Apolda
A Practical Tool for Semantic Annotation

Christian Wartena

Rogier Brussee

Luit Gazendam

Willem-Olaf Huissen

Telematica Instituut
Enschede
The Netherlands

Overview

- Introduction: concepts in texts
- Problem
- Approaches
- Gate/Apolda
- Applications

Introduction: concepts in texts

- Find occurences of ontology concepts in a text
- Problems
  1. Spelling variants, inflection, encoding, etc.
  2. Synonyms
  3. Homonyms/Ambiguity
  4. Concept is ‘present’ but not mentioned
- Problems 3 and 4 scientifically very interesting
  – Problem 1 often underestimated and solved in a ad-hoc manner

Typical problems

- Encoding ‘special’ characters
- Capitalization
  - Concepts in ontology often capitalized!
  - *Typical* = *typical* (in title or sentence beginning)
  - *die Deutsche Bank ≠ eine deutsche Bank*
  - *national aeronautics and space administration = National Aeronautics and Space Administration*
- Inflection
  - *Museums = museum*
  - *British museums ≠ British Museum*
- *Interpunction*
  - *Art-works ≠ Art – works*
    - as in “The Metropolitan Museum of Art - Works of Art”

More Problems

- Synonymy
  - Especially names
    - *Bill Clinton, William Jefferson Clinton, William J. Clinton, governor Clinton, president Bill Clinton, president Clinton, Mr. Clinton*, *Clinton, W.J.*
- Multiword Expressions

Level of Abstraction

- At which level should look for concepts?

- Matching at high-level of abstraction
  - Consider text as a list of tokens or even lemmas
  - Advantage: No problems with line breaks, encoding, etc
  - Disadvantage: Loss of essential information

- Matching at low-level
  - Consider text as character sequence
  - All information directly available
  - No dependency on other analysis tools
  - But: lot of basic problems to solve

Approaches for detecting concepts

- Lexicalized ontologies
  - Easy to use in small projects
  - Easy to maintain
- Extending lexical resources with references to ontology
  - Natural separtion of lexical and ontological knowledge with clear interface
  - Problematic for multiword expressions
    - If they cannot be motivated from the lexicon
- Many project specific solutions

Apolda

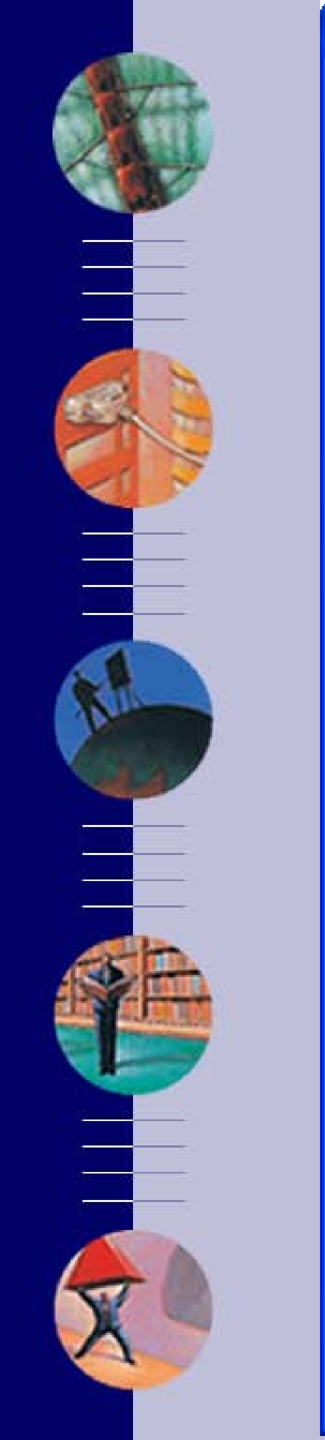
- **A**utomated **P**rocessing of **O**ntologies with **L**exical **D**enotations for **A**nnotation
- General open source solution
- For Lexicalized ontologies
- Concentrate on detecting all mentioned concepts
  - Leave all disambiguation to other tools

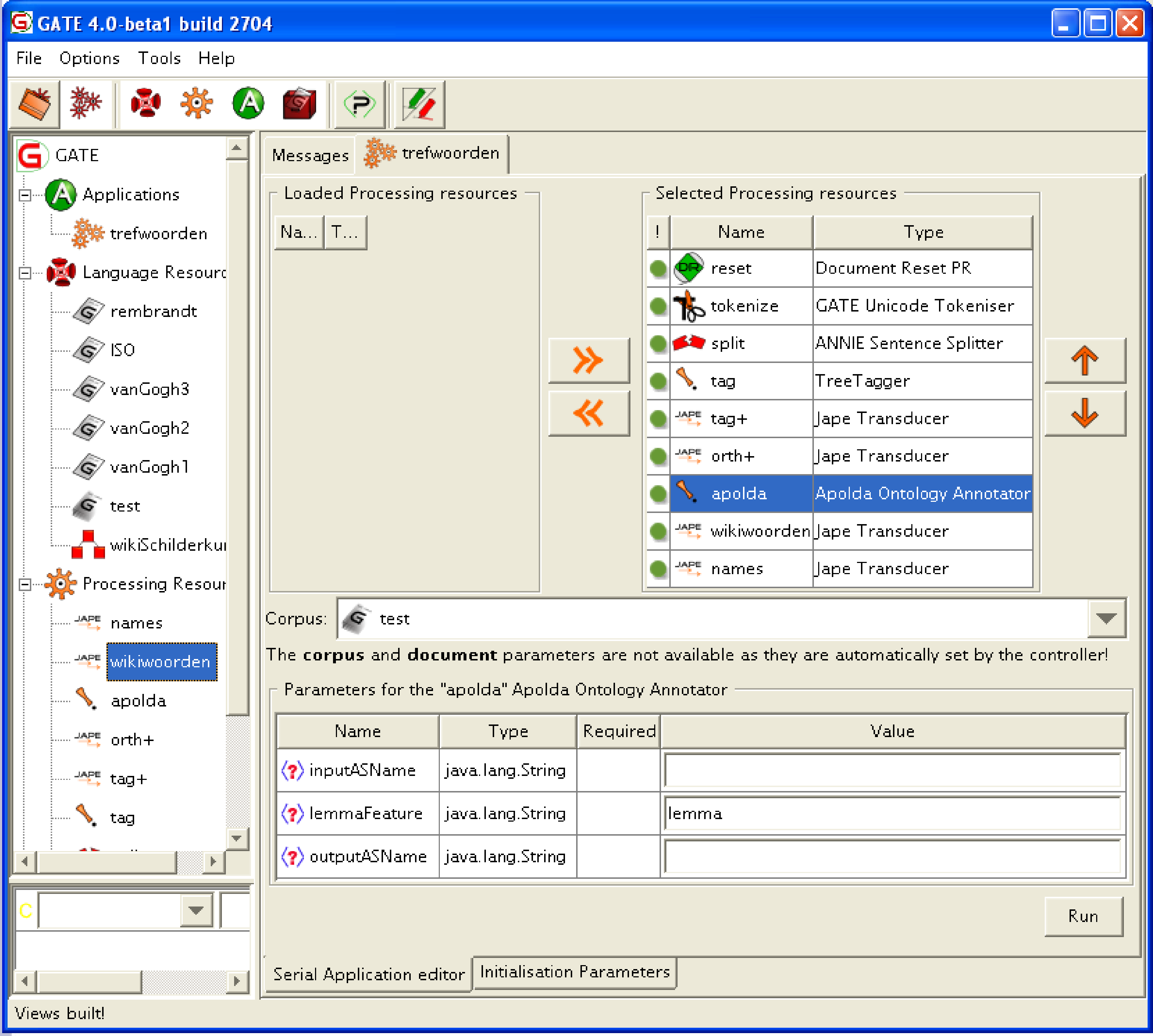
General Solution

- GATE plugin

- GATE:
  - General Architecture for Text Engineering
  - “an architecture, a free open source framework (or SDK) and graphical development environment”
  - Developed by NLP group of Sheffield University
  - http://gate.ac.uk/

- Apolda download:
  - http://apolda.sourceforge.net





GATE 4.0-beta1 build 2704

- GATE
  - Applications
    - trefwoorden
  - Language Resourc
    - rembrandt
    - ISO
    - vanGogh3
    - vanGogh2
    - vanGogh1
    - test
    - wikiSchilderku
  - Processing Resour
    - names
    - wikiwoorden
    - apolda
    - orth+
    - tag+
    - tag

Loaded Processing resources

Na...	T...

Selected Processing resources

!	Name	Type
●	reset	Document Reset PR
●	tokenize	GATE Unicode Tokeniser
●	split	ANNIE Sentence Splitter
●	tag	TreeTagger
●	tag+	Jape Transducer
●	orth+	Jape Transducer
●	apolda	Apolda Ontology Annotator
●	wikiwoorden	Jape Transducer
●	names	Jape Transducer

Corpus: test

The **corpus** and **document** parameters are not available as they are automatically set by the controller!

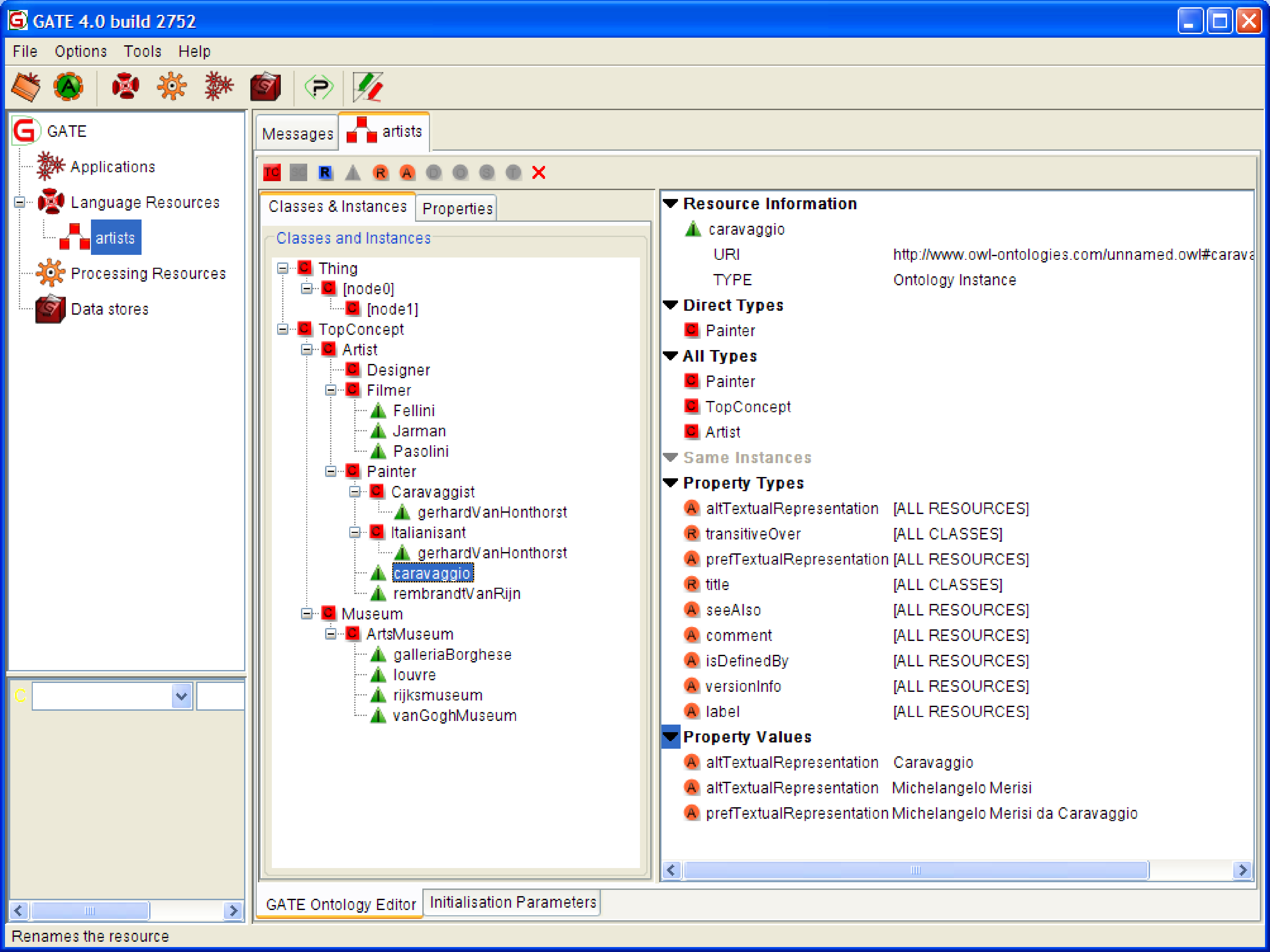
Parameters for the "apolda" Apolda Ontology Annotator

Name	Type	Required	Value
inputASName	java.lang.String		
lemmaFeature	java.lang.String		lemma
outputASName	java.lang.String		

Views built!

Lexicalized Ontologies

- Using Gate’s OWLIM ontology interface
  - Supports RDF/OWL, turtle, ntriples
- Expects explicit definition of textual representations (as annotation properties)
  - If no such labels are present identifiers are used
  - Deals with two different labels
  - If more than 2 types of labels are used make them `rdfs:subProperty` of one property.
- Supports different representations for different languages.



Classes and Instances

```
Thing
    [node0]
        [node1]
TopConcept
    Artist
        Designer
        Filmer
            Fellini
            Jarman
            Pasolini
        Painter
            Caravaggist
                gerhardVanHonthorst
            Italianisant
                gerhardVanHonthorst
            caravaggio
            rembrandtVanRijn
    Museum
        ArtsMuseum
            galleriaBorghese
            louvre
            rijksmuseum
            vanGoghMuseum
```

Resource Information

caravaggio

URI	http://www.owl-ontologies.com/unnamed.owl#carava
TYPE	Ontology Instance

Direct Types

Painter

All Types

Painter
TopConcept
Artist

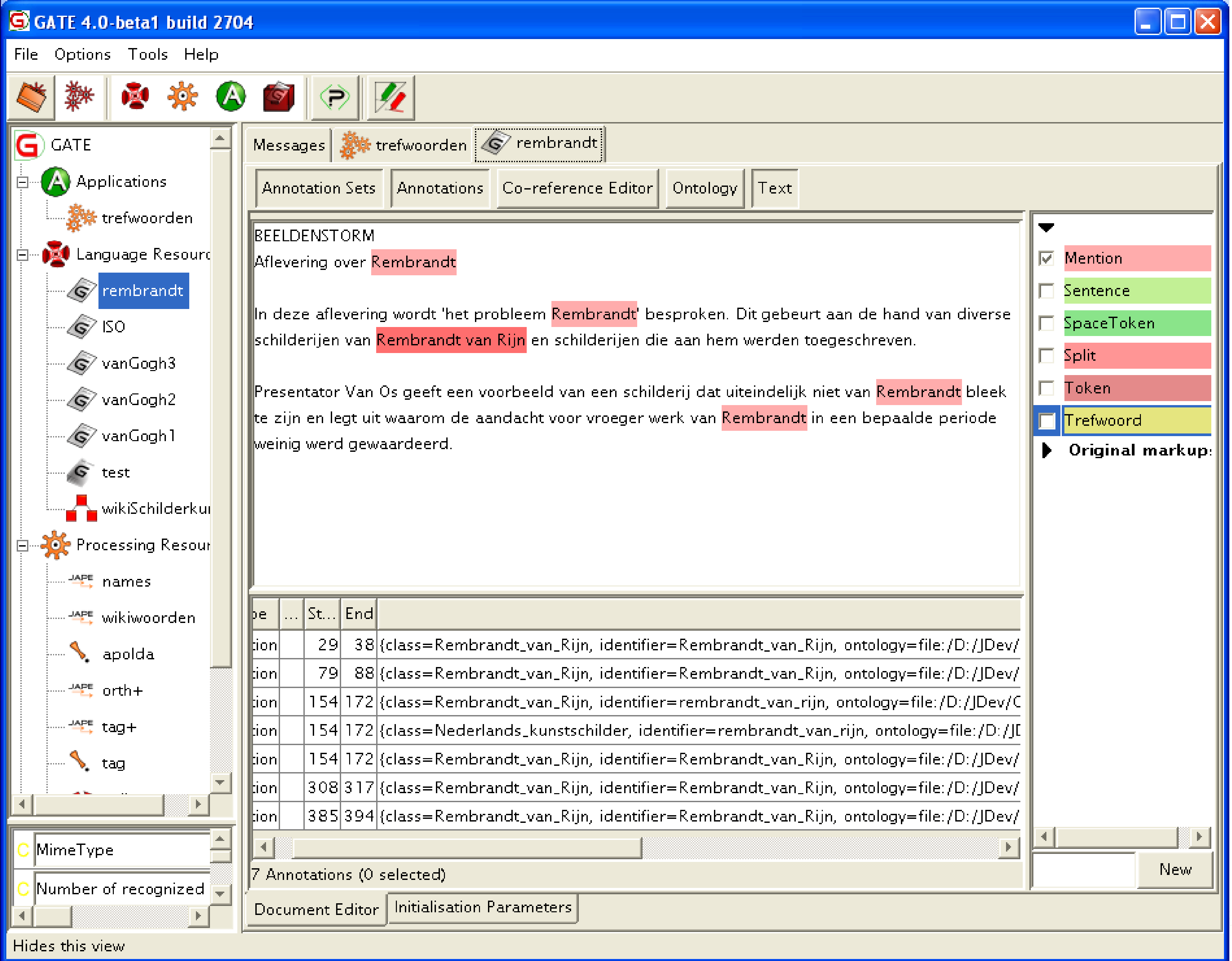
Same Instances

Property Types

altTextualRepresentation	[ALL RESOURCES]
transitiveOver	[ALL CLASSES]
prefTextualRepresentation	[ALL RESOURCES]
title	[ALL CLASSES]
seeAlso	[ALL RESOURCES]
comment	[ALL RESOURCES]
isDefinedBy	[ALL RESOURCES]
versionInfo	[ALL RESOURCES]
label	[ALL RESOURCES]

Property Values

altTextualRepresentation	Caravaggio
altTextualRepresentation	Michelangelo Merisi
prefTextualRepresentation	Michelangelo Merisi da Caravaggio



BEELDENSTORM
Aflevering over Rembrandt

In deze aflevering wordt 'het probleem Rembrandt' besproken. Dit gebeurt aan de hand van diverse schilderijen van Rembrandt van Rijn en schilderijen die aan hem werden toegeschreven.

Presentator Van Os geeft een voorbeeld van een schilderij dat uiteindelijk niet van Rembrandt bleek te zijn en legt uit waarom de aandacht voor vroeger werk van Rembrandt in een bepaalde periode weinig werd gewaardeerd.

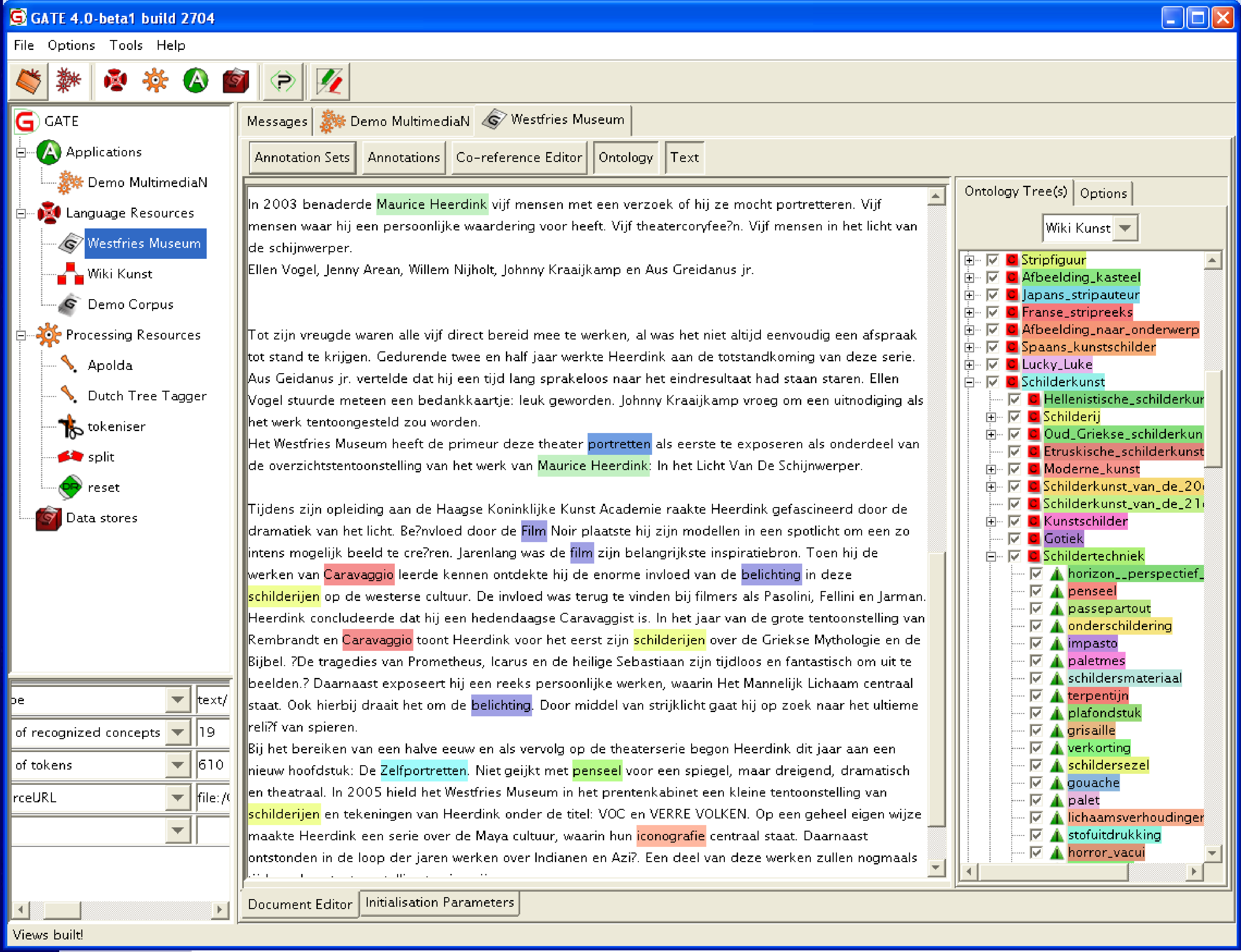
☑ Mention
□ Sentence
□ SpaceToken
□ Split
□ Token
□ Trefwoord
Original markup:

pe	...	St...	End	
tion		29	38	{class=Rembrandt_van_Rijn, identifier=Rembrandt_van_Rijn, ontology=file:/D:/JDev/
tion		79	88	{class=Rembrandt_van_Rijn, identifier=Rembrandt_van_Rijn, ontology=file:/D:/JDev/
tion		154	172	{class=Rembrandt_van_Rijn, identifier=rembrandt_van_rijn, ontology=file:/D:/JDev/C
tion		154	172	{class=Nederlands_kunstschilder, identifier=rembrandt_van_rijn, ontology=file:/D:/JI
tion		154	172	{class=Rembrandt_van_Rijn, identifier=Rembrandt_van_Rijn, ontology=file:/D:/JDev/
tion		308	317	{class=Rembrandt_van_Rijn, identifier=Rembrandt_van_Rijn, ontology=file:/D:/JDev/
tion		385	394	{class=Rembrandt_van_Rijn, identifier=Rembrandt_van_Rijn, ontology=file:/D:/JDev/

7 Annotations (0 selected)

MimeType
Number of recognized

Hides this view



In 2003 benaderde Maurice Heerdink vijf mensen met een verzoek of hij ze mocht portretteren. Vijf mensen waar hij een persoonlijke waardering voor heeft. Vijf theatercoryfee?n. Vijf mensen in het licht van de schijnwerper.
Ellen Vogel, Jenny Arean, Willem Nijholt, Johnny Kraaijkamp en Aus Greidanus jr.

Tot zijn vreugde waren alle vijf direct bereid mee te werken, al was het niet altijd eenvoudig een afspraak tot stand te krijgen. Gedurende twee en half jaar werkte Heerdink aan de totstandkoming van deze serie. Aus Geidanus jr. vertelde dat hij een tijd lang sprakeloos naar het eindresultaat had staan staren. Ellen Vogel stuurde meteen een bedankkaartje: leuk geworden. Johnny Kraaijkamp vroeg om een uitnodiging als het werk tentoongesteld zou worden.
Het Westfries Museum heeft de primeur deze theater portretten als eerste te exposeren als onderdeel van de overzichtstentoonstelling van het werk van Maurice Heerdink: In het Licht Van De Schijnwerper.

Tijdens zijn opleiding aan de Haagse Koninklijke Kunst Academie raakte Heerdink gefascineerd door de dramatiek van het licht. Be?nvloed door de Film Noir plaatste hij zijn modellen in een spotlicht om een zo intens mogelijk beeld te cre?ren. Jarenlang was de film zijn belangrijkste inspiratiebron. Toen hij de werken van Caravaggio leerde kennen ontdekte hij de enorme invloed van de belichting in deze schilderijen op de westerse cultuur. De invloed was terug te vinden bij filmers als Pasolini, Fellini en Jarman. Heerdink concludeerde dat hij een hedendaagse Caravaggist is. In het jaar van de grote tentoonstelling van Rembrandt en Caravaggio toont Heerdink voor het eerst zijn schilderijen over de Griekse Mythologie en de Bijbel. ?De tragedies van Prometheus, Icarus en de heilige Sebastiaan zijn tijdloos en fantastisch om uit te beelden.? Daarnaast exposeert hij een reeks persoonlijke werken, waarin Het Mannelijk Lichaam centraal staat. Ook hierbij draait het om de belichting. Door middel van strijklicht gaat hij op zoek naar het ultieme reli?f van spieren.
Bij het bereiken van een halve eeuw en als vervolg op de theaterserie begon Heerdink dit jaar aan een nieuw hoofdstuk: De Zelfportretten. Niet gejikt met penseel voor een spiegel, maar dreigend, dramatisch en theatraal. In 2005 hield het Westfries Museum in het prentenkabinet een kleine tentoonstelling van schilderijen en tekeningen van Heerdink onder de titel: VOC en VERRE VOLKEN. Op een geheel eigen wijze maakte Heerdink een serie over de Maya cultuur, waarin hun iconografie centraal staat. Daarnaast ontstonden in de loop der jaren werken over Indianen en Azi?. Een deel van deze werken zullen nogmaals

pe	text/
of recognized concepts	19
of tokens	610
rceURL	file:/(

Matching

- Match all textual representations, including overlapping ones
  - Exception: overlap of representations of same concept. E.g. *Rembrandt* / *Rembrandt van Rijn*

- Match case insensitive, literal strings, normalized strings and lemmas

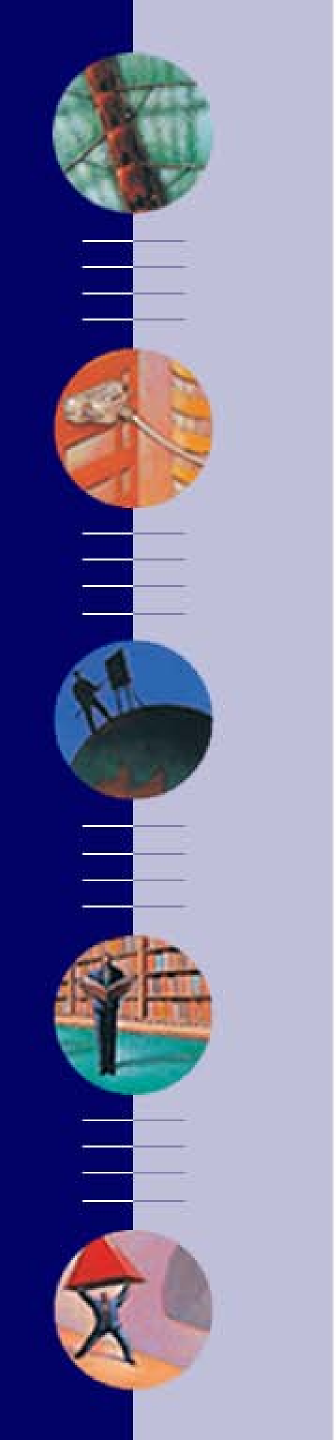
- Match only literal strings if quotation marks are used
  - (Remember the different matching levels)

Be cautious

- Reusing existing resources is always problematic
- Example:
  - Ontology of category structure, articles and redirects from wikipedia
  - Lemmatization
  - “*landdden*” (*landed*) → lemmat. → “*landen*” (to land)
  - “*landen*” *(countries)* → Apolda → “*land*” *(country)*
  - “*landdden*” (*to land*) • “*land*” (*country*)

Applications

- MultimediaN Video Tagging Project at Telematica Instituut
- Catch project at Dutch Institute for Sound and Vision
  - Keyword extraction from descriptions of videos
- Other, e.g. Master thesis project at Rotterdam University.
- About 140 downloads

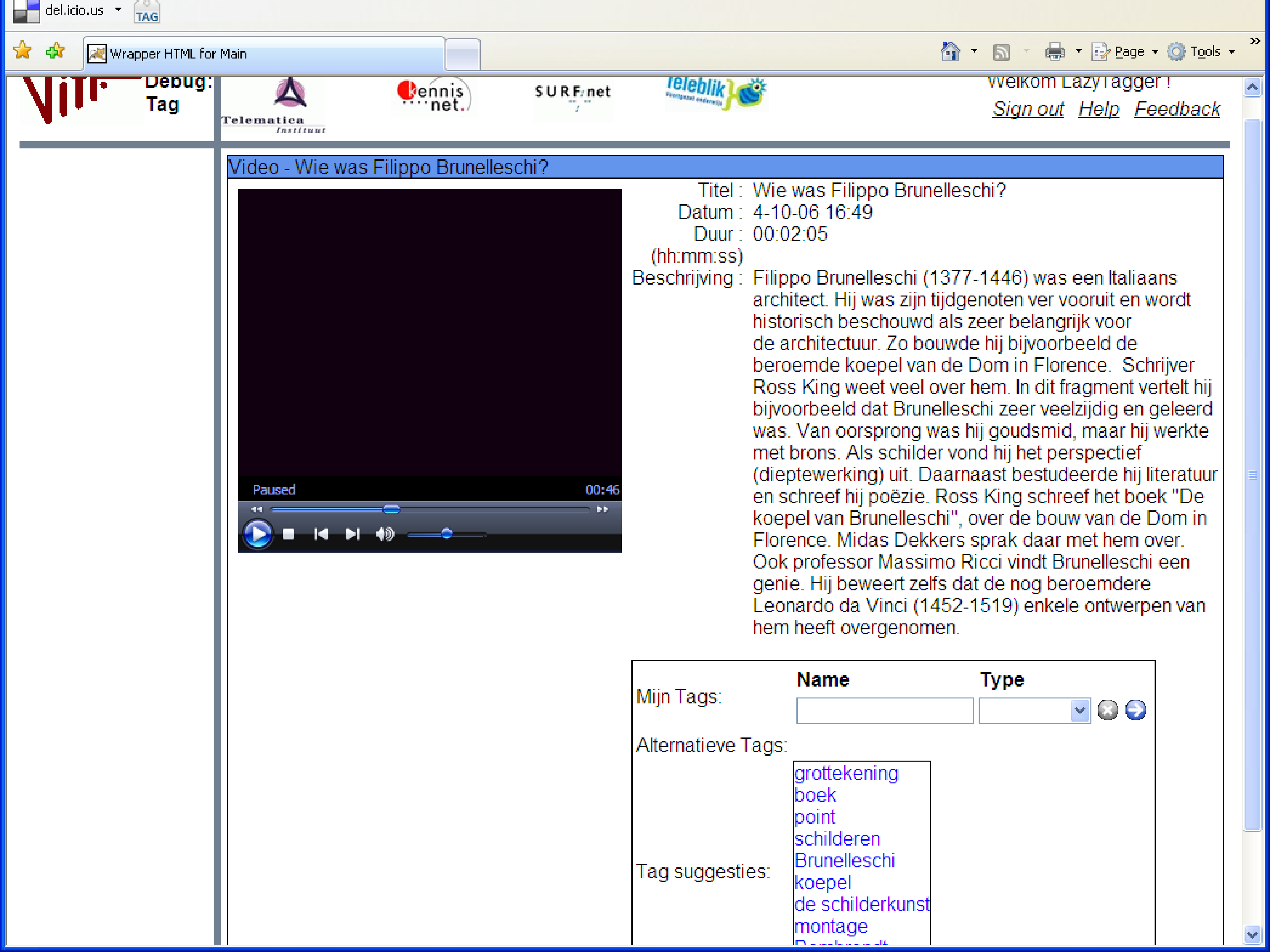


‘Cataloguing pipeline for TV-programs using Apolda’

- RDF\OWL representation of cataloguers ontology enriched with singular/plural form, synonyms, alternative spellings etc.
- collect textual resources related to the TV-program
- Use Apolda to highlight concept from the cataloguers ontology.
- Use TF.IDF and ontology relationships between found concepts to identify the semantically most relevant concepts in the texts.
- Recommend these concepts for the catalogue description of the TV-program.

MultimediaN

Telematica Instituut



Welkom LazyTagger !

Sign out Help Feedback

Video - Wie was Filippo Brunelleschi?

Titel :	Wie was Filippo Brunelleschi?
Datum :	4-10-06 16:49
Duur : (hh:mm:ss)	00:02:05
Beschrijving :	Filippo Brunelleschi (1377-1446) was een Italiaans architect. Hij was zijn tijdgenoten ver vooruit en wordt historisch beschouwd als zeer belangrijk voor de architectuur. Zo bouwde hij bijvoorbeeld de beroemde koepel van de Dom in Florence. Schrijver Ross King weet veel over hem. In dit fragment vertelt hij bijvoorbeeld dat Brunelleschi zeer veelzijdig en geleerd was. Van oorsprong was hij goudsmid, maar hij werkte met brons. Als schilder vond hij het perspectief (dieptewerking) uit. Daarnaast bestudeerde hij literatuur en schreef hij poëzie. Ross King schreef het boek "De koepel van Brunelleschi", over de bouw van de Dom in Florence. Midas Dekkers sprak daar met hem over. Ook professor Massimo Ricci vindt Brunelleschi een genie. Hij beweert zelfs dat de nog beroemdere Leonardo da Vinci (1452-1519) enkele ontwerpen van hem heeft overgenomen.

	Name	Type
Mijn Tags:		

Alternatieve Tags:

Tag suggesties:

grottekening
boek
point
schilderen
Brunelleschi
koepel
de schilderkunst
montage

Summary

- Find representations of concepts in a text
- Inventarisation of problems
- Implementation of a general solution
- Available for a widely used framework
- Usefull in several projects