Named Entity Identification / Disambiguation

Intelligent information access to linked data — weaving the cultural heritage web

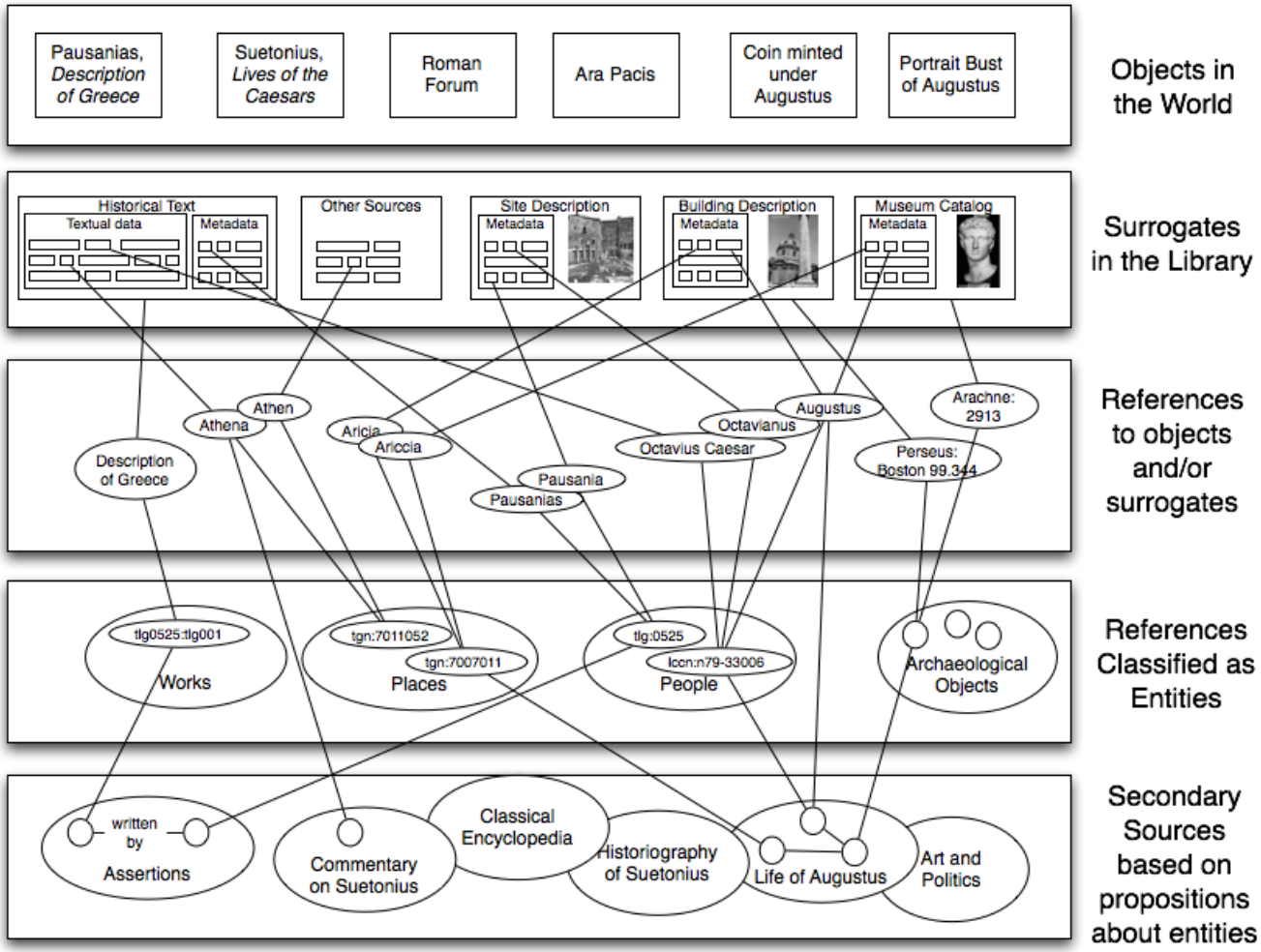
Robert Kummer <r.kummer@uni-koeln.de>

Research Archive for Ancient Sculpture
Universität of Cologne, Germany

18. September 2007

Outline

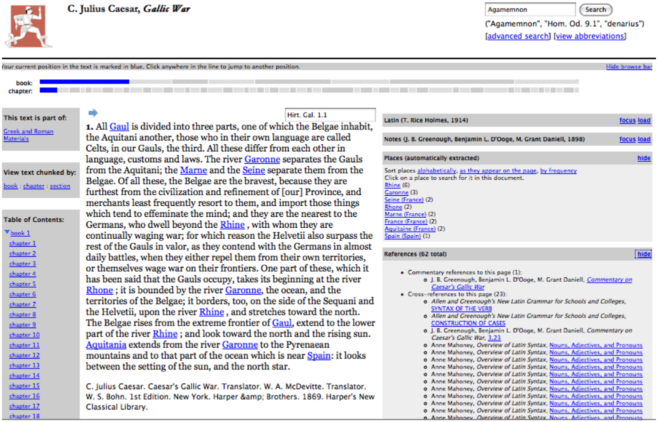
A model of digital scholarship (in the humanities)



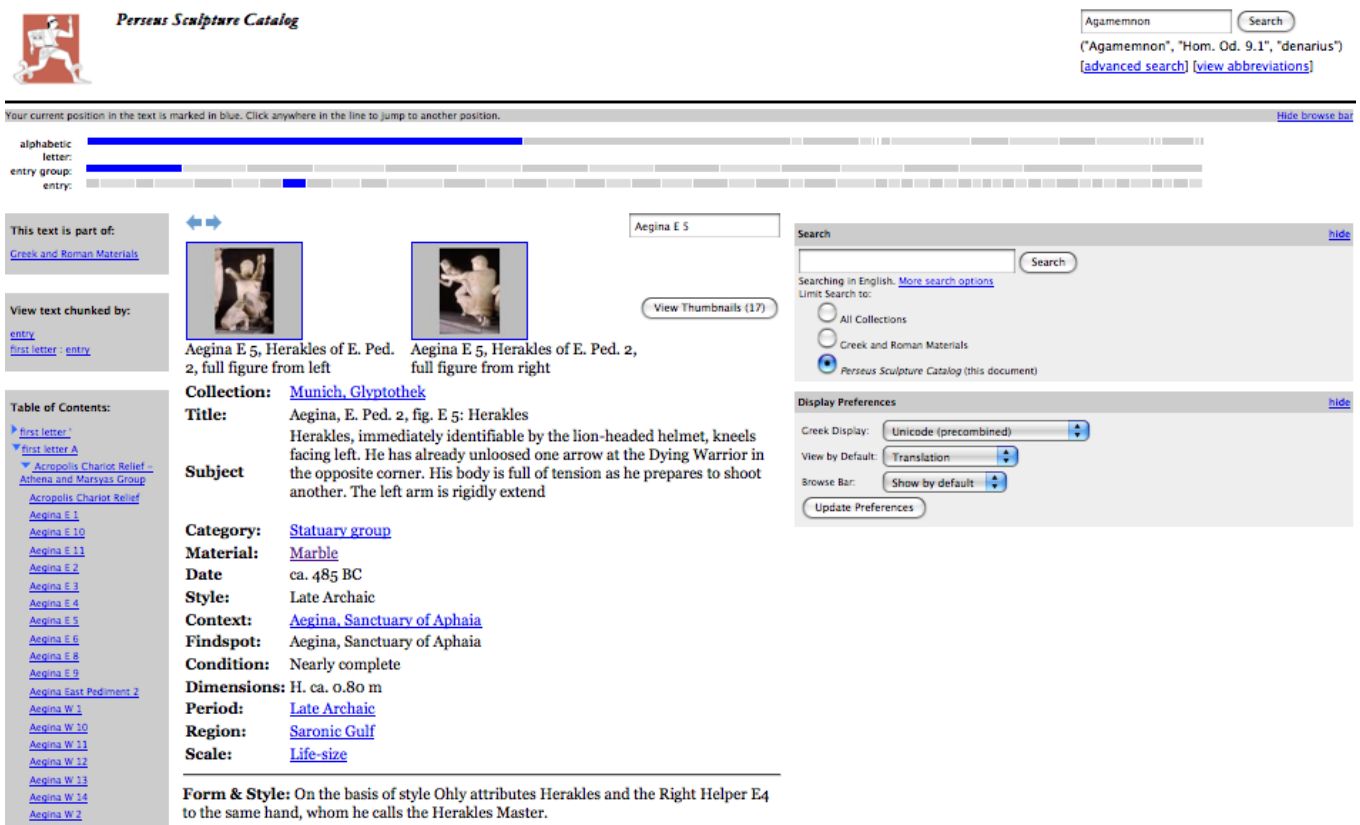
Books (and objects) talking to each other

- 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.

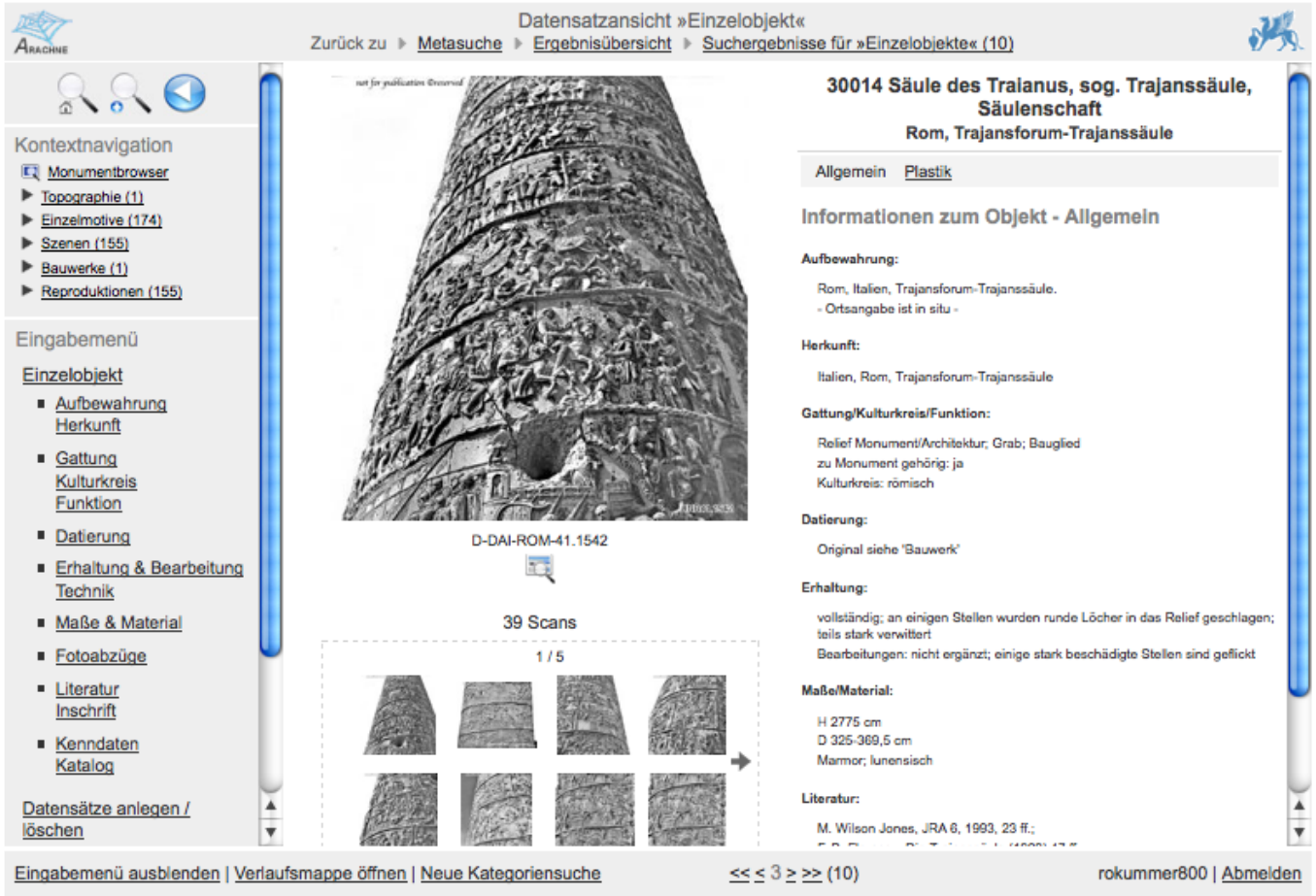
Perseus (Reading Environment – Texts)



Perseus (Reading Environment – Artifacts)



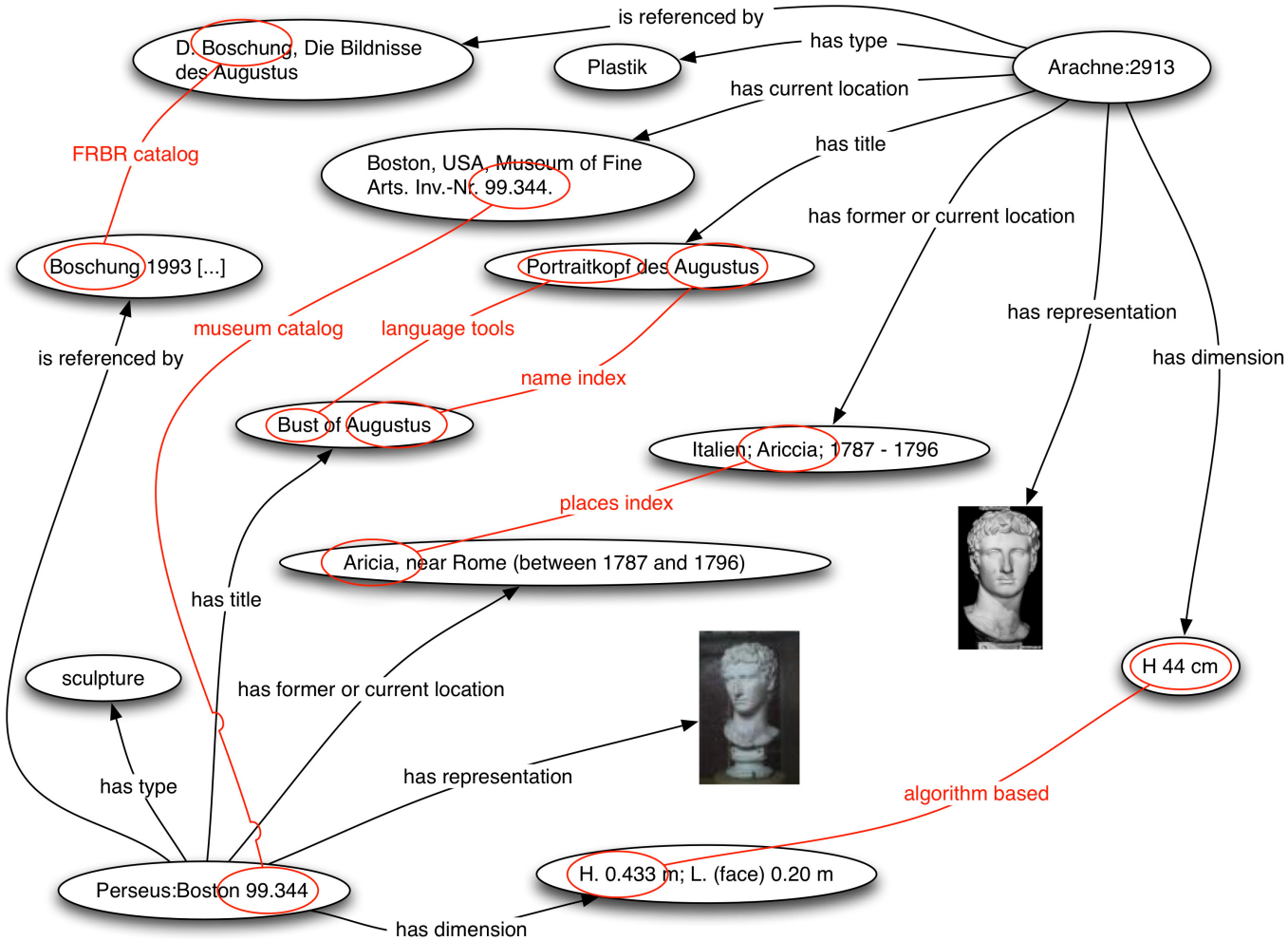
Arachne


Datensatzansicht »Einzelobjekt«
Zurück zu ▸ Metasuche ▸ Ergebnisübersicht ▸ Suchergebnisse für »Einzelobjekte« (10)
Kontextnavigation
Monumentbrowser
Topographie (1)
Einzelmotive (174)
Szenen (155)
Bauwerke (1)
Reproduktionen (155)
Eingabemenü
Einzelobjekt
Aufbewahrung
Herkunft
Gattung
Kulturkreis
Funktion
Datierung
Erhaltung & Bearbeitung
Technik
Maße & Material
Fotoabzüge
Literatur
Inschrift
Kenndaten
Katalog
Datensätze anlegen / löschen
D-DAI-ROM-41.1542
39 Scans
1 / 5
30014 Säule des Tralanus, sog. Trajanssäule, Säulenschaft
Rom, Trajansforum-Trajanssäule
Allgemein Plastik
Informationen zum Objekt - Allgemein
Aufbewahrung:
Rom, Italien, Trajansforum-Trajanssäule.
- Ortsangabe ist in situ -
Herkunft:
Italien, Rom, Trajansforum-Trajanssäule
Gattung/Kulturkreis/Funktion:
Relief Monument/Architektur; Grab; Bauglied
zu Monument gehörig: ja
Kulturkreis: römisch
Datierung:
Original siehe 'Bauwerk'
Erhaltung:
vollständig; an einigen Stellen wurden runde Löcher in das Relief geschlagen; teils stark verwittert
Bearbeitungen: nicht ergänzt; einige stark beschädigte Stellen sind geflickt
Maße/Material:
H 2775 cm
D 325-369,5 cm
Marmor; lunensisch
Literatur:
M. Wilson Jones, JRA 6, 1993, 23 ff.;
Eingabemenü ausblenden | Verlaufsmappe öffnen | Neue Kategoriensuche
<< < 3 > >> (10)
rokummer800 | Abmelden

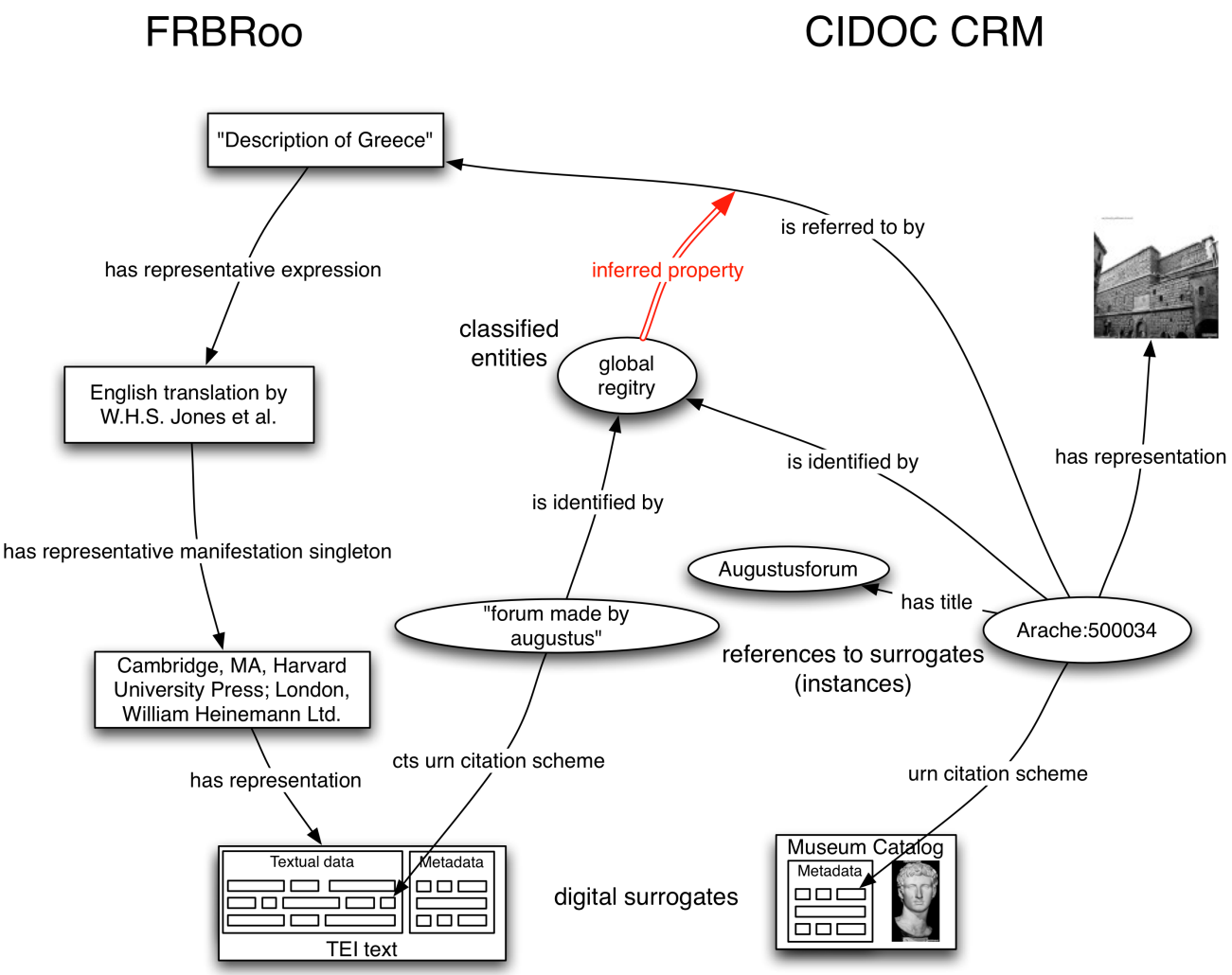
Entities of the historical world

- 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)

Linking overlapping collections



Linking books and objects



A Personal Name Entry in the PE XML file

```
<div1 type="entry" id="abderus">
  <head>Abderus</head>
  <div2 type="subentry" id="abderus-1">
    <head><persName>
    <surname>Abderus</surname> </persName>,son of <persName><surname>Hermes</surname>
    </persName></head>
      <div3 type="index"><list type="index">
      <item>killed by the mares of Diomedes: <bibl n="Apollod. 2.5.7">Apollod.2.5.7</bibl></item>
      <item>the city of Abdera founded by Herakles beside his grave: <bibl n="Apollod.
              2.5.7">Apollod. 2.5.7</bibl></item>
          </list></div3></div2></div1>
```

A Personal Name Entry in Smith

```
<div2 type="entry" id="abderus-bio-1">
    <head>
        <label>ABDE&prime;RUS</label>
    </head>
    <p>(<label lang="greek">*)/Abdhros</label>), a son of Hermes, or according to others of Thromius
        the Locrian. (Apollod. ii. 5. &sect; 8; Strab. vii. p. 331.) He was a favourite of
        Heracles, and was torn to pieces by the mares of Diomedes, which Heracles had given him to
        pursue the Bistones. Heracles is said to have built the town of Abdera to honour him.
        According to Hyginus, (<hi rend="ital">Fab.</hi> 30,) Abderus was a servant of Diomedes. the
        king of the Thracian Bistones, and was killed by Heracles together with his master and his
        four men-devouring horses. (Compare Philostrat. <hi rend="ital">Heroic.</hi> 3. &sect;
        1; 19. &sect; 2.) </p>
    <byline>[<ref target="author.L.S">L.S</ref> ]</byline>
</div2>
```

Approach

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

First steps...

```
<namedEntity>
    <id>2100009</id>
    <location>BeschreibungBauwerk</locaton>
    <token>agora</token>
    <disambiguationInfo>
        <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>
    </disambiguationInfo>
    <context>
        <name>Attalos</name>
        <name>Agora</name>
    </context>
</namedEntity>
```

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)

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?

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.

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`