

# Automatic Ontology-Based Knowledge Extraction from Web Documents

Stefan Bischof, 0327033

SS2005

This report resumes the paper “*Automatic Ontology-Based Knowledge Extraction from Web Documents*” [AKM<sup>+</sup>03]. So this should be a short critical text, highlighting the main ideas, concepts and components of the referenced paper and some thoughts of the author about it.

## 1 Initial situation

When you search the web nowadays for broad information of a specific topic you often have to visit several pages and then combine the information found there to have it in a usable form.

The examples used in this project are biographies of impressionist artists. If you want extensive information about a specific artist you probably won't find it on one single page.

So the aim of this project is providing a website which is able to **generate biographies dependent on users needs**.

## 2 Project structure

The described system, *ArtEquAKT*, automatically extracts knowledge about artists from the Web, populates a knowledge base, and uses it to generate personalized biographies.

The system combines the following projects (the name, *ArtEquAKT*, is composed by these projects):

- **Artiste:** Database of art images <http://www.artisteweb.org/>
- **The Equator IRC:** Interdisciplinary Research Collaboration (IRC); uses narrative techniques <http://www.equator.ac.uk/>
- **The AKT IRC:** Advanced Knowledge Technologies; examine knowledge life cycle <http://www.aktors.org/>

The system uses a few tools and programs:

- **Protégé:** used for creating the ontology; also stores the knowledge base (KB) <http://protege.stanford.edu/>
- Knowledge Extraction
  - **Apple Pie Parser:** groups grammatically related phrases <http://nlp.cs.nyu.edu/app/>
  - **WordNet:** a general purpose lexical database <http://wordnet.princeton.edu/>
  - **GATE:** General Architecture for Text Engineering <http://www.gate.ac.uk/>
- **Auld Linky:** link server <http://www.equator.ecs.soton.ac.uk/technology/linky/index.shtml>

The whole system, or process, can be divided in three main parts, with its own problems and specific solutions:

- **Knowledge extraction**
- **Information management**
- **Narrative generation**

## 2.1 Knowledge extraction

The most interesting, complex and largest part is the knowledge extraction subsystem thus I want to amplify this component somewhat.

After the web page is retrieved, the Apple Pie Parser syntactically and semantically analyzes the text. The result is passed to **GATE** which categorizes the information (e.g. *Rembrandt* is a person, *July 15, 1606* is a date, *Netherlands* is a location). GATE (General Architecture for Text Engineering) provides a specification of architecture, an implementation of this architecture and a graphical development environment. GATE itself provides only this and no specific tools.

In the next step **WordNet** combines the information of GATE with the ArtEquAKT ontology. WordNet is a combination of a dictionary and a thesaurus. It groups English words in synonym sets (called synsets). Every synset is connected with other synsets via different types of relations (e.g. for nouns: synonyms, hypernyms, hyponyms, holonym, meronym and coordinate terms). So WordNet finds out that *Leiden* is a city, and *Netherlands* are a country.

## 2.2 Information management

The ontology is populated automated with the extracted information. This means that the information is inserted in the KB following the ontology domain representation.

Frequently used data (for biography generation) is stored in a RDBMS. This step only provides some speed and (probably) a standardized interface for the narrative generation.

## 2.3 Narrative generation

In the last step, the system generates a personalized biography of the specified artist. The story is created from the KB and a template based mechanism (Auld Linky) and some servlets.

## 3 Conclusion

When I started reading I was confused by the several projects the paper links to and depends on. I had the impression the system was just plugged together like lego. What I missed most was a definition or an explanation of what ontologies are and concrete examples of its implementation (like you can find it in the paper of my colleague).

Remarkable is also the fact that numbers of success or failure of the system are completely lacking.

In general I think this could be a good concept, although it's very complex and therefore error-prone. Problems are mainly in knowledge extraction, automated ontology population and automatic narrative generation. A working demo webpage was also missing.

## 4 Reference

- [AKM<sup>+</sup>03] Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, Nigel R. Shadbolt. *Automatic Ontology-Based Extraction from Web Documents*. IEEE Intelligent Systems, Volume 18, Issue 1, January 2003.