# Named Entity Identification / Disambiguation Intelligent information access to linked data — weaving the cultural heritage web

#### Robert Kummer <r.kummer@uni-koeln.de>

Research Archive for Ancient Sculpture Universität of Cologne, Germany

18. September 2007

- 4 回 ト 4 ヨ ト 4 ヨ ト

## Outline

#### Digital Scholarship

#### Books (and objects) talking to each other Existing systems Linking hybrid collections Approach / Next steps

#### Conclusion

伺下 イヨト イヨト

# A model of digital scholarship (in the humanities)



Robert Kummer <r.kummer@uni-koeln.de>

Named Entity Identification / Disambiguation

Э

Existing systems Linking hybrid collections Approach / Next steps

# Books (and objects) talking to each other

- Digital libraries get over the paradigm of static books.
- They will offer sophisticated services to add value to their underlying content.
- They should deal with recombinant and dynamic data.
- Combine and link hybrid data, for example textual data and data about archaeological objects.

イロト イポト イヨト イヨト

Existing systems Linking hybrid collections Approach / Next steps

# Perseus (Reading Environment – Texts)



C. Julius Caesar, Gallic War

Agamemnon Search ("Agamemnon", "Hom. Od. 9.1", "denarius") [advanced search] [view abbreviations]

イロト イヨト イヨト イヨト

2

Hide browse be

our current position in the text is marked in blue. Click anywhere in the line to jump to another position.

This text is part of:	<ul> <li>All Gaul is divided into three parts, one of whi</li> </ul>	Hirt. Gal. 1.1 ch the Belgae inhabit.	Latin (T. Rice Holmes, 1914)	focus loa		
Materials	the Aquitani another, those who in their own lan Celts, in our Gauls, the third, All these differ from	guage are called n each other in	Notes (J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, 1898) focus is			
	language, customs and laws. The river Garonne	separates the Gauls	Places (automatically extracted)	hic		
View text chunked by:	from the Aquitani; the Marne and the Seine sepa	rate them from the	Sort places alphabetically, as they appear on the page, by frequency			
book : chapter : section	Belgae. Of all these, the Belgae are the bravest, b	ecause they are	Click on a place to search for it in this document. Rhine (6)			
	furthest from the civilization and refinement of	our] Province, and	Garonne (3) Seine (Franze) (2)			
	merchants least frequently resort to them, and in	nport those things	Rhone (2)			
Table of Contents:	Cermans who dwell beyond the Phine, with wh	om they are	Marine (Hanne) (2) France (France) (2)			
Theok 1	continually waging war: for which reason the He	lvetii also surnass the	Aquitaine (France) (2) Soain (Soain) (1)			
chapter 1	rest of the Gauls in valor, as they contend with th	e Germans in almost				
chapter 2	daily battles, when they either repel them from t	heir own territories,	References (62 total)	his		
chapter 3	or themselves wage war on their frontiers. One p	art of these, which it	<ul> <li>Commentary references to this page (1):</li> </ul>			
chapter 5	has been said that the Gauls occupy, takes its be	ginning at the river	<ul> <li>J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, Comp.</li> </ul>	tentary on		
chapter 6	Rhone ; it is bounded by the river Garonne, the c	cean, and the	Cross-references to this page (23):			
chapter 7	territories of the Belgae; it borders, too, on the si	de of the Sequani and	<ul> <li>Allen and Greenough's New Latin Grammar for Schools and C SYNTAX OF THE VERB</li> </ul>	olleges,		
chapter 8	the Helvetii, upon the river Rhine, and stretches	toward the north.	<ul> <li>Allen and Greenough's New Latin Grammar for Schools and C</li> </ul>	tolleges.		
chapter 9	The Beigae rises from the extreme frontier of Ga	in, extend to the lower	<ul> <li>L B. Greenouch, Benjamin L. D'Ooge, M. Grant Daniell, Comp.</li> </ul>	tentary on		
chapter 10	part of the river <u>knine</u> ; and look toward the hor	Puroposon	Caesar's Gallic War, 3.23			
chapter 12	mountaine and to that part of the ocean which is	near Spain: it looks	<ul> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> </ul>	and Pronouns		
chapter 13	between the setting of the sun and the north sta	r	<ul> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives, Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> </ul>	and Pronouns		
chapter 14	between the setting of the sun, and the north su	••	<ul> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> </ul>	and Pronouns		
chapter 15	C. Julius Caesar. Caesar's Gallic War. Translator. W. A.	McDevitte. Translator.	<ul> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> </ul>	and Pronouns and Pronouns		
chapter 16	W. S. Bohn. 1st Edition. New York. Harper & amp; Broth	ers. 1869. Harper's New	<ul> <li>Anne Mahoney, Overview of Latin Syntax, Nours, Adjectives,</li> </ul>	and Pronouns		
chapter 17	Classical Library.		<ul> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> <li>Anne Mahoney, Overview of Latin Syntax, Nouns, Adjectives,</li> </ul>	and Pronouns		
chapter 18	1		Anne Mahneer Overview of Latin Suntax, Noune, Adjustices	and Propount		

Existing systems Linking hybrid collections

## Perseus (Reading Environment – Artifacts)

K Perus	s Sculpture Cata	leg				Agamemnon Search ("Agamemnon", "Hom. Od. 9.1", "de [advanced search] [view abbreviatio	narius") <u>ns</u> l
Your current position in the text	is marked in blue. Click ar	synthese in the line to jump	to another position.			144	e browse bar
alphabetic hetteri entry group: entry:	-						
This text is part of:	<b>++</b>			Argina E S	Search		Nete
Creek and Roman Materials View text chunked by: Entry Entry Inter : entry	Aegina E 5, He 2, full figure fi	erakles of E. Ped. rom left	Argina E 5, Herakles of E. Ped full figure from right	(Vew Thursbasis (17))	Searching in finglish. <u>Marie scarch approse</u> Lives Search to: O All Calencions O Creek and Koman Materials O Prinsus Sculpture Catalog (this document)	)	
Table of Contents:	Collection:	Munich, Glyptot	hek		Display Preferences		hide
Anno letter *     Anno letter *	Title: Subject	Aegina, E. Ped. 2 Herakles, immed facing left. He ha the opposite corr another. The left	r, fig. E 5: Herakles liately identifiable by the lion-h is already unloosed one arrow a ner. His body is full of tension a arm is rigidly extend	eaded helmet, kneels t the Dying Warrior in s he prepares to shoot	Creek Display. Unicode (precombined) Vew by Default Translation Browne Bar. Show by default Update Proferences	•	
Aceina 5.10	Category:	Statuary group					
Aceina E 11	Material:	Marble					
Argina 6.3	Date	ca. 485 BC					
Argna E.4	Style:	Late Archaic	and a shade				
Acons E.S.	Context:	Aegina, Sanctuar	v of Aphaia				
Aceina E.B	Condition	Nearly complete	y or Apriana				
Argina E.9	Dimensions	H ca o Som					
Acqua tast Pedment 2 Acqua IV 1	Period:	Late Archaic					
Argina W 10	Region:	Saronic Gulf					
Acons # 11	Scale:	Life-size					
Angena IV 13 Angena IV 13 Angena IV 14 Angena IV 2	Form & Style	e: On the basis of s nd, whom he calls	tyle Ohly attributes Herakles as the Herakles Master.	nd the Right Helper E4			

・ロン ・回と ・ヨン ・ヨン

Existing systems Linking hybrid collections Approach / Next steps

#### Arachne



Robert Kummer <r.kummer@uni-koeln.de>

Named Entity Identification / Disambiguation

イロト イヨト イヨト イヨト

Existing systems Linking hybrid collections Approach / Next steps

# Entities of the historical world

- Single objects
   "Bust of Augustus", "Portraitstatue des Augustus"
- Places, topographical units "Aricia", "Ariccia"
- people (hist. relevant people)
   "Augustus", "Octavius Caesar", "Gaius Octavius"
- buildings, parts of buildings
   "Basilica Aemilia", "Basilica Aemilia et Fulvia"
- concepts (archaeological types, reception)

イロト イポト イヨト イヨト

Existing systems Linking hybrid collections Approach / Next steps

#### Linking overlapping collections



Existing systems Linking hybrid collections Approach / Next steps

## Linking books and objects

#### FRBRoo

CIDOC CRM



Robert Kummer <r.kummer@uni-koeln.de>

Named Entity Identification / Disambiguation

Э

Existing systems Linking hybrid collections Approach / Next steps

## A Personal Name Entry in the PE XML file

イロン イヨン イヨン イヨン

3

Existing systems Linking hybrid collections Approach / Next steps

## A Personal Name Entry in Smith

イロト イヨト イヨト イヨト

3

Existing systems Linking hybrid collections Approach / Next steps

# Approach

- Prospective methods
  - Subject headings
  - Thesauri
  - Gazetteers
- Retrospective methods
  - Statistical approaches
  - Semantic webs
  - Heuristics, rules
  - Users annotate / vote
- Expansible infrastructure (plug-ins)

(4月) イヨト イヨト

Existing systems Linking hybrid collections Approach / Next steps

#### First steps...

```
<namedEntity>
    <id>2100009</id>
    <location>BeschreibungBauwerk</locaton>
    <token>aaora</token>
    <disambiauationInfo>
        <entities>
            <name probability='1.00'>agora-geo</name>
        </entities>
    </disambiguationInfo>
    <context>
        <name>Attalos</name>
        <name>Eumenes</name>
    </context>
</namedEntity>
<namedEntity>
    <id>2100009</id>
    <location>BeschreibungBauwerk</locaton>
    <token>eumenes</token>
    <disambiguationInfo>
        <entities>
            <name probability='0.52'>eumenes-bio-3</name>
            <name probability='0.30'>eumenes-bio-5</name>
            <name probability='0.08'>eumenes-bio-4</name>
            <name probability='0.06'>eumenes-bio-1</name>
            <name probability='0.02'>eumenes-2</name>
            <name probability='0.01'>eumenes-bio-2</name>
            <name probability='0.01'>eumenes-1</name>
        </entities>
    </disambiauationInfo>
    <context>
        <name>Attalos</name>
        <name>Agora</name>
    </context>
</namedEntity>
```

Robert Kummer <r.kummer@uni-koeln.de> Named Entity Identification / Disambiguation

イロト イヨト イヨト イヨト

Existing systems Linking hybrid collections Approach / Next steps

# Named Entity Identification / Disambiguation

- Markup of primary sources: Multiple editions / translations of the same work can be aligned, carful markup can be projected to multiple editions, different works can be linked by identification of source citations (Thuc. 1.86).
- Author indices: Thousands of indices exist for Greek and Latin authors, they store judgements of experts as to which Alexander or Alexandria is meant in a given passage. These can be exploited for building authority files that enable scholars to refer to unambiguously identified entities.
- Reference works: Citation schemes are used to associate reference articles with particular passages of particular texts.
  - Smith Dictionary of Greek and Roman Geography (11,564 entries, 25,748 citations)
  - Smith Dictionary of Greek and Roman Biography (20,336 entries, 37,549 citations)

Existing systems Linking hybrid collections Approach / Next steps

# Tools available

- GATE: General Architecture for Text Engineering (Java, http://gate.ac.uk/demos/annie/annie.html)
  - Already in use by perseus
  - Understands ontologies
  - Gazetteer plugin
  - Nominal Coreference Resolution Component
- NTLK: Natural Language Toolkit (Python)
- openNLP (Java)
- Mallet (Java)

Questions: Can I adapt these toolkits to my needs?

イロト イポト イヨト イヨト

# Addressing the entity problem

- Incorporate (canonical) citation schemes to facilitate pointing to granular entities.
- Use standards like FRBRoo and CIDOC CRM to unify and connect large collections of heterogeneous data.
- Support entity identification and record linkage by providing document indices, encyclopedias, gazetters and other reference tools machine actionable.

・ 同 ト ・ ヨ ト ・ ヨ ト

# Examples for semantic searching

- Arachne und Perseus in Sesame: http: //highgrass.uni-koeln.de:8080/openrdf-workbench/
- Falcon-S: http://www.falcons.com.cn/
- Eculture Project: http://e-culture.multimedian.nl/
- Arachne und Perseus in Longwell: http://athena.perseus.tufts.edu/
- CRM-Browser: http://highgrass.uni-koeln.de: 8080/CRM-Browser/EntityBrowser

(4月) イヨト イヨト