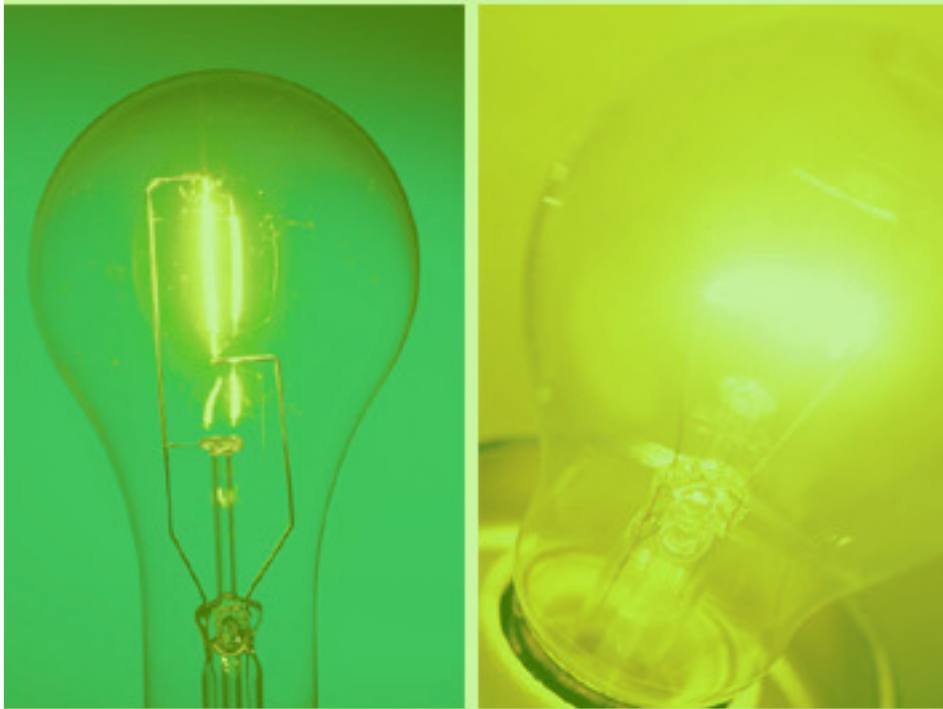


Automatic Ontology- Based Knowledge Extraction from Web Documents



Stefan Bischof
PSWIE SS 2005



Overview

The Big Picture

Introduction

Architecture

Knowledge Extraction

WordNet

GATE

References



The Big Picture

Problem: You want to have several information about an artist (e.g. Rembrandt)

Traditional Solution: Normally you search the web with any search engine, sort out the misses and collect the desired information from the remaining strewn pages.

This is a very time consuming and unamusing way to solve this problem.

ArtEquAKT: Ontologybased KE method



Introduction (1)

The **ArtEquAKT** system uses three separate projects

- **Artiste**
- The **Equator** IRC
- The **AKT** IRC

IRC means Interdisciplinary Research Centre



*art***ISTE**

EQUATOR ÷

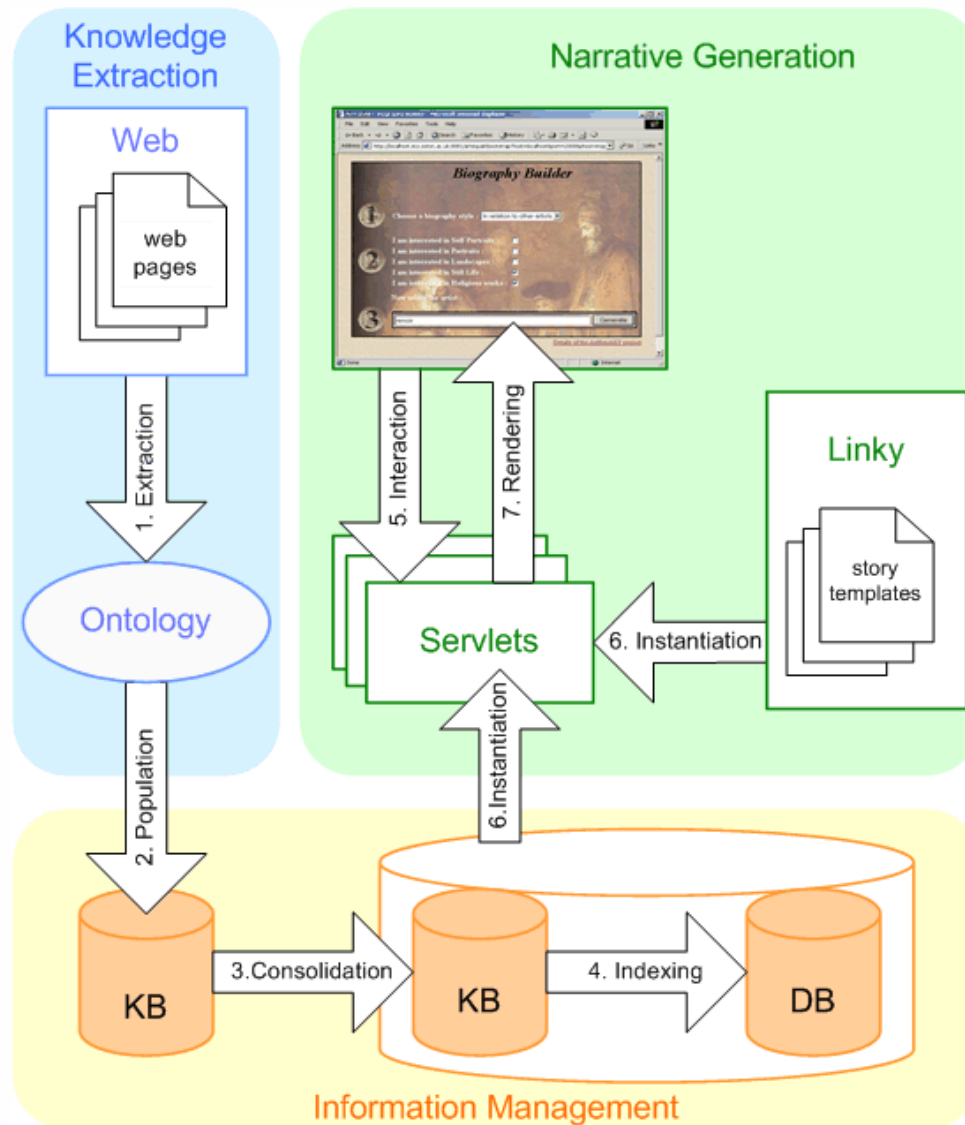
ADVANCED KNOWLEDGE
AKT
TECHNOLOGIES



ArtEquAKT Solution

- Use NLT to extract information about the life and work of artists from online documents
- Feed this information automatically to an ontology
- Ontology was created before for this domain
- Generate stories from the knowledge base

Architecture



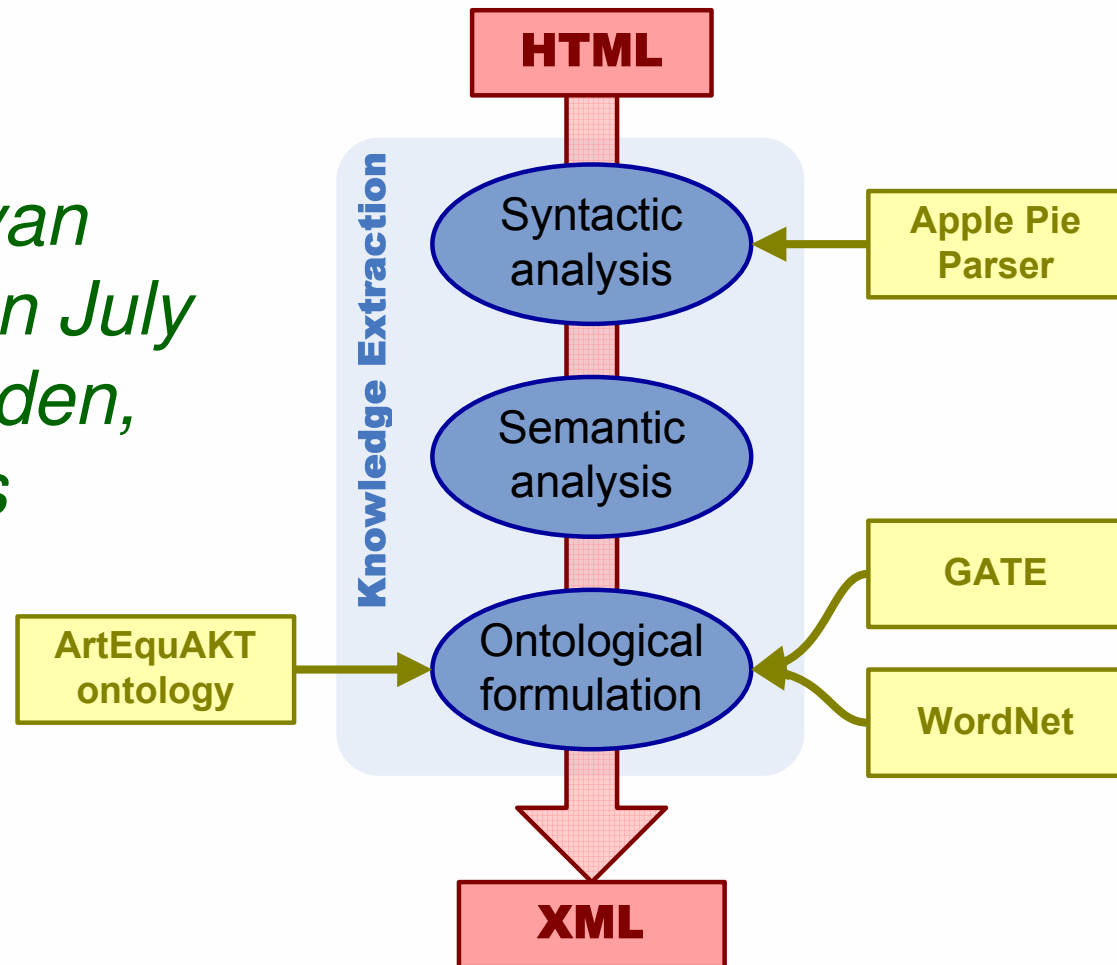
Most complex tasks:

- **Knowledge extraction**
- Automatic ontology population
- Narrative generation

Knowledge Extraction

HTML

*Rembrandt
Harmenszoon van
Rijn was born on July
15, 1606, in Leiden,
the Netherlands*



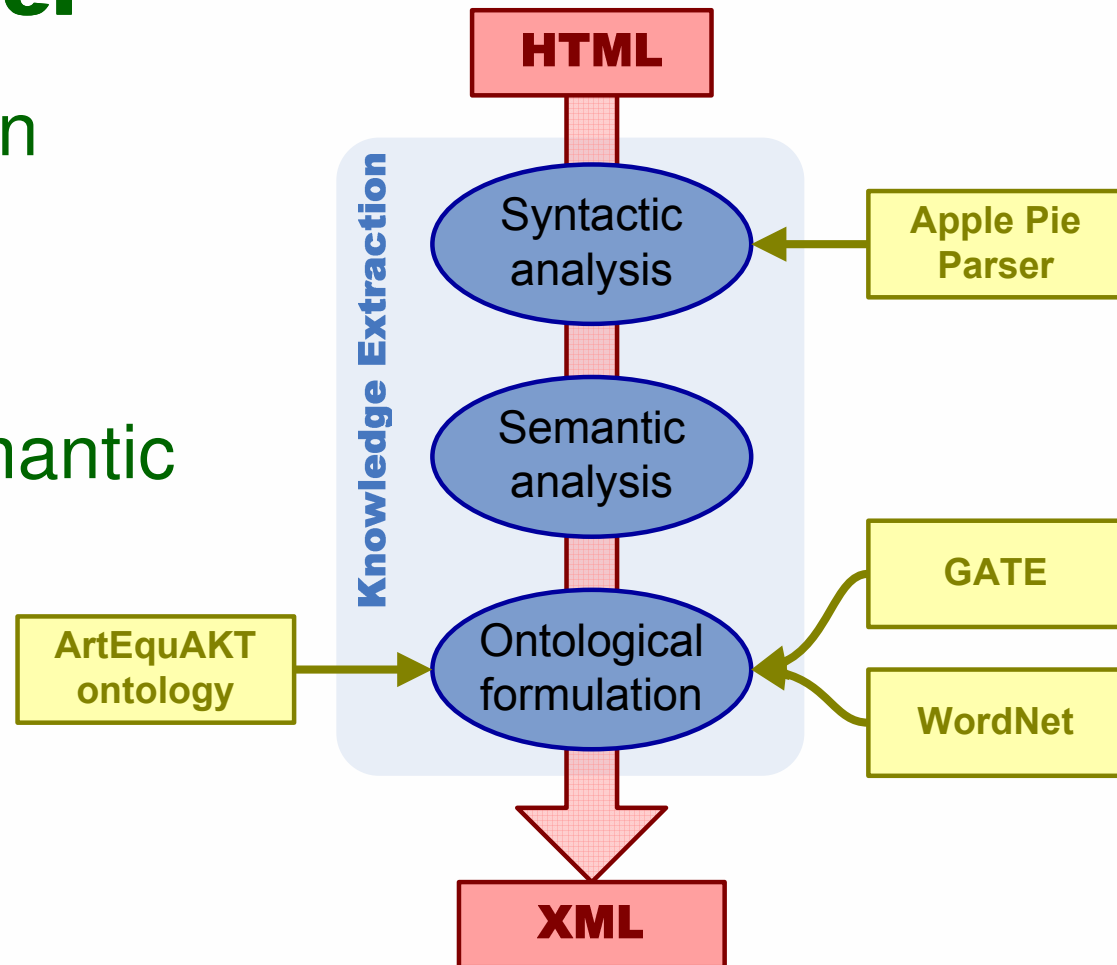
Knowledge Extraction

Apple Pie Parser

Rembrandt ⇒ Noun

Born ⇒ Verb

Syntactic and semantic analysis



Knowledge Extraction - GATE

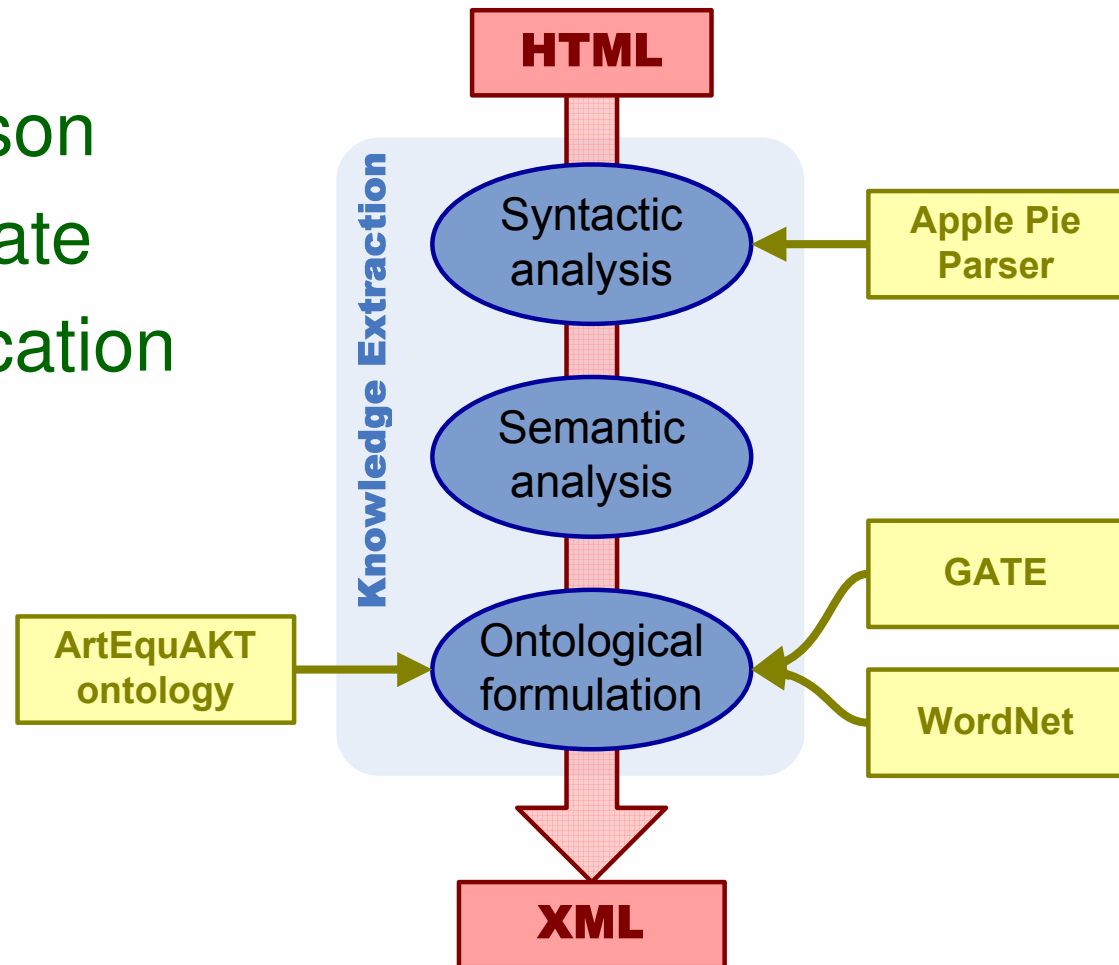
GATE

Rembrandt ⇒ Person

July 15, 1606 ⇒ Date

Netherlands ⇒ Location

Categorizes
Information





GATE

A **G**eneral **A**rchitecture for **T**ext **E**ngineering

- Identifies knowledge fragments - entities and relations between them
- Infrastructure for developing and deploying software components that process human language





GATE - Aims

- Architecture for language processing software
- Provide implementation of the architecture
- Provide graphical development environment



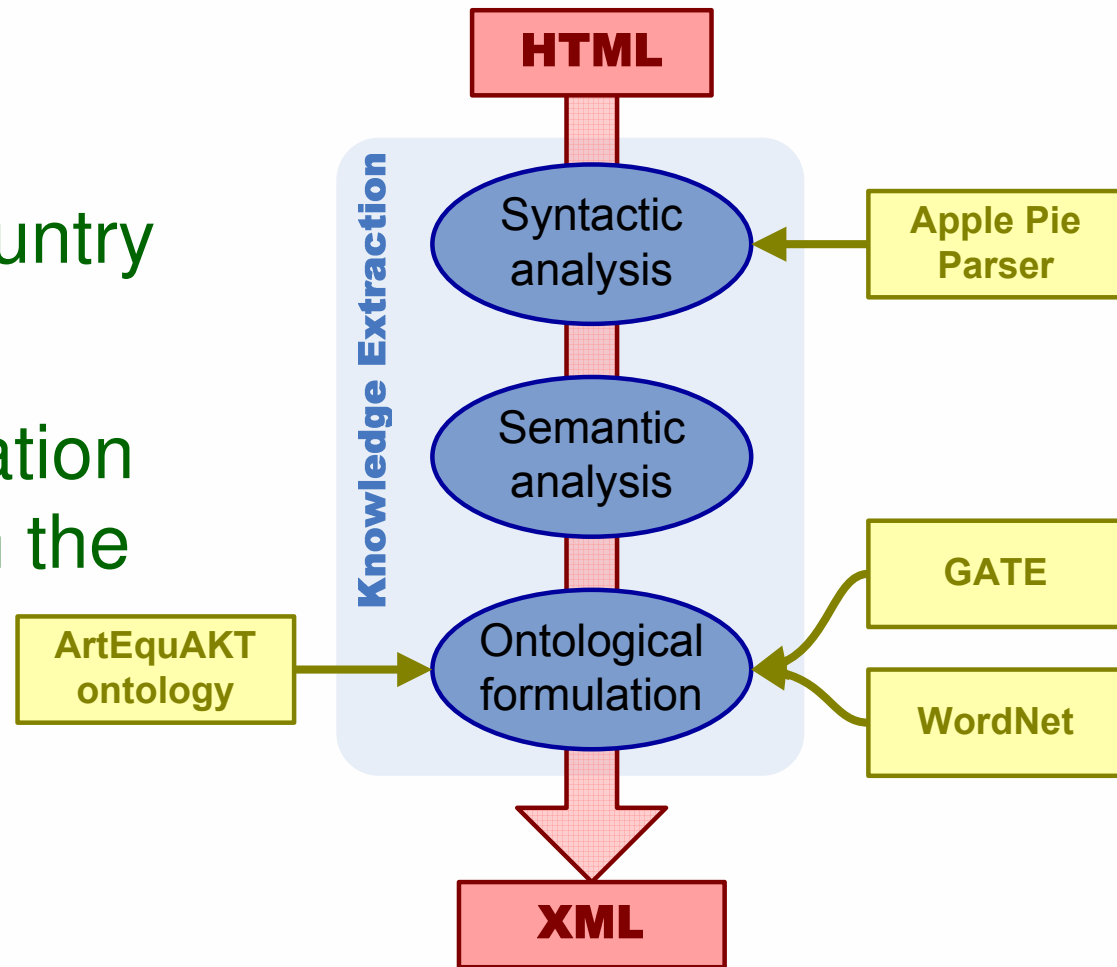
Knowledge Extraction - WordNet

WordNet

Leiden ⇒ City

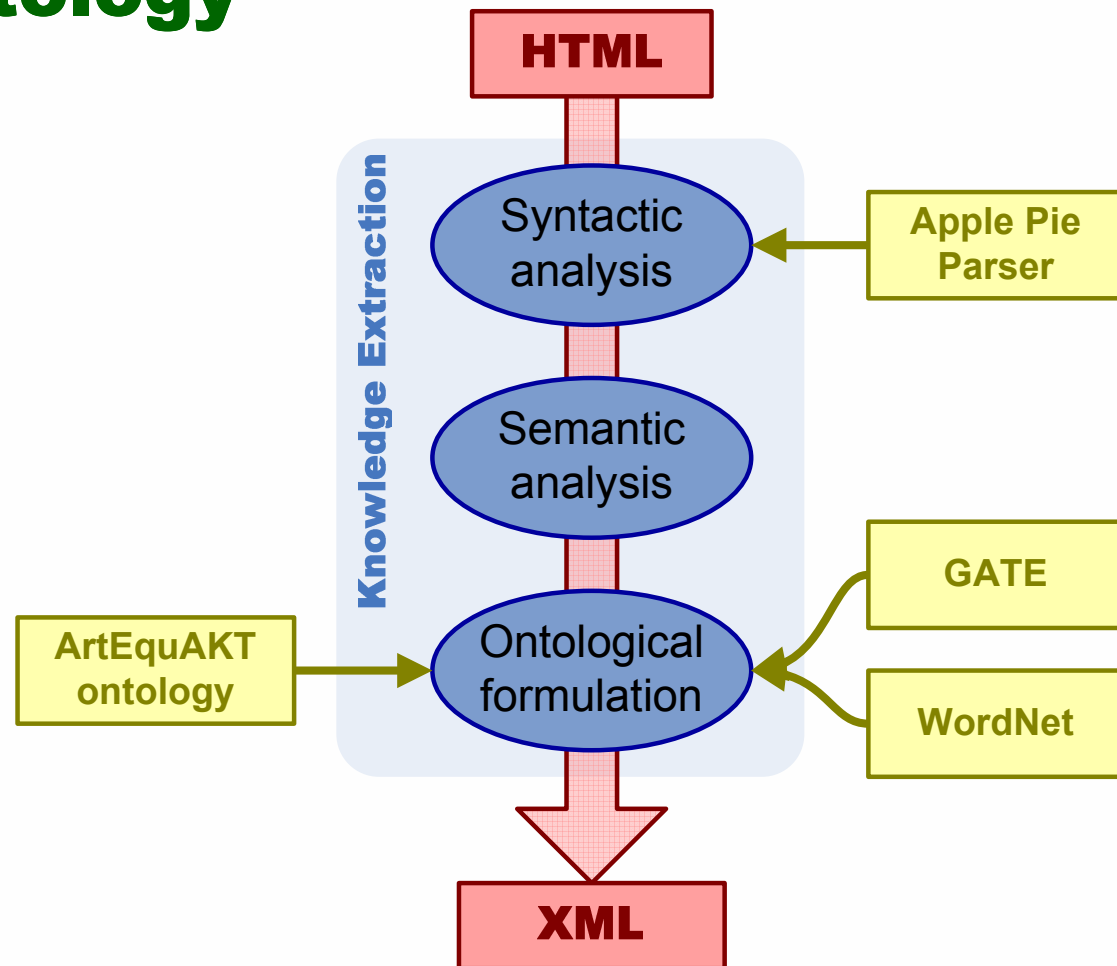
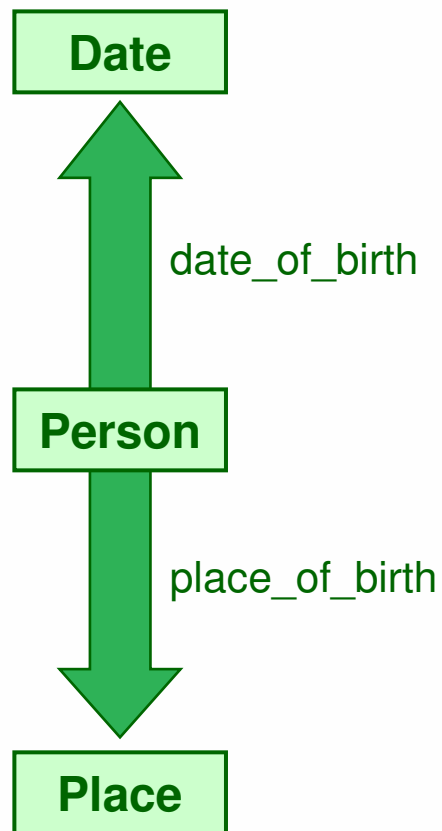
Netherlands ⇒ Country

Combines information
from GATE with the
ArtEquAKT
ontology



Knowledge Extraction – ArtEquAKT ontology

ArtEquAKT ontology

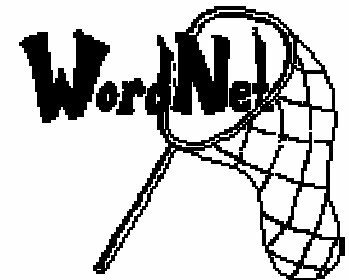




WordNet

A lexical database (semantic lexicon) for the English language

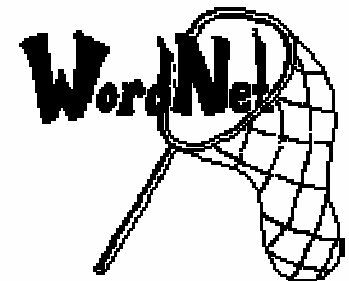
- groups English words, nouns, verbs, adjectives and adverbs, into sets of synonyms called *synsets*
- provides short definitions
- records the various semantic relations between the synonym sets





WordNet - Aims

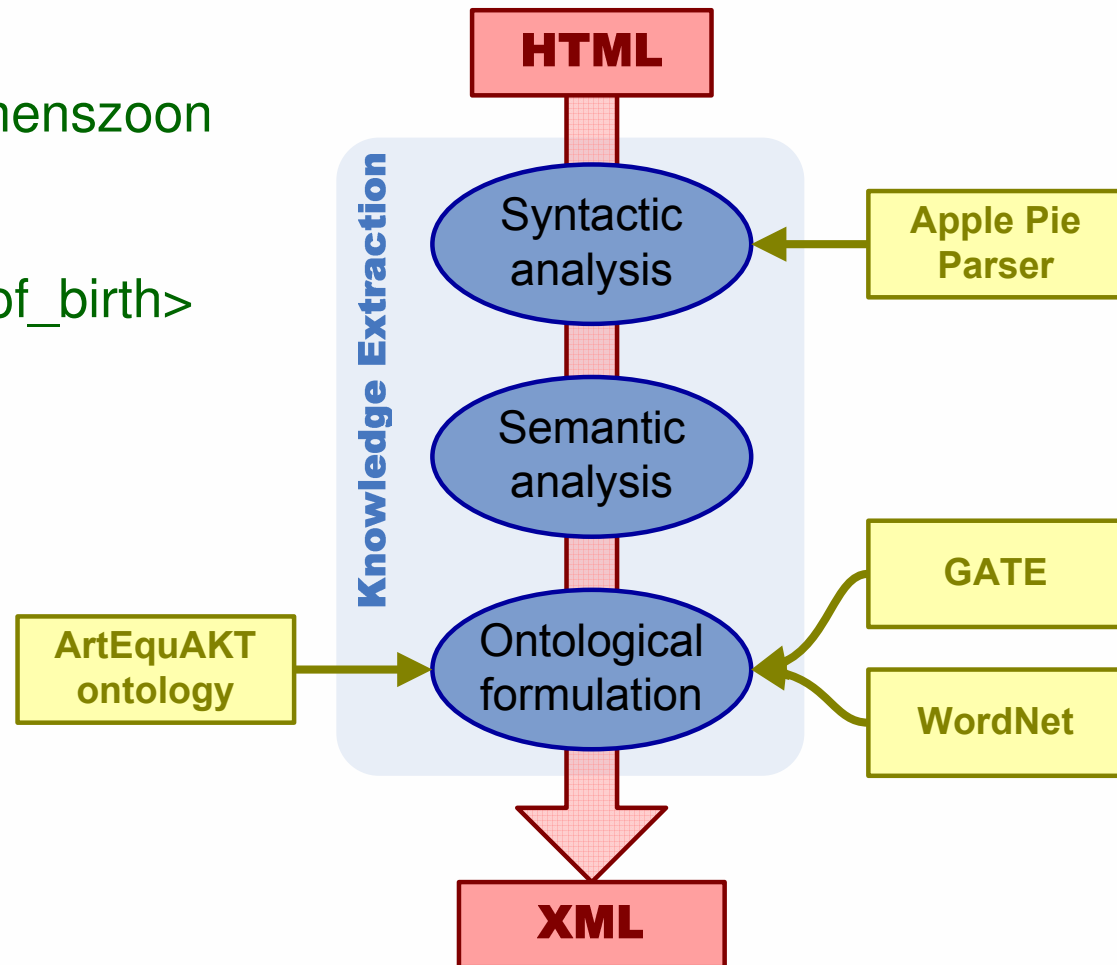
- combination of dictionary and thesaurus
- support automatic text analysis and artificial intelligence applications



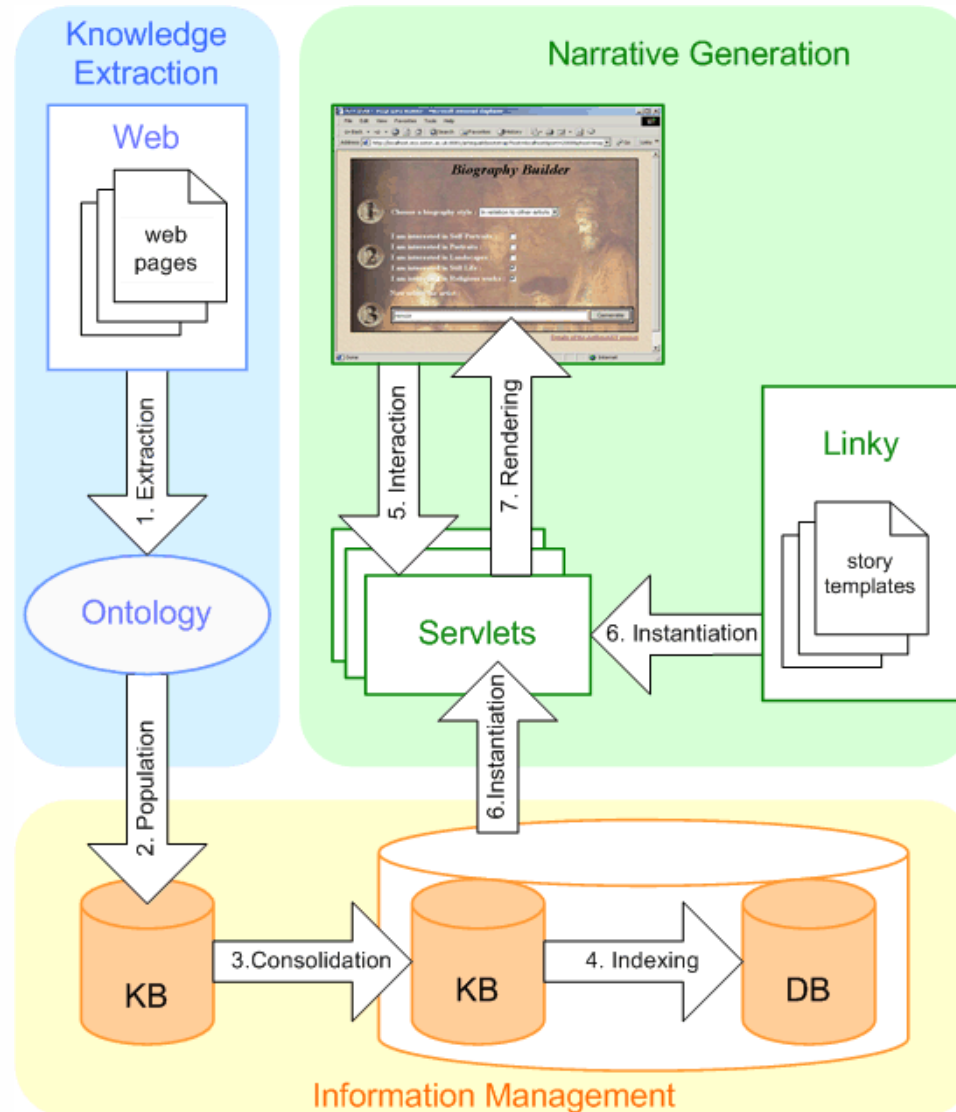
Knowledge Extraction

Output

```
<person>
<name>Rembrandt Harmenszoon
  van Rijn</name>
<place_of_birth>Leiden,
  Netherlands </place_of_birth>
<date_of_birth>
<day>15</day>
<month>July</month>
<year>1606</year>
</date_of_birth>
</person>
```



Architecture



Gate:(YearSpan2) -	1730-1809	Gate:(YearSpan1) -	the 1800s
Gate:(TempYear3) -	3591	Gate:(TempYear3) -	3592
Gate:(TempYear3) -	3756	Gate:(YearContext1) -	1864
Gate:(YearContext1) -	1882	Gate:(YearContext1) -	1887
Gate:(Person) -	H. Bartlett	Gate:(TempYear3) -	1875
Gate:(Location) -	Italy	Gate:(Person) -	Donato di Niccol
Gate:(Person) -	Di Betto	Gate:(TempYear3) -	1386
Gate:(YearSpan2) -	1475-1564	Gate:(DateNameRev) -	Nov 17, 1917
Gate:(Location) -	France	Gate:(Person) -	Rose
Gate:(GazDate Words) -	today	Gate:(Location) -	Western Civilization
Gate:(Organization) -	Paris Salon		

WordNet: Madison (location_type)
 WordNet: Britain (location_type)
 WordNet: Ireland (location_type)
 WordNet: Mother (person_type)
 WordNet: Paris (location_type)
 WordNet: Lorraine (location_type)
 WordNet: Sculptors (job_title)

C:\Documents and Settings\dem\My Documents\artequakt\example output.html - Microsoft Internet Ex...

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media Print Mail

Address C:\Documents and Settings\dem\My Documents\artequakt\example output.html Go Links

Rembrandt HARMENSZOOM

Summary Biography

Rembrandt Harmenszoon van Rijn was born on July 15, 1606, in Leiden, the Netherlands. His father was a miller who wanted the boy to follow a learned profession, but Rembrandt left the University of Leiden to study painting. His early work was devoted to showing the lines, light and shade, and color of the people he saw about him.

In 1636 Rembrandt began to depict quieter, more contemplative scenes with a new warmth in color. During the next few years three of his four children died in infancy, and in 1642 his wife died. In the 1630s and 1640s he made many landscape drawings and etchings. His landscape paintings are imaginative, rich portrayals of the land around him. Rembrandt was at his most inventive in the work popularly known as The Night Watch, painted in 1642. It depicts a group of city guardsmen awaiting the command to fall in line. Each man is painted with the care that Rembrandt gave to single portraits, yet the composition is such that the separate figures are second in interest to the effect of the whole. The canvas is brilliant with color, movement, and light. In the foreground are two men, one in bright yellow, the other in black. The shadow of one color tones down the lightness of the other. In the center of the painting is a little girl dressed in yellow.

Rembrandt HARMENSZOOM died 4 October 1669 in amsterdam.



References

<http://www.artequakt.ecs.soton.ac.uk/>

<http://www.artisteweb.org/>

<http://www.equator.ecs.soton.ac.uk/>

<http://www.aktors.org/>

<http://wordnet.princeton.edu/>

<http://gate.ac.uk/>

Comparison ...