CLiMB: Computational Linguistics for Metadata Building

Judith L. Klavans, Ph.D.
Carolyn Sheffield
Tandeep Sidhu

University of Maryland
What is CLiMB?

- An innovative project combining techniques of computational linguistics and computer science applied to metadata extraction for image access
- A tool to enhance metadata records for digital images, resulting in improved end-user retrieval of images
- A cataloger’s workbench, allowing catalogers to work quickly and efficiently from electronic texts
• Funded by Andrew W. Mellon Foundation

• CLiMB-1 : Columbia University 2002-2004
  Prototype toolkit
  Initial selection criteria for texts and images
  Formative testing with users

• CLiMB-2 : University of Maryland 2005-2007
Current Collaborations

• ARTstor

• Getty Vocabulary Institute
  • AAT
  • TGN
  • ULAN
Where CLiMB is today…

• Unique experience selecting
  • appropriate image collections and associated textual material
  • appropriate textual material and associated image collections

• Completed improved Toolkit
• Completed initial user studies
• Established collection criteria
Judith Klavans, PI, UMD
Jimmy Lin, UMD
Eileen G. Abels, UMD and Drexel
Dagobert Soergel, UMD
Lisbeth Herer, UMD
Rachael Bradley, UMD
Rachel Stazi, UMD
Carolyn Sheffield, UMD
Tandeep Sidhu, UMD
Jack Sullivan, UMD
Jeff Cohen, Bryn Mawr, U Penn
Laura Jenemann, Drexel
Joan Beaudoin, Drexel
Angela Giral, CU (ret)
Rebecca Passonneau, CU
Tae Yano, CU
Computational Linguistics for Metadata Building
**Record from AICT website**

Minimal metadata for image, no descriptive terms.

<table>
<thead>
<tr>
<th>Title or Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queen Nefertiti (front view), ca. 1350 B.C. [Dynasty XVIII]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>[Location] • AICT Photo ID #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aegyptisches Museum, Berlin #EN010</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams 3: pl. 5 22 [alt.]</td>
</tr>
<tr>
<td>Adams 4: pl. 5 22 [alt.]</td>
</tr>
<tr>
<td>Adams AAT: pl. 4.42 [alt.]</td>
</tr>
<tr>
<td>Adams AAT2: pl. 3.38 [alt.]</td>
</tr>
<tr>
<td>Gardner 11: pl. 3-33</td>
</tr>
<tr>
<td>Gardner 12: pl. 3-33</td>
</tr>
<tr>
<td>Hatt 4: pl. 3-42</td>
</tr>
<tr>
<td>H&amp;F 4: pl. 3.16 [alt.]</td>
</tr>
<tr>
<td>H&amp;F 5: pl. 3.16 [alt.]</td>
</tr>
<tr>
<td>Janson 5: pl. 82 [alt.]</td>
</tr>
<tr>
<td>Janson 5 R: pl. 79</td>
</tr>
<tr>
<td>Janson 6: pl. 2-28 [alt.]</td>
</tr>
<tr>
<td>Stokstad: pl. 3-38</td>
</tr>
<tr>
<td>Stokstad R: pl. 3-38</td>
</tr>
<tr>
<td>Stokstad 2: pl. 3-38</td>
</tr>
</tbody>
</table>

To download image:

Mac: hold down option and click photo size.

PC: hold down command and click desired image size.

- x250 pixels
- or
- x312 pixels
The famous painted limestone bust of Akhenaton’s queen, Nefertiti (fig. 3-33), exhibits a similar expression of entranced musing and an almost mannered sensitivity and delicacy of curving contour. The piece was found in the workshop of the queen’s official sculptor, Thutmose, and is a deliberately unfinished model very likely by the master’s own hand. The left eye socket still lacks the inlaid eyeball, making the portrait a kind of before-and-after demonstration piece. With this elegant bust, Thutmose may have been alluding to a heavy flower on its slender stalk by exaggerating the weight of the crowned head and the length of the almost serpentine neck...
The famous painted limestone bust of Akhenaton’s queen, Nefertiti (fig. 3-33), exhibits a similar expression of entranced musing and an almost mannered sensitivity and delicacy of curving contour. The piece was found in the workshop of the queen’s official sculptor, Thutmose, and is a deliberately unfinished model very likely by the master’s own hand. The left eye socket still lacks the inlaid eyeball, making the portrait a kind of before-and-after demonstration piece. With this elegant bust, Thutmose may have been alluding to a heavy flower on its slender stalk by exaggerating the weight of the crowned head and the length of the almost serpentine neck…
Issues this example demonstrates...

- Terms informed by art historical criteria:
  - deliberately unfinished model
- Ability to find related images
  - elongated neck
  - bust
- Potential for using Art & Architecture Thesaurus (AAT)
  - painted limestone bust
The CLiMB Toolkit

Tandeep Sidhu
• lawn (textile) (<textile materials by weaving technique>, <textile materials by process or technique>, ... Materials)
• Note: Sheer, lightweight plain-woven textile, originally of linen now also of fine combed cotton yarn, filled with starch or sizing. Often used for handkerchiefs, aprons, and curtains.
• Terms: lawn (textile) (preferred, C,U,D, American English-P)
Facet/Hierarchy Code: M.MT
Hierarchical Position:

Materials Facet
.... Materials
........ materials
........... <materials by form>
............. <materials by physical form>
.......................... <fiber and fiber products>
............................ <fiber products>
............................ <textile materials>
............................ <textile materials by process or technique>
............................ <textile materials by weaving technique>
................................ lawn (textile)
• **lawns (landscaped grass)** (<landscaped-site elements>, <site elements>, ... Built Environment)

• **Note:** Areas of cultivated grass or other ground cover maintained for aesthetic quality or recreation.

• **Terms:** lawns (landscaped grass) (**preferred**, C, U, D, American English-P) lawn (landscaped grass) (C, U, AD, American English)

Facet/Hierarchy Code: **V.RM**

Hierarchical Position:

Objects Facet
.... Built Environment
........ Open Spaces and Site Elements
............... <site elements>
................... <landscaped-site elements>
....................... lawns (landscaped grass)
Homographs

• A standard and difficult problem in computational linguistics
• Known as disambiguation

• One sense per discourse, but which sense?
• Developed algorithms to select a preferred sense
• CLiMB - 76% accurate
  • Better than 60%!
Art & Architecture Thesaurus (AAT)

- structured vocabulary
- ~34,000 records
- represents
  - art
  - architecture
  - material culture
  - archival materials
- Getty Vocabularies Institute

<table>
<thead>
<tr>
<th>Senses</th>
<th>Terms</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1097</td>
<td>lawn</td>
</tr>
<tr>
<td>3</td>
<td>215</td>
<td>abaci</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>boards</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>ocher</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>carmine</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>slate</td>
</tr>
</tbody>
</table>
Considerations

• Types of collections:
  • Structured: collection catalogs
  • Unstructured: textbooks, lectures

• Availability:
  • Of electronic text
  • Of digitized images
  • Of existing catalog records

• Copyright issues
Users

Image Catalogers

Image Searchers

Carolyn Sheffield
Cataloger Studies

Workflow without CLiMB
Goal: Establish general cataloger workflow
- Step by step process
- Number and type of subject terms
- Time spent on subject description

Workflow with CLiMB
Goal: Measure the impact of the Toolkit
- Thoroughness of subject metadata
- Efficiency of subject description
- Cataloger reaction
Image Searcher Studies

Original catalog records (before CLiMB)
Goal: Establish general cataloger workflow
- Retrieval
- End user satisfaction

CLiMB-enhanced catalog records
Goal: Measure the impact of the Toolkit
- Retrieval
- End user satisfaction
Summary

• CLiMB Toolkit to aid in scholarly digital library cataloging
• CL useful for the text associated with images

Research Issues

• Use of machine-learning to perfect filtering (needs data)
• Evaluation for two user groups
• Ontological associations
Select

• User decides
  which of the proposed subject descriptors are actually good index terms

Who are the initial users?
• Cataloger’s Toolkit
  Computationally-assisted Task
• Non-invasive surgery
• Surgeon still does the work!
• Image catalogers
• Reference librarians
• Tool builders for image libraries
• Research: jklavans@umd.edu
• Student Projects: Contact any of us
• All else:

Carolyn Sheffield
csheffie@umd.edu

• www.umiacs.umd.edu/~climb
Computational Linguistics for Metadata Building