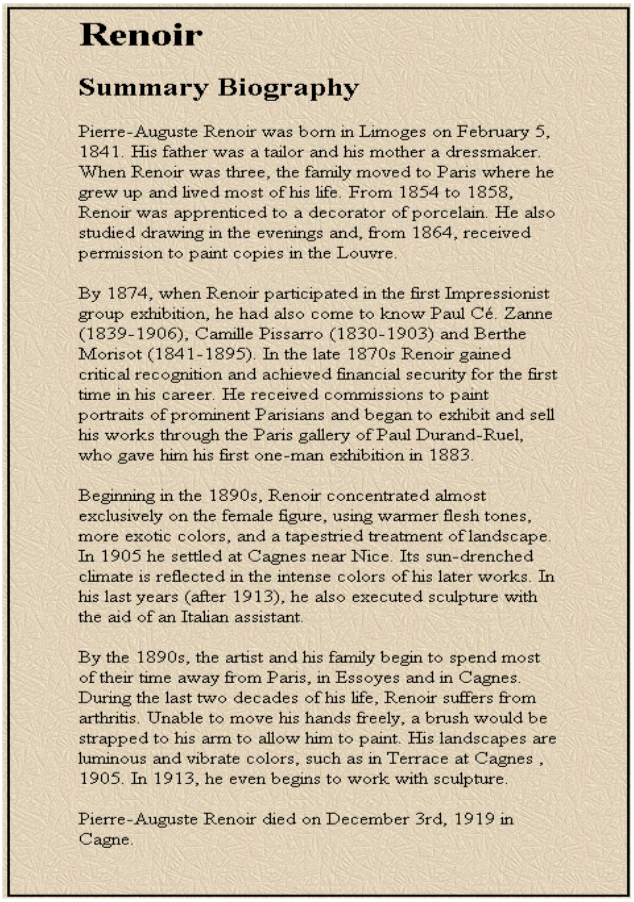


Figure 3. Artequakt ontology with some instantiations.



Renoir

Summary Biography

Pierre-Auguste Renoir was born in Limoges on February 5, 1841. His father was a tailor and his mother a dressmaker. When Renoir was three, the family moved to Paris where he grew up and lived most of his life. From 1854 to 1858, Renoir was apprenticed to a decorator of porcelain. He also studied drawing in the evenings and, from 1864, received permission to paint copies in the Louvre.

By 1874, when Renoir participated in the first Impressionist group exhibition, he had also come to know Paul Cé. Zanne (1839-1906), Camille Pissarro (1830-1903) and Berthe Morisot (1841-1895). In the late 1870s Renoir gained critical recognition and achieved financial security for the first time in his career. He received commissions to paint portraits of prominent Parisians and began to exhibit and sell his works through the Paris gallery of Paul Durand-Ruel, who gave him his first one-man exhibition in 1883.

Beginning in the 1890s, Renoir concentrated almost exclusively on the female figure, using warmer flesh tones, more exotic colors, and a tapestried treatment of landscape. In 1905 he settled at Cagnes near Nice. Its sun-drenched climate is reflected in the intense colors of his later works. In his last years (after 1913), he also executed sculpture with the aid of an Italian assistant.

By the 1890s, the artist and his family begin to spend most of their time away from Paris, in Essoyes and in Cagnes. During the last two decades of his life, Renoir suffers from arthritis. Unable to move his hands freely, a brush would be strapped to his arm to allow him to paint. His landscapes are luminous and vibrate colors, such as in Terrace at Cagnes , 1905. In 1913, he even begins to work with sculpture.

Pierre-Auguste Renoir died on December 3rd, 1919 in Cagne.

Figure 4. Biography for Renoir generated by Artequakt

CONCLUSIONS

This paper described a system that automatically extracts knowledge from the web to instantiate an ontology held in Protégé, then reassembles the knowledge in the form of biographies. This project explored the potential of integrating Protégé with IE tools, narrative generation templates, a lexicon to extend its terminology, and with a variety of programs to manipulate its knowledge.

Using an ontology in this context is aimed at increasing the system's portability to other domains. Building a cross-domain system is one of the aims of this project, and will be fully investigated in the next stage of development.

ACKNOWLEDGEMENTS

This research is funded in part by EU Framework 5 IST project "Sculpteur" IST-2001-35372, EPSRC IRC project "Equator" GR/N15986/01 and EPSRC IRC project "AKT" GR/N15764/01.

REFERENCES

[1] Alani, H., S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis and N. Shadbolt. Automatic Ontology-based Knowledge Extraction and Tailored Biography Generation from the Web. IEEE Intelligent Systems, 18(1): 14-21, 2003.

[2] Alani, H., S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis and N. Shadbolt. Automatic Extraction of Knowledge from Web Documents. Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd Int. Semantic Web Conf. Sanibel Island, Florida, USA, 2003.

[3] Alani, H., S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis and N. Shadbolt. Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation. 2nd Int. Conf. Knowledge Capture (K-Cap'03), Workshop on Knowledge Markup and Semantic Annotation, Sanibel Island, FL, USA, 2003.

[4] Ciravegna, F., S. Chapman, A. Dingli, and Y. Wilks, Learning to Harvest Information for the Semantic Web. Proc. 1st European Semantic Web Symposium, Heraklion, Greece, May 2004

[5] Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan. GATE: a framework and graphical development environment for robust NLP tools and applications. Proc. 40th Anniversary Meeting of the Association for Computational Linguistics, Phil, USA, 2002.

[6] Michaelides, D.T., D.E. Millard, M.J. Weal, D. De Roure. Auld Leaky: A Contextual Open Hypermedia Link Server. Proc. 7th Hypermedia: Openness, Structural Awareness, and Adaptivity, Heidelberg, 2001.

[7] Millard, D.E., H. Alani, S. Kim, M.J. Weal, P.H. Lewis, W. Hall, and N. Shadbolt. Generating Adaptive Hypertext Content from the Semantic Web. 1st Int. Workshop on Hypermedia and the Semantic Web, HyperText'03, Nottingham, UK. 2003.

[8] Musen, M. A., R. W. Fergerson, W. E. Grosso, N. F. Noy, M. Y. Grubezy and J. H. Gennari. Component-based support for building knowledge-acquisition systems. Intelligent Info. Processing Conf. (IIP), pp 18-22, Beijing, China, 2000.