Mining texts for image terms: the CLiMB project

Judith L. Klavans
University of Maryland, College Park
jklavans@umd.edu

Eileen Abels
Drexel University
eileen.abels@ischool.drexel.edu

Jimmy Lin
University of Maryland, College Park
jimmylin@umd.edu

Rebecca Passonneau
Columbia University
becky@cs.columbia.edu

Carolyn Sheffield
University of Maryland, College Park

Dagobert Soergel
University of Maryland, College Park
dsoergel@umd.edu

The CLiMB (Computational Linguistics for Metadata Building) project addresses the existing gap in subject metadata for images, particularly for the domains of art history, architecture, and landscape architecture. Within each of these domains, image collections are increasingly available online yet subject access points for these images remain minimal. In an observational study with six image catalogers, we found that typically 1 – 8 subject terms are assigned, and that many legacy records lack subject entries altogether. Studies on end users' image searching indicate that this level of subject description is often insufficient. In a study of the image-searching behaviors of faculty and graduate students in American history, Choi and Rasmussen 2003 found that 92% of the 38 participants in their study considered the textual information associated with the images in the Library of Congress' American Memory Collection to be inadequate. The number of subject descriptors assigned to an image in this collection is comparable to what we found in the exploratory CLiMB studies. Furthermore, these searchers submitted more subject-oriented queries than known-artist and title queries. Similar results demonstrating the importance of subject retrieval have been reported in other studies, including Keister, Collins, and Chen 1994.

Under the hypothesis that searchers do not find images they seek partly due to inadequate subject description in metadata fields, the CLiMB project was initiated to address this subject metadata gap by applying automatic and semi-automatic techniques to the identification, extraction, and thesaural linking of subject terms. The CLiMB Toolkit processes text associated with an image through natural language processing (NLP), categorization using machine learning (ML), and disambiguation techniques to identify, filter, and normalize high-quality subject descriptors. Like Pastra et al. 2003 we use NLP techniques and domain-specific ontologies, although our focus is on associated texts such as art historical surveys or curatorial essays rather than captions; unlike generic image search, such as in Google, we analyze beyond keywords and we use text which is specifically and clearly related to an image. For this project, we use the standard Cataloging Cultural Objects (CCO) definition of subject metadata[1] as including terms which provide "an identification, description, or interpretation of what is depicted in and by a work or image."

In order to understand the cataloging process and to inform our system design, we conducted studies on the image cataloging workflow and the process of subject term assignment. Our goal was to collect data on the humanities-driven process as a whole to be able to incorporate our results into an existing workflow and thus assist a portion of the workflow with automatic techniques. An additional purpose of studying the cataloging process was to permit the development of system functionality, i.e., the implementation of rules or the use of statistical methods to identify high-quality subject descriptors in scholarly texts. As part of the CLiMB evaluation, we have established a series of test collections in the fields of art history, architecture, and landscape architecture. These three domains were selected in part because of the existing overlap in domain-specific vocabulary. Testing with distinct but related domains enables us to test for disambiguation issues which arise in the context of specialized vocabularies. For example, the Getty Art & Architecture Thesaurus (AAT) provides many senses of the term *panel* which apply to either the fine arts, architecture, or both, depending on context. In the context of fine arts, *panel* may refer to a *small painting on wood* whereas in the context of architecture, *panel* may refer to a *distinct section of a wall, within a border or frame*.

Figure 1 shows the CLiMB architecture which produces subject term recommendations that can be used into the image cataloging workflow observed in visual resource centers:

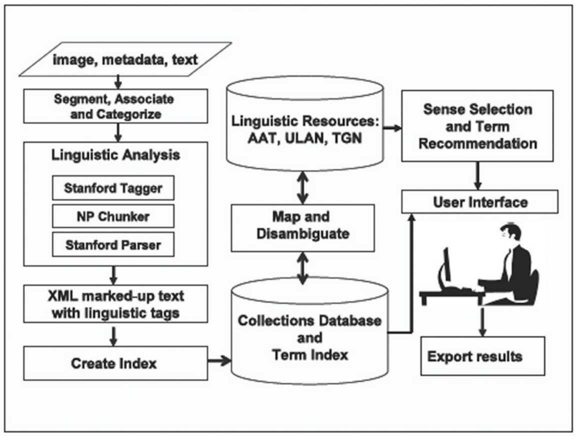

image, metadata, text
Segment, Associate and Categorize
Linguistic Analysis
Stanford Tagger
NP Chunker
Stanford Parser
XML marked-up text with linguistic tags
Create Index
Linguistic Resources: AAT, ULAN, TGN
Map and Disambiguate
Collections Database and Term Index
Sense Selection and Term Recommendation
User Interface
Export results

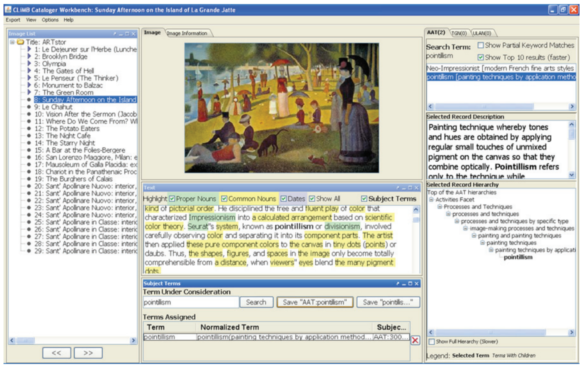
Figure 1. The CLiMB toolkit architecture

CLiMB combines new and pre-existing technologies in a flexible, client-side architecture which has been implemented in a downloadable toolkit and which can be tailored to the user's needs. In addition to matching segments of texts to referenced images, we are developing methods to categorize spans of text (e.g., sentences or paragraphs) as to their semantic function relative to the image. For example, a sentence might describe an artist's life events (e.g. "during his childhood", "while on her trip to Italy", "at the death of his father") or the style of the work ("impressionism"). A set of seven categories – Image Content, Interpretation, Implementation, Historical Context, Biographical Information, Significance, and Comparison – has been initially proposed through textual analysis of art survey texts. These categories have been tested through a series of labeling experiments. Full details are available in Passonneau et al. 2008. The output of this categorization will be incorporated in future versions of the Toolkit, and will be used as part of the disambiguation component.

An important contribution of the CLiMB project is the development of a disambiguation component, enabling the system to move beyond standard keyword-based indexing by associating words and terms that have multiple meanings which correspond to different descriptors with the correct meaning in context. The ability to select one sense from many is referred to as lexical disambiguation. Results of our ongoing studies on sense disambiguation using hierarchically structured faceted thesauri and lexical resources, such as the Art and Architecture Thesaurus and WordNet, will be presented**.** We have experimented with the use of WordNet, with different levels of the facets of the AAT, and with different degrees of filtering for modifiers in noun phrases. We also have results on setting weights for each of these factors to determine the most accurate disambiguation techniques.

One of the most vexing problems in word sense disambiguation is the fact that often several senses could be considered correct within a given context. Therefore, evaluation can be a challenge since there may be no clear-cut right or wrong. The need for fuzzy evaluation will be discussed in our presentation, with a demonstration of different ways to measure precision and recall against a "moving target" baseline.

Figure 2 shows the CLiMB interface in its current state as of Fall 2008; as we use the results of our experimental research, this interface may change as of the time of presentation of the paper.



Title: ARTstor
1: Le Dejeuner sur l'Herbe (Lunche
2: Brooklyn Bridge
3: Olympia
4: The Gates of Hell
5: Le Penseur (The Thinker)
6: Monument to Balzac
7: The Green Room
8: Sunday Afternoon on the Island
9: Le Chahut
10: Vision After the Sermon (Jacob
11: Where Do We Come From? Wl
12: The Potato Eaters
13: The Night Cafe
14: The Starry Night
15: A Bar at the Folies-Bergere
16: San Lorenzo Maggiore, Milan: e
17: Mausoleum of Galla Placidia: ex
18: Chariot in the Panathenaic Proc
19: The Burghers of Calais
20: Sant' Apolinare Nuovo: interior,
21: Sant' Apolinare Nuovo: interior,
22: Sant' Apolinare Nuovo: interior,
23: Sant' Apolinare Nuovo: interior,
24: Sant' Apolinare Nuovo: interior,
25: Sant' Apolinare in Classe: interio
26: Sant' Apolinare in Classe: interio
27: Sant' Apolinare in Classe: interio
28: Sant' Apolinare in Classe: interio
29: Sant' Apolinare in Classe: interio

Highlight ☑ Proper Nouns ☑ Common Nouns ☑ Dates ☑ Show All ☑ Subject Terms

kind of pictorial order. He disciplined the free and fluent play of color that characterized Impressionism into a calculated arrangement based on scientific color theory. Seurat"s system, known as **pointillism** or divisionism, involved carefully observing color and separating it into its component parts. The artist then applied these pure component colors to the canvas in tiny dots (points) or daubs. Thus, the shapes, figures, and spaces in the image only become totally comprehensible from a distance, when viewers" eyes blend the many pigment dots.

Term Under Consideration
pointilism

Terms Assigned

Term	Normalized Term	Subjec...
pointillism	pointillism(painting techniques by application method...	AAT:300...

Search Term: pointilism
□ Show Partial Keyword Matches
☑ Show Top 10 results (faster)

Neo-Impressionist [modern French fine arts styles
pointilism [painting techniques by application metho

Selected Record Description
Painting technique whereby tones and hues are obtained by applying regular small touches of unmixed pigment on the canvas so that they combine optically. **Pointillism** refers only to the technique while

Selected Record Hierarchy
Top of the AAT hierarchies
Activities Facet
Processes and Techniques
processes and techniques
processes and techniques by specific type
image-making processes and techniques
painting and painting techniques
painting techniques
painting techniques by applicati
pointillism

□ Show Full Hierarchy (Slower)
Legend: **Selected Term** *Terms With Children*

Figure 2: The CLiMB Toolkit Interface

Note in Figure Two that the collection under review is found in the left panel, the image is in the center, the analyzed text is shown to the cataloger, with the searchable Getty thesaural resources (AAT, Thesaurus of Geographical Names (TGN) and Union List of Artist Names (ULAN)) in the right panel. The cataloger can select subject terms, and when possible, normalize according to the Getty unique identifier. All interface panes are flexible, and can be hidden or enlarged, as required by the user. Cataloger subject term selections can be exported in a range of formats (see Export button in upper left hand corner of Figure 2) for incorporation into an existing catalog record.

To sum, in this paper we will present:

- The problem of subject term access in image retrieval
- **The CLiMB system**, which utilizes computational linguistics and machine learning to improve basic keyword search through:
  - Semantic categorization of text segments
  - Disambiguation

- User evaluation studies and findings

Selected References

Chen, H. (2001) An Analysis of Image Retrieval Tasks in the Field of Art History. Information Processing & Management, Vol. 37: 701-720.

Choi, Y. and E. Rasmussen (2003) Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History. Journal of the American Society for Information Science and Technology, Vol. 54: 498-511.

Collins, K. (1998) Providing Subject Access to Images: A Study of User Queries. The American Archivist, Vol. 61: 36-55.

Keister, L.H. (1994) User Types and Queries: Impact on Image Access Systems. In: Fidel, R., T.B. Hahn, E. Rasmussen, P. J. Smith (eds.): Challenges in Indexing Electronic Text and Images. Learned Information for the American Society of Information Science, Medford: 7-22.

Klavans, Judith L, Carolyn Sheffield, Eileen Abels, Jimmy Lin, Rebecca Passonneau, Tandeep Sidhu, and Dagobert Soergel (2009) Computational Linguistics for Metadata Building (CLiMB): Using Text Mining for the Automatic Identification, Categorization, and Disambiguation of Subject Terms for Image Metadata. Journal of Multimedia Tools and Applications, Special issue on Metadata Mining for Image Understanding (MMIU) 42(1):115-138. Elsevier: Paris.

Passonneau, R., T. Yano, T. Lippincott, J. Klavans (2008) Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning. 3rd International Conference on Computer Vision Theory and Applications, Workshop on Metadata Mining for Image Understanding: 13-22.

Pastra, K., H. Saggion, Y. Wilks, (2003) Intelligent Indexing of Crime-Scene Photographs. In: IEEE Intelligent Systems: Special Issue on Advances in Natural Language and Processing, Vol. 18, Iss. 1: 55-61.

Notes

[1] http://vraweb.org/ccoweb/cco/parttwo_chapter6.html.

Library Collaboration with Large Digital Humanities Projects

William A. Kretzschmar, Jr.
University of Georgia
kretzsch@uga.edu

William G. Potter
University of Georgia

At DH2007 a special session on "finishing" large humanities research projects (now forthcoming as a cluster in *DHQ*) suggested, in part, that particular stages of such projects might be completed, but that continuing institutional support was important for the long-term sustainability of the projects and their products. At DH2008 a special session was devoted to "Aspects of Sustainability in Digital Humanities," in which technical, organizational, and scholarly dimensions were discussed with reference to a museum project, along with metadata and the issue of portability in other settings. In this paper, we would like to continue the theme of sustainability. We will discuss issues of institutional support for a large digital humanities project, and then propose collaboration with the university library as the only realistic option for long-term sustainability in our environment. We believe that our experience is typical of the situation for other projects, large and small, that many digital humanities faculty now face at their institutions, and therefore that our experience is also typical of the demands that will be placed on libraries to sustain faculty digital research for the long-term.

As for many digital projects, the Linguistic Atlas Project (LAP) began with computing resources located within the research office itself, first personal computers and later servers. When the university created a research computing service (as an addition to the institution's administrative and instructional services), LAP was one of the first clients--the editor of LAP was even asked to help design the service. However, over the course of several years the funding structure for the research computing service changed from an essentially institutional budget with additions from externally funded research, to a fee-based service much more dependent on research with annual external funding. This meant that humanities projects like LAP, while not excluded from the research computing service, either needed to find consistent external funding or hope for sufferance from the paying customers. Neither of these options appeared