Using Protégé for Automatic Ontology Instantiation

Harith Alani, Sanghee Kim, David Millard, Mark Weal,
Paul Lewis, Wendy Hall, Nigel Shadbolt

7th International Protégé Conference
ArtEquAKT

- Aims:
  - Use NLT to automatically extract relevant information about the life and work of artists from online documents
  - Feed this information automatically to an ontology designed for this domain
  - Generate stories by extracting and structuring information from the knowledge base in the form of biographical narratives
Motivation

• The knowledge is out there!
  – Available on the web, buried in text documents, not understood by machines!

• Semantic annotation might help
  – Annotations are rare
  – In the near future, annotations will probably not be rich or detailed enough to support the capture of extended amounts of content

• Knowledge extraction
  – There will always be a need for tools that can locate and extract specific types of knowledge, and store it in a KB for further inference and use
Architecture
ArtEquAKT Ontology

• Based on the Conceptual Reference Model (CRM) ontology

• Developed by CIDOC and promoted as an ISO standard

• CRM models the concepts and relationships used in cultural heritage documentation

• CRM is extended in ArtEquAKT to cover the life and work of artists
User Interface

Biography Builder

Choose a biography style: A Summary of the Life of an Artist

1. I am interested in Style:
2. I am interested in Family:
3. I am interested in Influences:
4. I am interested in Paintings:

Now select the artist:

rembrandt

Generate

Details of the ArtEquAKT project
Search and Filter Documents

- Documents are selected following these steps:
  1. Query search engine (Google) with the given artist name
  2. Calculate the similarity of the returned documents to some example documents about artists
  3. Apply some heuristics (e.g. minimum paragraph length) to filter out documents containing mainly tables or hyperlinks
  4. Send the remaining documents to the information extraction process
Knowledge Extraction Process
Send the identified triples to the ontology server:

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

1. Person_1 → Rembrandt Harmenszoon van Rijn
2. Person_1 → 15 July 1606
3. Person_1 → Leiden

RDF:

```xml
<kbrdf:about="&kb;Person_1"
    kb:name="Rembrandt Harmenszoon van Rijn"
    rdfs:label="Person_1">
    <kbrdf:has_information_text
        rdf:resource="&kb;Paragraph_1"/>
</kbrdf:Person>

<kbrdf:about="&kb;Date_1"
    kb:day="15"
    kb:month="7"
    kb:year="1606"
    rdfs:label="Date_1"/>
</kbrdf:Date>

<kbrdf:about="&kb;Place_1"
    kb:name="Leiden"
    rdfs:label="Place_1"/>
</kbrdf:Place>
```
Knowledge Management Component

Knowledge Extraction

Web

IE Tool

GATE

gated

named entities

paintings info

Armadillo

query expansion

WordNet

geographic names

Protégé Server

RDF

<kb:Person

kb:name="Renoir"
rdfs:label="P_1">

</kb:Person>

RDF triples

domain representation

instantiate ontology

Narrative Generation

retrieve knowledge

generate biographies
Knowledge Management Process

- Provide guidance to the extraction process
- Receives extracted knowledge in RDF format
- Instantiate the ontology with the given knowledge triples (add to the KB)
- Consolidation the knowledge
- Verify inconsistencies
- Ontology server providing a set of inference queries
Knowledge Consolidation
Types of Duplication

1. Duplicate instances of the same artist:
   - Rembrandt (1606, Leiden)
   - Rembrandt van Rijn (1606, Leiden)

2. Duplicate attribute values:
   - Rembrandt (1606, Leyden)
   - Rembrandt van Rijn (15 July 1606, Leiden)

3. Duplicate instances and attribute values:
   - Rembrandt (1606, Leyden)
   - Rembrandt van Rijn (15 July 1606, Leiden)
Consolidation Procedure

- **Unique Name Assumption**
  - e.g. all “Rembrandts” are merged
  - Not fool-proof, but works well in this limited domain

- **Information Overlap**
  - Merge similarly named artists if they share specific attribute values
  - e.g. Rembrandt, and Rembrandt Harmenszoon share a date of birth and a place of birth

- **Merge less specific information into more detailed ones**
  - This is mainly performed for dates and places
    - e.g. 1606 is merged into 15/7/1606; Netherlands is merged into Leiden
  - Place names are expanded with WordNet
    - Synonyms: Leiden = Leyden
    - Holonyms (part of): Leiden is part of The Netherlands
    - What if there is more than one Leiden? How do we know which to select?
      - Use the specificity variation of the given place for disambiguation
      - e.g. we are here looking for a Leiden that is related to the Netherlands
Verifying Inconsistencies

Hans Holbein the Elder (1465–1524). Born in Augsburg, Bavaria, the elder Holbein, like his brother Sigmund, painted richly colored religious works in the late Gothic style. In addition to the altar paintings that are his principal works, he designed church windows and also made a number of portrait drawings that foreshadow the work of his famous son. His later paintings show the transition from the late Gothic to the Renaissance style. He died in Isenheim, Alsace.
Verifying Inconsistencies

• We don’t aim for “the right answer”, but for some sort of a confidence value

• But which answer is more likely to be the correct one?
  – *Trust*: certain sources can be more trusted than others, but how do we judge that?
  – *Frequency*: certain facts might be extracted more often than others
  – *Extraction*: some extraction rules are more reliable than others!
Instantiated Ontology
Narrative Generation Component

Knowledge Extraction

Web

IE Tool

GATE

named entities

paintings info

Armadillo

query expansion

WordNet

domain representation

geographic

RDF

<kb:Person
  kb:name="Renoir"
  rdfs:label="P_1">;
</kb:Person>

RDF triples

Protégé Server

instantiate ontology

Narrative Generation

retrieve knowledge

generate biographies
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.

Best option is to have one paragraph that contains both pieces of information.
Intro paragraph:
DOB + place

Otherwise need a sequence of two fragments (DOB and place).

Either use a paragraph for each fragment, or construct out of raw facts

**Constructed sentence:**
*Rembrandt* was born on *July 15, 1606*. 
Rembrandt Harmenszoon van Rijn

Summary Biography

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

Rembrandt married Saskia van Uylenburgh, the cousin of a successful art dealer, in 1634. The marriage enhanced his career and brought him into contact with wealthy patrons who eagerly sought out his services. She had a substantial fortune, and this allowed Rembrandt to indulge in an extravagant lifestyle. He spent a great deal of money on an extensive collection of art. At auctions, he would often place an opening bid so high that no one else could place a bid. He bought a large house that later would place further strain on his resources. [1]

In 1625 Rembrandt returned to Leiden and set up his own independent studio and in 1629 he began the first of his numerous self-portraits (Self Portrait, 1629). The paintings he completed in his Leiden studio show a more mature use of chiaroscuro technique than his master, Lastman. Instead of using the lighting effect to create drama, Rembrandt used it to compose the entire painting. His use of light and dark gave his subjects a physical presence which for the first time involved the viewer in the painting. In his Leiden studio Rembrandt worked alongside Jan Lievens (1607 - 74) also a former student of Lastman. The two often worked on identical themes. In 1629 when Rembrandt produced paintings focusing on the apostle Paul (Peter and Paul in Conversation), Lievens painted the same apostle. [4]

Rembrandt Harmenszoon van Rijn, Dutch baroque artist, painted The Return of the Prodigal Son shortly before his death in 1669. It now hangs in the Hermitage in St. Petersburg, Russia. The subject comes from the Gospel According to Luke, Chapter 15. Gaze upon this painting and you will appreciate a sense of great tragedy. The contrast between the light and the dark, the two mysterious figures that appear dimly in the background, the woman behind the father, the stoic appearance of the old man who stands on the side, and the dishabille of the younger brother, all contribute to this sense of tragedy. One author has said that Rembrandt has elevated tragedy to a symbol of universal significance in this painting. I couldn’t agree more. This image represents the epitome of Rembrandt’s psychological mastery. [1]

Rembrandt was the first artist to practice self-portraiture as a specialty. In doing so, he created a medium for self-fashioning that has since inspired many artists. In 1668 a committee of Dutch art historians known as the Rembrandt Research Project took upon itself to distinguish between real Rembrandt paintings and all others. The project has met with less acceptance than anticipated. [8]

Rembrandt Harmenszoon van Rijn died 4 October 1669 in Amsterdam.

More detail available (2).

More detail available (2).

More detail available (2).

More detail available (2).

More detail available (2).

<table>
<thead>
<tr>
<th>Style</th>
<th>More detail available (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>[1] 3</td>
</tr>
</tbody>
</table>
ArtEquAKT Challenges

• Extraction
  – Some facts are too complex to extract
  – Rule-based IE is not always sufficient
  – Mapping of ontology terms to those in the text is unreliable (better for the ontology editor to include synonymous terms)

• Generation
  – A much wider range of facts should be extracted to be able to generate the biographies from scratch
  – Narrative construction may require richer semantic support (e.g., ontology of narrative)
  – Generation is not error free. We rely on people’s ability to parse and understand text
  – Difficult to track what facts have been included in the biography if these facts have not been identified

• Consolidation
  – Unreliable if the facts are extracted incorrectly
  – Could be inaccurate with sparse information
  – Geographical expansion can be wrong for places with same names

• Planning a bid for a second generation of ArtEquAKT
  – Entirely ontology driven
  – Domain independent
  – Much better text generation
Questions you may want to ask!

1. So does this system work with other domains?
2. Why bother with biographies anyway! There are many out there already!
3. Why extract knowledge, then use whole paragraphs in your biographies?!
4. Did you evaluate any of this?
5. What kind of knowledge did you manage to extract?
6. What did you say that Armadillo thing does?
7. How can we get GATE to recognise different entities?
8. How much rubbish does your system extract?
9. Can we use this system?! ….. please?

10. How would you like me to fund you? cash or check?