

Results of CLiMB Advisory Board Meeting **June 22, 2007**

1. Executive Summary

On June 22, 2007, the Computational Linguistics for Metadata Building research project held its Annual Advisory Board Meeting at the Morrison-Clark Inn of Washington, D.C. Librarians, museum curators, computer scientists and information professionals from across the nation gathered to review and provide feedback on our results. During this meeting, CLiMB team members provided an overview of the project and reported on accomplishments in the following areas:

- **Image and Text Collections:** Refinement of selection criteria and characteristics of current collection sets
- **Demonstration of Toolkit:** Current interface and functionality of the CLiMB Toolkit; live demonstration of CLiMB
- **Disambiguation and Thesauri:** Statistical success rate for selections of appropriate word sense from the Art and Architecture Thesaurus
- **Semantic analysis:** Method and design of an algorithm for automatically assigning selected semantic categories to a body of text processed by the Toolkit
- **User studies:** Findings from observations, interviews, and written surveys that examined cataloger workflows and term preferences for assigning subject descriptors.

Following each presentation and in a brainstorming session toward the end of the day, attendees were invited to provide feedback on each of these topics. Selected contributions made from that open discussion are provided below.

2. Contributions from Attendees in Discussion Session

Image and Text Collections

Highlights: Leona Faust, Head of Technical Services for the United States Senate Library, confirmed interest in participating in the CLiMB cataloger studies. This gives us a well-rounded beta test site, local to the Washington area with major public access needs within which to test the interface and functionality of the Toolkit.

Marcia Bates and Joan Stahl recommended additional sites that would enable us to broaden the scope of our results to reflect variations in practice across different types of institutions. Furthermore, Joan Stahl pointed out that catalogers in the Registrar's office of a museum may describe an image differently than someone working in the education department because their audiences are different.

Bill Ying of ArtStor brought up that by extending our scope to include some of the individual collections maintained by ARTstor, CLiMB could test the limits of the text-image collection approach. These collections cover a range of topics and could provide a broad perspective on the types of information available for extraction for different areas of art history. He

suggested that follow-up with ArtStor would help to clarify how CLiMB can have the most benefit for catalogers, and for end users. Jack Sullivan added that we have the potential to integrate multiple resources through CLiMB's text mining and data exporting functions. In particular, he suggested making use of existing relationships between image collections for teaching and innovative text resources that bridge thesauri and scholarly literature, such as Therese O'Malley's recent illustrated landscape architecture dictionary.

John Unsworth suggested utilizing electronic texts that are available through the web. He pointed out that many catalogers do not have electronic copies of scholarly writings readily available for mining and may find useful texts describing specific images on the web.

Toolkit Interface and Functionality

Highlights: The redesign of the CLiMB Toolkit--which is now Java compliant, easily downloadable, and promotes open access--was well-received by attendees. In the case where resources are licensed, each institution remains responsible for maintaining their own licenses, but a clear statement of what needs to be done is now available from the CLiMB team and facilitates implementation of the Toolkit at any institution.

Neal Johnson and Brian Schottlaender provided insightful suggestions for refining details of the existing functionality of the Toolkit. Schottlaender recommended adding a default mouse behavior that would enable selecting and dragging descriptor candidates into the export window. Johnson pointed out that there are a number of metadata schemas that define how dates should be presented and noted that incorporating these rules into the Toolkit could facilitate the current export process even further. John Unsworth made several suggestions related to the interface which would reflect and possibly enhance existing cataloging practices.

Catherine Plaisant commented that now that the initial functionality is complete, we have the potential to not only test different user interface designs but to also initiate and test for intuitive workflow interaction for both catalogers and end users.

Disambiguation and Thesauri

Highlights: Marcia Bates pointed out the potential of incorporating frequency analysis into the disambiguation algorithm. She noted that this would enable us to increase the likelihood of automatically selecting the correct sense of a given term. In addition, Bates pointed us to BIOSIS (a medical thesaurus) and other resources that we are currently reviewing to expand the functionality of integrated thesauri within the Toolkit.

Susan Schreibman reminded us that there is an inherent relationship and mapping between images and texts that should not be overlooked. She pointed out that we could utilize these relationships to link together multiple CLiMB-enhanced records to address the one-to-many relationships that exist between so many works and images of works. We are now investigating how to not only make use of the principles of inheritance available through the integrated thesauri but also how to utilize the principles of inheritance which apply to a group of works or even an individual work. For example, we are looking at how frequently terms

which apply to a given structure, such as a building, will apply to its parts, such as a door. One of our next steps will be to incorporate a feature on the Toolkit interface that will enable catalogers to “broaden or narrow” their description based on terms selected for related works.

Semantic analysis:

Highlights: John Unsworth initiated a discussion on the application of the semantic categories at the sentential level versus the paragraph level. As a result, we are investigating similar research on the scope of linked semantic categories, particularly the *BECHAMEL Markup Semantics Project* based at the Graduate School for Library and Information Science at the University of Illinois, Urbana-Champaign.

User studies

Highlights: Carol Mandel and Lauree Sails noted that CLiMB has the potential to enhance current practices and techniques in more ways than one, especially if we are careful to avoid hardening non-optimal practices that are already in existence. We are re-examining our initial findings from the cataloger studies to identify potential pitfalls of cataloging methods. Additionally, we are looking for new ways to present candidate terms that occupy the same semantic space as opposed to the more traditional--and less informative--alphabetical arrangement.

Kari Kraus pointed out that there are key differences between how catalogers and end users approach images. She reminded us that by focusing only on the catalogers, the end user perspective becomes lost.

Angela Giral suggested picture researchers as a potential end user group to consider for the studies. These researchers are experts in finding pictures for a variety of uses such as advertisements and illustrations. Jeff Cohen commented that once we have gathered data from this group and others, we could develop an elitist folksonomy to correlate with the types of search strategies currently being employed by end users.

3. Conclusions

The CLiMB team appreciated concrete feedback and next steps, summarized as:

- **Image and Test Collections:**
 - i. The Senate collections (museum, libraries) wants to explore becoming a beta-test site.
 - ii. ArtStor will follow up with CLiMB to test use of subcollections.
- **Demonstration of Toolkit:**
 - i. The new downloadable and easy-to-use toolkit encourages wider use.
 - ii. The next steps involve cataloger interaction for ease of use.
- **Disambiguation and Thesauri:**
 - i. CLiMB should consider incorporating frequency analysis.
 - ii. Developing evaluation metrics will ensure better accuracy and adoption by catalogers.
- **Semantic analysis:**

- i. Methods to analyze feature inheritance will be explored.
- ii. Application of features to text units of different length using machine learning will help in user studies to evaluate accuracy.
- **User studies:**
 - i. Studies with cataloger users confirm CLiMB hypothesis about utility.
 - ii. Next steps involve closer integration into cataloger workflow in order to be able to test the Toolkit for wide use.

4. Attendees

CLiMB team

Eileen Abels, Drexel University
Joan Beaudoin, Drexel University
Jeff Cohen, Bryn Mawr College
Angela Giral, Avery Art and Architecture Library, Columbia University (retired)
Xiaoli Huang, University of Maryland, College Park
Judith L. Klavans, University of Maryland, College Park
Jimmy Lin, University of Maryland, College Park
Rebecca Passonneau, Columbia University
Carolyn Sheffield, University of Maryland, College Park
Tandeep Sidhu, University of Maryland, College Park
Jack Sullivan, University of Maryland, College Park
Tae Yano, Columbia University

Advisory Board members

Marcia Bates, University of California, Los Angeles (retired)
Leona Faust, United States Senate Library
Neal Johnson, National Gallery of Art
Kari Kraus, University of Maryland, College Park
Carol A. Mandel, New York University
Lauree Sails, University of Maryland, College Park
Brian E. C. Schottlaender, University of California, San Diego
Susan Schreibman, University of Maryland, College Park
Joan Stahl, University of Maryland, College Park
Scott Strong, United States Senate Library
John Unsworth, University of Illinois, Urbana-Champaign
William Ying, ARTstor

5. Regrets

CLiMB team

Laura Jenemann, Drexel University

James Masciuch, University of Maryland, College Park

Dagobert Soergel, University of Maryland, College Park

Advisory Board members

Joseph Danks, University of Maryland, College Park

Allison Druin, University of Maryland, College Park

David Fenske, Drexel University

Neil Freistat, University of Maryland, College Park

Jon Furner, University of California, Los Angeles

Chris Higgins, University of Maryland, College Park

Sheila Intner, University of Maryland, College Park

Matt Kirschenbaum, University of Maryland, College Park

Charles Lowry, University of Maryland, College Park

Clifford Lynch, Coalition for Networked Information

Tom Moritz, Getty Research Institute

Therese O'Malley, National Gallery of Art

Doug Oard, University of Maryland, College Park

Jenny Preece, University of Maryland, College Park

Philip Resnik, University of Maryland, College Park

Elizabeth Barlow Rogers, Bard Graduate Center (retired)

James Shulman, ARTstor

Martha Nell Smith, University of Maryland, College Park

Megan Winget, University of Texas