Abstract of presentation at CNI, April 2005, Washington, DC

Squeezing Metadata Out of Scholarly Texts: Extending the CLiMB Project

Judith L. Klavans, University of Maryland
Marilyn Domas White, University of Maryland
Angela Giral, Columbia University (ret.)

This project briefing will consist of two parts: first, we will review past progress from the Mellon-funded CLiMB project initiated at Columbia University at the Center for Research on Information Access (CRIA) within the Columbia Libraries. CLiMB refers to new techniques using Computational Linguistics for Metadata Building. The goal of the CLiMB project is to extract automatically potential subject descriptors from text written about images. In phase one at Columbia University, known as CLiMB-1, we succeeded in establishing the criteria for potential collections, in building an initial CLiMB toolkit prototype, in extracting information from a few sample collections, and in performing several initial evaluations of the toolkit for catalogers. Our initial results demonstrated that the approach has clear potential, but that it is necessary to carefully filter terms and phrases, and then to link them to existing vocabularies. We also showed that the nature of the collection and the text describing the images can radically affect results. Particular challenges of this approach are determining which segments of texts refer to which images and then relating key phrases to the image or the object depicted in the image.

The second part of the briefing will present the goals and desired outcomes for CLiMB-2, which is currently in planning stages at the University of Maryland, College of Information Studies (CLIS). Building on the success of proof of concept from CLiMB-1, the second phase of the CLiMB project will extend the CLiMB toolkit, build a cataloger's workbench, and test the techniques in a more robust way. In addition to evaluating usefulness for catalogers, CLiMB-2 will incorporate results into an image access platform to be able to test the impact of CLiMB generated terms on searching by several types of user groups, from the image professional to the knowledgeable, but perhaps lay, user. Although reference to existing related thesauri exists in the CLiMB-1 platform, the full incorporation and integration of these valuable resources into the filtering process has yet to occur.

To achieve our goals in CLiMB-2, we are in the process of establishing several partnerships with museums and libraries. Partnering will enable us to work directly with collections and user groups (both catalogers and end searchers) of relevance to the project. Partners are being selected for their strengths in areas such as: image collections, image cataloging, metadata schema creation,

computational linguistics, user studies, or complex image access issues in general.

The focus of this briefing will be on collections and user groups, rather than on computational linguistic techniques to be tested, although discussion of methods planned will be presented. In addition to more classical text analysis techniques, such as named entity identification, noun phrase chunking, and term and phrase extraction, we plan to test machine learning techniques to address issues such as text segmentation and discourse reference. We also plan to research ways to extract definitions from text, and fold them into existing authoritative vocabularies. Finally, we will examine ways that collaborative work spaces can be established given user groups accessing collections using CLiMB terms.

Webpages for CLiMB can be found at:

- CLiMB-1 : www.columbia.edu/cu/libraries/inside/projects/climb
- CLiMB-2: www.umiacs.umd.edu/~climb

In its new home at the University of Maryland, the CLiMB-2 team consists of faculty from the College of Information Sciences (CLIS), including Judith Klavans, co-PI College of Information Studies (CLIS), University of Maryland (UMD); Marilyn White, co-PI, CLIS at UMD; Eileen Abels, CLIS at UMD; Jimmy Lin, CLIS at UMD; and Dagobert Soergel, CLIS at UMD. Also included are two members of CLiMB-1: Angela Giral, former director of Avery Architectural and Fine Arts Library, and Rebecca Passonneau, Department of Computer Science, Columbia University. Close connections with the University of Maryland Libraries is ensuring that both the libraries, information science and computer science perspectives are represented.