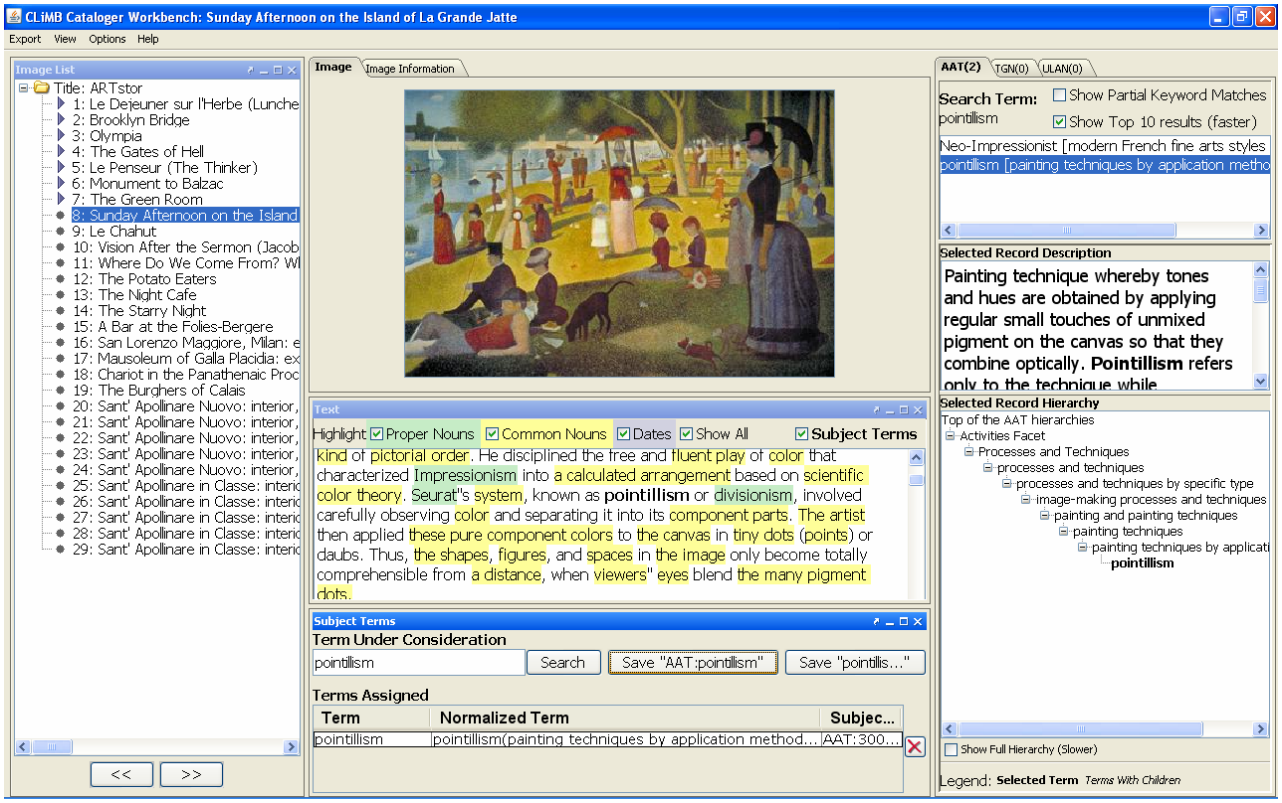
Computational Linguistics for Metadata Building

Final Report
2004-2008



Screenshot of CLiMB Toolkit for the ARTstor *Art History Survey Collection*

Principal Investigator: Judith L. Klavans, Ph.D.
Project Manager: Carolyn Sheffield

Project Team:
University of Maryland: Mairead Hunter, Tatyana Lavut, Jimmy Lin, Tandeep Sidhu, Dagobert Soergel, Rachel Wadsworth, Deborah Wallace

Drexel University: Eileen Abels, Joan Beaudoin, Laura Jenemann, Dwight Swanson

Columbia University: Tom Lippincott, Rebecca Passonneau, Tae Yano

University of Maryland, College Park

Executive Summary

In this report, we describe the CLiMB Toolkit, including system specifications, the results of our usabilty and work flow experiments, and descriptions of disambiguation and categorization tools.

Over the course of the second phase of the CLiMB project (CLiMB-2), we have implemented refinements to the Toolkit developed in the first phase (CLiMB-1). Changes were executed in an iterative fashion and influenced by:

- User feedback
- Expert recommendations
- Iterative testing
- Developments in the field

An overarching goal throughout the project has been to develop a mechanism for the description of works of art that enhances access and exceeds keyword extraction. A frequent reaction from the image indexing community has been to compare our work to Google Image and similar image keyword-based search engines available for general use. Our goal throughout the project has been to enrich the capabilities of such search and retrieval mechanisms through the integration of existing image indexing standards from the museum and library community and to develop and utilize refined computational linguistic techniques.

We exceed simple keyword search by:

- Consulting controlled vocabularies and enabling hierarchical query expansion
- Incorporating a process for the semi-automatic disambiguation of terms using hierarchical thesauri to facilitate description
- Categorizing text segments by their semantic function to further enable query refinement and to assist with disambiguation

1. Project Background

The Computational Linguistics for Metadata Building (CLiMB) project was initially funded at Columbia University to the Center for Research on Information Access. A description of the project and information on results of the first phase can be found at www.columbia.edu/cu/libraries/inside/projects/climb/. In sum, this initial project (now referred to as CLiMB-1) enabled a proof-of-concept, an implemented client-server prototype, and an initial set of resources for testing with the toolkit. The second phase of the project enabled a full implementation and testing with users; this second phase (CLiMB-2) www.umiacs.umd.edu/~climb/ is the topic of this report. The next phase will involve further user evaluation, and the association of social tags and trust with terms from CLiMB into a platform we have termed T^3: Tags, Terms, and Trust.

2. The CLiMB Toolkit

Over the course of CLiMB-2, the Toolkit has evolved into a packaged, downloadable application that reflects the image cataloging workflow we observed at museums and academic visual resource centers. We iteratively improved the Toolkit design through an ongoing literature review, user studies, and outreach efforts.

Highlights of our accomplishments include:

- Reimplementing the Toolkit in a client-side Java-based architecture
- Packaging the Toolkit for easy public download
- Improving overall usability and interface design based on user feedback
- Developing disambiguation and categorization techniques to enrich the description and search processes
- Determining future directions for enabling Toolkit use in repositories based on cataloging practices

Each of these accomplishments is discussed in more detail in the body of the report.

2.1 System Specifications

The CLiMB-1 prototype Toolkit and testbed collections operated on a server-based platform too slow for practical real-world implementation, particularly in digitally-driven image access environments. One of the major goals accomplished in CLiMB-2 was to redesign the Toolkit in a client-side architecture to speed up the download and use of the software, while preserving and extending its original functionality. We chose a Java-based platform. Java provides a rich API to develop fully-featured client-side applications with responsive graphical user interfaces. The cross-platform nature of Java also simplifies many of the issues associated with distribution and deployment of

applications. Additionally, the Toolkit and a public domain testbed collection are available for download as a set of JAR files which require minimal storage space on a personal computer, thereby limiting the burden of several server calls by multiple users. The Toolkit, along with one sample collection, is now available for public download from the CLiMB website (www.umiacs.umd,edu/~climb).

2.2 Workflow Studies and Usability Testing

Our goal was a CLiMB-2 Toolkit interface that was intuitive, required a minimum of training, and integrated smoothly with existing cataloging practices and standards. We conducted both workflow and usability studies with this aim.

Workflow studies conducted over the course of the project examined existing cataloging practices at four different institutions. We interviewed visual resource curators and museum registrars to identify indexing practices and to investigate the process flow of image cataloging in place at each institution; we also collected opinions on the perceived strengths and weaknesses of existing metadata standards. We observed participants as they cataloged a series of images in their own workspaces to better understand user needs and to establish the image indexing workflow into which CLiMB will be integrated.

Usability testing involved observations of image catalogers and museum registrars using the Toolkit. Each participant used the Toolkit to assign subject terms to a given image and then to an image of their choosing. We were surprised and pleased to find out that, despite receiving no instruction in the use of the system, participants discovered the most important features of the Toolkit on their own and immediately went to work assigning terms to image records. They clearly indicated interest in using the tool, both for incoming works to be cataloged as well as retrospective indexing to enrich subject access and clean up legacy records.

Highlights from the workflow and usability studies are summarized below.

2.3 Importing Collections

As a tool designed for repositories providing access to authoritative and documented image collections for the study of art history, CLiMB requires three inputs: (1) digital images to be uploaded along with (2) basic metadata (Title, Creator Name, etc.) and (3) digital texts (expository, prose) about those images in digital format. From our workflow studies, we learned that a range of metadata standards are implemented in different institutions, including locally developed schemas. We provide import options for the most frequently referenced metadata schemas referenced on the professional listservs (VRA CCO, MARC) as well as more general formats such as text and xml. Texts should be clearly associated with specific images and marked up in TEI code. Detailed instructions for preparing materials for ingest are provided in the User Manual found in Appendix A.

For testbed purposes, we assembled six image-text sets (see Table 1) plus one additional image collection that we associate with dispersed texts located online. From our workflow studies, we observed that image catalogers consult a range of materials depending on their knowledge of a particular work and of relevant sources. In future work, we are interested in facilitating the import process so catalogers can bring in relevant and authoritative texts from online resources, import them into the Toolkit, and mine them for subject terms. Examples of other authoritative, online textual resources of interest from the art and architecture domain include:

- Grove Dictionary of Art
- Bibliography of the History of Art
- ArtIndex and ArtIndex Retrospective
- Avery Index to Architectural Periodicals
- ArchNet
- Individual museum websites

As with all of our text resources, to prevent a breach of copyright or intellectual property rights, only the assigned terms and source citation for online resources would be presented to end users.

Table 1. Testbed of Image Collections and Associated Texts

Image Collection	**Associated Text**	**Image/Text Relationship**
National Gallery of Art (NGA) Online Collection	Narratives associated with images on the NGA Web site	Pre-established
U.S. Senate Museum	Catalogue of Fine Arts	Pre-established
Art History Survey Collection (AHSC), ARTstor	Janson and Gardner	Needs to be established
The Vernacular Architecture Forum (VAF)	VAF Field Guides:	Pre-established
The Society of Architectural Historians (SAH): World Architectural Survey and the American Architectural Survey	*Buildings Across Time: An Introduction to World Architecture*. Marian Moffett, et al	Needs to be established
Landscape Architecture Image Resource (LAIR)	*Landscape Design: A Cultural and Architectural History* Elizabeth B. Rogers	Needs to be established

2.4 Examining Images

Since visual resource curators and museum registrars typically work from an image list, we included a scrollable image list as a key component of the layout. Catalogers especially appreciated the indication of which images had already been worked on as this helped orient them on their progress.

Flexibility of the layout was paramount for our users. Catalogers appreciated the system's support of their multiple workflows via multiple methods for transitioning between detailed views of the image, detailed views of the metadata, and combined views. Our users also preferred a large view of the image, even at the expense of other screen elements. They requested the option to view the image in its larger size while annotating, and also to view the existing metadata assigned to the image. Catalogers also suggested that future versions might allow the cataloger to change the way existing metadata and newly added metadata are displayed in the interface vis a vis the image itself.

2.5 Reviewing Texts

The CLiMB Toolkit extracts nouns and noun phrases, the parts of speech traditionally used for indexing and thesauri.[1] We search for these nouns and noun phrases in the combined Getty Vocabularies[2], a domain-specific resource for assigning normalized terms to works of art. The cataloger then selects the best match from the results.

As shown in the screenshots in the user guide, our display uses colored highlighting to distinguish between common nouns, proper nouns, and dates extracted from this text. This allows the cataloger to quickly determine which of the Getty Vocabularies to consult: The Art and Architecture Thesaurus (common), the Union List of Artist Names (proper), or the Thesaurus for Geographic Names (proper). In response to requests for a simplified interface, we limited the use of highlighting and provided the option to turn highlighting off completely. We also removed the highlighting from terms as they were added to the record, instead displaying them in bold text. This simplified the display while helping catalogers track their progress through the text.

The system is flexible and supports a variety of workflows. We observed some catalogers indexing term-by-term by reading the text, picking a term, mapping it to a thesaural resource, returning to their place in the text, picking another term, and repeating. Others preferred to read the entire text, create a list of terms, and then go back and map the list to

[1] According to the ANSI/NISO Z39.19-2005. *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, an indexing term is defined as "The representation of a concept in an indexing language, generally in the form of a noun or noun phrase. Terms, subject headings, and heading-subheading combinations are examples of indexing terms. Also called descriptor." (p. 18) Further, "The grammatical form of a term should be a noun or noun phrase. Verbs expressed as infinitives (without "to") or participles should not be used alone as terms. Activities should be represented by nouns or gerunds." (p. 37).

[2] www.getty.edu//research/conducting_research/ vocabularies/

a thesaural resource all at once. We also observed variation in the types of terms added to the image record. Some highlighted proper nouns and dates only, and noted that terms in the title were not often included in text. Others primarily assigned terms that were not already in the text, suggesting that if terms were in the text already, a search should retrieve them that way. This variation also appeared at an interaction level: some users favored typing terms out, while others preferred to click to select terms in the text. Our current interface easily accommodates each of these styles.

In the future, catalogers would like the Toolkit to take spelling and morphological variants into account when highlighting/bolding terms. For example, adding the term "nymph" should cause any instances of "nymphs" to be bolded. Catalogers also requested a fourth tab where they could view results from all three thesauri in a combined format. Catalogers also expressed a desire to occasionally edit a term after adding it to their list of terms to export. Currently, the only way to do this is to delete, retype and then add the term again. Catalogers would prefer to edit directly in the list of terms pending export. They also expressed a desire to annotate free text terms with definitions and scope notes to help guide future use of the term within the collection. Another suggestion was to display all terms added by a cataloging team in an additional term list so that they can be found again and used consistently. Finally, catalogers requested the functionality to search for a specific image in the collection, rather than scrolling through the image list. Since the original architecture was not designed to support search, we did not put this step into the process, but will incorporate it in the future.

2.6 Consulting Thesaural Resources

Word sense disambiguation is generally performed using a particular lexical resource, i.e., a thesaurus, dictionary, or glossary. As shown in Figure 2, the CLiMB Toolkit uses the Getty Vocabularies, a set of well-established and widely used resources for the cataloging and indexing of art, architecture, artifactual, and archival materials. The Vocabularies are comprised of:

- The **Art & Architecture Thesaurus** (AAT), a structured vocabulary for describing art objects, architecture, and other cultural or archival materials. The AAT's structure is comprised of seven major facets from which multiple hierarchies descend. In the AAT, each concept is described through a record containing a unique ID, preferred name, record description, variant names, and other information that relate a record to other records. In total, AAT has 31,000 such records. Of these, 1,400 are homonyms, i.e., ambiguous terms.
- The **Union List of Artist Names** (ULAN), a name authority that includes the given names of artists, as well as any known pseudonyms, variant spellings, and name changes (e.g., married names). The structure of this resource corresponds to the Agents facet of the AAT in that it contains Person and Corporate Body as its primary subdivisions.
- The **Thesaurus of Geographic Names** (TGN), an authority for place names, including place names as they appear in English as well as in other languages, historical names, and names in natural order and inverted order.

We observed that catalogers at some testing sites frequently consult other domain-specific name authority files and would like to see these accessible through the Toolkit. Resources of interest included: ICONCLASS, the Library of Congress Thesaurus for Geographic Names I & II, and the Library of Congress Name Authority Files. The technical manual includes details on how to convert and call licensed copies of authority files for use in the Toolkit.

Catalogers indicated they would like to see the inclusion of more tools for controlling term usage. For example, to normalize a term's usage across the collection, it would be helpful to be able to pull all previous entries of the term. To this end, catalogers would like to know which terms are used, how often, and which form of the term should be considered “preferred” within their collection. One possible approach would be to supply a tree of terms, with indication of how often each was used in a particular collection, for catalogers to browse through. If necessary, the cataloger could drill down to see the records themselves. Catalogers would also like to be able to determine how many times a term has been used as a descriptor across the collection, and what other terms were used in conjunction with it. Reports of this data could be used not only to guide further cataloging, but also to improve or even create thesaural resources.

2.7 Exporting CLiMB Metadata

Current export functionality reflects established standards observed in use at visual resource centers and museums and further supported by use in the community as documented on listservs and advocated by partners and colleagues. Examples include VRA CCO and MARC. It is important to note for future refinements that catalogers requested more flexible export options, possibly in other formats that would integrate with their existing local database structures. Catalogers agree that they would like a batch-export function.

2.8 Usability Improvements

Our usability tests yielded the following usability and functionality requests:

- The initial design included icons for representing the term review functions of the Toolkit, including adding normalized terms, looking terms up in thesaural resources, and adding free text terms. During the usability tests, participants expressed confusion over which icons were to be used for which functions. We replaced the icons with labeled buttons.

- Users would like the Toolkit to include standard editing tools familiar to them from other applications. They requested keyboard shortcuts and an editing menu accessed by right-clicking with options like copy, paste and delete.

- Catalogers expressed uncertainty as to whether the system was actually looking up the term, especially if no results showed. System feedback in the form of an hourglass or a similar convention would alleviate this confusion.

2.9 Disambiguation

Discourse from specialized domains such as art and architecture draws from specialized vocabularies that present special challenges for automatic word sense disambiguation. For instance, everyday terms often have different senses within art historical discourse. Conversely, everyday terms do not always appear in the terminological resources used for indexing images. For example, WordNet[3], a general purpose lexical resource, contains eleven different senses of the term *foot*. A subset of these are displayed below:

- S: (n) **foot, human foot, pes** (the part of the leg of a human being below the ankle joint) *"his bare feet projected from his trousers"; "armored from head to foot"*
- S: (n) **foot, ft** (a linear unit of length equal to 12 inches or a third of a yard) *"he is six feet tall"*
- S: (n) **foot** (the lower part of anything) *"curled up on the foot of the bed"; "the foot of the page"; "the foot of the list"; "the foot of the mountain"*
- S: (n) **animal foot, foot** (the pedal extremity of vertebrates other than human beings)
- S: (n) **foot** (travel by walking) *"he followed on foot"; "the swiftest of foot"*
- S: (n) **foot** (a member of a surveillance team who works on foot or rides as a passenger)

On the other hand, the Art and Architecture Thesaurus (described in more detail in Section 2.6), contains over forty entries that include the term *foot*, a subset of which appear below:

- **feet, ball (ball feet)** (<feet: furniture components>, <feet and foot components: furniture>, ... Components (Hierarchy Name)) [300040600]
- **feet, block (block feet)** (<feet: furniture components>, <feet and foot components: furniture>, ... Components (Hierarchy Name)) [300040601]
- **feet, bootjack (bootjack feet)** (<feet: furniture components>, <feet and foot components: furniture>, ... Components (Hierarchy Name)) [300121599]
- **feet, bun (bun feet)** (<feet: furniture components>, <feet and foot components: furniture>, ... Components (Hierarchy Name)) [300040609]
  feet, onion (onion feet)
  feet, bulbous (bulbous feet)
- **feet (object components)** (<components by general context>, components (objects), Components (Hierarchy Name)) [300233911]

Note that none of the senses in the AAT refer to *foot (measurement)* or to *foot (body part -- human or animal)*, the most common usages in everyday discourse. Rather, the majority of senses (39 out of 42) refer to furniture or object components. This problem is representative of the challenge of disambiguation in a discourse characterized by ambiguous terms whose senses are a small semantic distance from each other. It is

[3] http://wordnet.princeton.edu/

difficult to distinguish between multiple senses of a single term in this domain automatically because the concepts to which they refer are highly specific and very similar.

The granular level of specificity of the AAT offers catalogers a range of semantically close term variants, enabling a high level of query refinement for end users. However, it does not accommodate description of more general concepts depicted in or by a work of art. For this reason, we sought to leverage relationships between the AAT and WordNet for use in disambiguation. In future work, we propose building on existing techniques developed during CLiMB and testing them with other thesaural resources.

We focused the development of our disambiguation techniques on the AAT because its specialized scope and highly detailed sense specificity provide a foundation for an accurate mapping of term instances to senses. To facilitate use of the Getty resources in the Toolkit process flow, we designed a custom Getty Thesaurus browser for the CLiMB project. As in the standard browser provided by Getty, a cataloger can look up subject terms in the Getty Thesauri. The primary difference, and advantage, of the CLiMB browser is that it allows the cataloger to quickly search all three thesauri at once from the same interface.

Our disambiguation process begins by analyzing texts to distinguish between proper and common nouns. We use the output of this analysis to choose between the three Getty Vocabularies, thereby presenting the cataloger with the resource that most likely contains a match. For all common nouns, the Toolkit shows results from the AAT, even if there are more matches in the TGN or ULAN. For proper nouns, the Toolkit favors both the TGN and ULAN over the AAT, even if more matches can be found in the AAT, choosing whichever of the TGN and ULAN has more matches for the term under consideration.

As part of CLiMB-2, we also evaluated more sophisticated disambiguation techniques for common nouns. These are described in detail in Sidhu et al. (Appendix B). The most effective technique, which was integrated with the Toolkit for further evaluation, is described below.

To select the most likely sense, the CLiMB disambiguation algorithm:

1. Analyzes noun phrases to separate modifiers and head nouns.
2. Searches the AAT for all possible matches of the head noun.
3. Searches the entire AAT record of each term in this result set for occurrences of modifiers.

The record with the closest head-term match and most modifier matches is returned as the most likely match.

2.10 Categorization

CLiMB applies linguistic technologies to semi-automatically categorize text segments according to their semantic relationship to the image(s) which they describe. The semi-automatic semantic categorization of texts enables novel methods for enhancing subject access to images by:

- allowing catalogers to specify the types of information they want to locate and assign to images;
- mitigating disambiguation challenges by solidifying the context in which an ambiguous term occurs; and
- facilitating both cataloging and disambiguation by automatically aligning text segments with existing metadata schema.

CLiMB's external categorization techniques apply semi-automatic linguistic technologies to classify text segments about a single image according to their semantic relationship to the image(s) which they describe. As in other computational linguistic research projects exploring automatic categorization, this involves a feature-based approach that distinguishes text segments based on semantic content. A feature-based approach offers a broader view of the text than is possible with only the lexical cohesion techniques applied in CLiMB's segmentation phase. We have identified a set of features pertinent to art history survey texts for use in their categorization techniques. The following section describes these features and their application in the CLiMB categorization process.

We have addressed the following tasks:

1. Identifying text features relevant to classification and development of a set of functional semantic categories (FSCs)
2. Performing machine learning experiments on three of the FSCs
3. Evaluating interannotator agreement and its relationship to machine learning performance.

2.11 Text Features

Many linguistic approaches rely on standard feature sets, based on lexical cohesion, term frequency, and term co-occurrence, which are fed into a probabilistic model. Differences between researchers arise in their choice of particular feature set, probabilistic model, and machine learning tool(s). In contrast, we established a feature set via manual review and analysis of art survey texts. Manually identifying features gave us tighter control over the number of features to be learned and allowed us to develop a clear set of parameters for assigning those features.

Our initial set of features included unigram word counts and hand-selected features drawn from approximately half a dozen semantic domains. For example, we created sets of key words and phrases characteristic of the art history domain, (e.g., *masterpiece*), and other words and phrases deemed discriminative by our experts and annotators, such as terms denoting parts of the human body. We later added more automatically derived

features, including tense information drawn from part-of-speech tagging. Experiments with WordNet's semantic distance metrics, which weigh terms with respect to various semantic domains, such as appearance, physical dimension, and so on, have been initiated to guide future work on this piece.

2.12 Functional Semantic Categories

After segmenting text extracted from art history textbooks into segments that each refer to a single image, we labeled these segments according to its membership in the set of seven categories. Each category represented a functional relation between a body of expository text and an image depicting a work of art.

A larger number of categories would lead to much sparser data; a smaller number would lead to categories that are less distinct. We selected categories that are maximally distinct, based on textual analysis and the input of two subject matter experts (rather than on existing indexing). This approach reflected the content of the collection, i.e. the classes covered all possible cases observed in the two to six paragraphs associated with an image.

In addition we took steps to make the categories easy to assign. From the perspective of designing an annotation task, the ideal set of labels is one that can be learned quickly by the same group from which actual annotators will be drawn. This indicated that annotators understand the criteria for each annotation label and feel confident about making the judgments involved. Furthermore, we also designed an interface so that annotators easily label, data can be collected quickly and then stored in a format convenient for further processing. The less training required, the fewer categories, and the simpler the decision-making process, the better the category set.

Table 2 below shows the seven semantic categories with a brief definition of each. We assigned one of these semantic categories to each text feature.

Table 2. Seven Functional Semantic Categories for Labeling Text Extracts

Image Content	A description of what is depicted in or by the image. Terms in this category may also provide an identification of specific persons, places, or things. For non-representational works and objects, Image Content may include terms which describe the function of the piece. Image Content terms may also describe a particular view including a description of the subject matter within a detail view of a larger work, the vantage point of that particular view, or other descriptions of how the view relates to the work as a whole.
Interpretation	An analysis of the meaning or meanings conveyed by the work as a whole or any of its components. Terms in this category refer to the themes which the artist is trying to convey through the work. Examples: *salvation, original sin, sacrifice, truth, power, innocence, idolatry*.
Implementation	Terms that describe the artistic methods, techniques, or materials a used to create a work. For example, the term chiarrascuro, a technique for representing light and shadow, would fall into this category. Terms indicating how an artist's use of particular materials, tools, or techniques is either unique or conventional may also be categorized as implementation. References to measurements and the work's scale (large scale, miniature, etc.) may also fall under this category.
Comparison	Terms, including titles and artist names, which occur in the context of comparing or contrasting the work to one ore more other works. Points of comparison or contrast may include the work's imagery, technique, subject matter, materials or other characteristics. This category will usually be assigned in conjunction with one or more other the first 5 categories.
Biographic	This category includes the proper names of the individuals (artists, patrons, subjects) involved in the work's creation and may also cover a description of their role. In addition, terms in this category may refer to the artist's/patron's/subject's life or views, including formative influences, or the influence of the these individuals on others.
Historical Context	This category encompasses both the historical context of creation for a specific work (i.e., a commission) as well as the broader social, political, or economic context in which the work was created (i.e., a war, the Great Depression, etc.). The broader context may provide insight into the subject matter and can help provide context for the artists' views on those events. Historical context terms may also refer to how a work has been interpreted over time.
Significance	Terms in this category refer to the aspects or characteristics of a work which distinguishes it as an ultimate example of its type. All the images are significant but this label applies primarily to terms that describe a unique aspect of the work that makes it superlative with respect to that aspect. It is often easy to identify such a term, because it will appear as a superlative like "this is the most. . .", "this is a quintessential . . ." and so on. This category will usually be assigned in conjunction with one or more of the other first 5 categories.

This approach yielded high interannotator consistency which is critical for reliable initial machine learning results.

One difficulty that stems from our choice of functional semantic (rather than topical) categories is that many of the content words were not relevant features, since they were different for descriptions of different images. However, we were able to identify content-independent features such as verb tense, which distinguishes the **Image Content** class from others, and thereby achieve high results on relatively small datasets. We hope that

our functional categories will transfer more easily to texts that are substantially different from our training and test materials than topical categories would.

Classifying text into functional semantic categories provides some control over the selection of metadata. Our categories correspond roughly to categories discussed in the image indexing literature[4]. It should be possible to align our categories with controlled vocabularies used in university visual resource centers, facilitating their use in automatic or semi-automatic cataloging.

2.13 Machine Learning Experiments

To establish a baseline for our machine learning experiments, we began with a classic text classification approach: Naïve Bayes classifiers using unigram word counts to represent the input texts, with separate classifiers for each semantic category. For our experiments, we used three text representation options:

1. Unigram word counts
2. Hand-selected features drawn from approximately half a dozen semantic domains. Input texts were represented using raw counts for these hand-picked features.
3. A combination of hand-picked features with the unigram word count representation.

We experimented with several learning methods, including Naive Bayes (NB), support vector machines (SVMs) , and tree-based learners (TBLs). We also used chi-square independence tests to prune the word vectors in order to reduce the noise in the training data, which smaller corpora such as ours are sensitive to.

NB was the best overall method. The Image Content classifier performed best, presumably due to the relatively greater amount of training data. The results for the Image Content and Historical Context categories are shown in Table 3 below. There were insufficient examples to assess performance over the other categories.

Table 3. Results of Initial Machine Learning Experiments

Category	Best Representation	Best Learning Method	Accuracy

[4] M. Baca. *Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information*. The Haworth Press, Inc., 2003. Available through Library Literature, last accessed July 25, 2006. ; H. Chen. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, pages 260–273, 2001. ; S. S. Layne. Some issues in the indexing of images. Journal of the American Society for Information Science, pages 583–8, 1994.

Image Content	3	Naïve Bayes	83%
Historical Context	1	Naïve Bayes	70%

2.14 Interannotator Agreement Studies

In our studies of inter-annotator agreement, we found that annotators with expertise are much more consistent with each other than non experts. Through our evaluations, we explored other factors that affect human agreement on a semantic labeling task in the art history domain. We conducted four experiments where we varied the number of labels annotators could assign, the number of annotators, the type and amount of training they received, and the size of the text span to be labeled. Using the labelings from one experiment involving seven annotators, we investigated the relation between interannotator agreement and machine learning performance. We constructed binary classifiers and varied the training and test data by swapping the labelings from the seven annotators. From our results, we found that machine learning performance was often quite good despite lower than recommended interannotator agreement. Additionally, we found that the machine learning performance for the functional semantic categories, on average, correlated with the overall agreement among the seven annotators. Finally, we found that learning performance on the data from a given annotator does not correlate directly with the quality of that annotator's labeling. Full details, including recommendations for the use of labeled data in machine learning and implications for large scale corpus annotation projects that deal with similarly subjective phenomena, can be found in Appendix C.

3. Sustainability

We are very pleased to report a recently-funded proposal submitted to the Institute for Museum and Library Services entitled "T^3: Text, Tagging and Trust to Improve Image Access for Museums and Libraries". [5] We proposed a project that would enrich metadata for image collections with both authoritative and unvetted metadata. The authoritative metadata would be generated using CLiMB's cataloger-mediated text mining capabilities. The unvetted metadata would be gathered using a social tagging mechanism developed by the steve.museum project. Furthermore, the proposed system would enable end users to not only submit metadata but to also indicate their level of trust for other sources of metadata, i.e., catalogers and other end users. The deliverable would be a product that enabled metadata population for image collections and encouraged engagement and collaboration among end users of the collections. As a result of this proposal, we developed a strong partnership with the Indianapolis Museum of Art (IMA) and, in anticipation of potential funding, we are working with them to implement the CLiMB system for use in tracking their collections. We have provided Rob Stein, Chief Information Officer at the IMA, with a temporary login and a copy of the CLiMB user

[5] http://www.imls.gov/news/2008/091008a_list.shtm#MD

manual. In return, he will provide periodic feedback as his staff gets up to speed with using the tool.

We have also received a great deal of interest in this proposed tool from a number of potential users including the Senate Museum, the University of Maryland Visual Resource Department and Architecture Library, and JSTOR.

Registrars from the Senate Museum participated in our usability studies and expressed interest in using the Toolkit as part of their collection management efforts. They were particularly interested in using the Toolkit to enhance legacy records containing notably less metadata than the records they currently create. Additionally, they envision the Toolkit as a tool for analyze existing metadata entries and updating them for consistency using the integrated authority files. They are currently in the process of obtaining permissions to install the Toolkit on their machines.

Lauree Sails, Director of the Visual Resources Center at the University of Maryland, has served on the CLiMB Internal Advisory Board and also participated in the usability testing. Her interests in a potential application for CLiMB center on the thesaural browser we developed as part of the Toolkit. In particular, she notes that the display of results and the ease of navigation between tabs for the AAT, ULAN, and TGN noticeably streamline the process of assigning normalized terms to image records. Sails has also initiated discussions between CLiMB and the Architecture Library at the University of Maryland to discuss potential applications for their collections as well.

Through our dissemination efforts, we established new contacts in the digital library community. Representatives from JSTOR approached us during a demo session at the Joint Conference on Digital Libraries (JCDL). We are in discussions with them about using CLiMB's text mining capabilities over a set of exhibition catalogues which JSTOR will be adding to their collections in the near future. This partnership would explore the efficacy of the CLiMB tool for indexing print materials, particularly in the domain of art and architecture.

Other potential future partnerships that we are exploring include the Maryland Historical Trust and the Mellon-funded Architecture Visual Resource Network (AVRN).

4. Appendices

The appendices include the following documents:

4.1 Appendix 1: CLiMB Toolkit User Manual

4.2 Appendix 2: List of CLiMB-2 Presentations

4.3 Appendix 3: CLiMB-2 Publications (full text)

4.4 Appendix 4: Brochures

4.5 Appendix 5: CLiMB-2 Posters

4.6 Appendix 6: Annotated Bibliography

4.7 Appendix 7: Internal Advisory Board Member List

4.8 Appendix 8: External Advisory Board Member List

Appendix 1: CLiMB Toolkit User Manual

Computational Linguistics for Metadata Building

University of Maryland, College Park

User Manual

CLiMB Toolkit v. 1.95

Principal Investigator: Judith L. Klavans
Project Manager: Carolyn Sheffield
Project Team: Eileen Abels, Joan Beaudoin, Mairead Hunter, Laura Jenemann, Tatyana Lavut, Jimmy Lin, Tom Lippincott, Rebecca Passonneau, Tandeep Sidhu, Dagobert Soergel, Dwight Swanson, Tae Yano

Funded by the Andrew W. Mellon Foundation

1. Intended User Group:

The goal of the Computational Linguistics for Metadata Building project is to address the cataloging bottleneck by enabling catalogers to leverage existing text about images. Our technologies address the issue of semi-automatic text data-mining using computational linguistic techniques. As an image cataloger's workbench, the CLiMB Toolkit was created to facilitate the review of terms extracted from scholarly texts about images, normalize those terms using disambiguation techniques and controlled vocabularies, and assign those terms to image records thereby improving access to digital image collections. Although the Toolkit was built for catalogers, this technology can be repurposed for end users, particularly in shared online image repositories and social tagging environments. We have performed onsite testing of the CLiMB Toolkit at several repositories with experienced catalogers and museum curators. The user interface is constantly evolving with feedback from our users. The version of this user manual is indicated in the footer.

2. Preparing Your Input:

The Toolkit requires as input:

- Images
- Minimal metadata about the work shown in the image
- Text associated with the images
- Resources (optional) such as Getty Vocabularies (www.getty.edu)

The Toolkit downloads with one pre-installed collection for trial purposes. Note that the publicly available Toolkit does not include the Getty Vocabularies. To access licensed copies of these or other authority files, follow the instructions provided in Section 2.3.

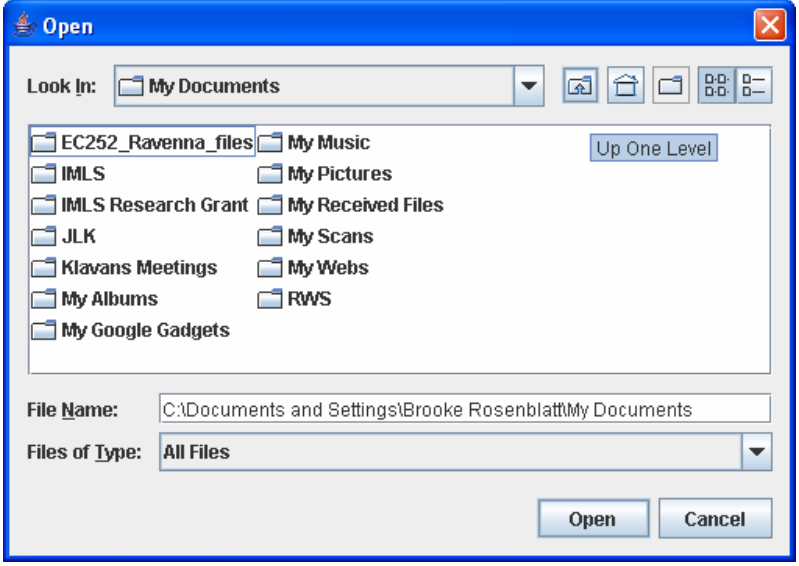
2.1. Preparing Your Images

- Type: Save images as JPEG or GIF
- Dimensions: No restrictions
- Resolution: No restrictions
- Image files must all be saved to one folder

2.2. Metadata and Text

Metadata and the text associated with these images must be in one flat file. Each line of this file corresponds to one image. Fields are separated by <tab> and must be entered in the following order:

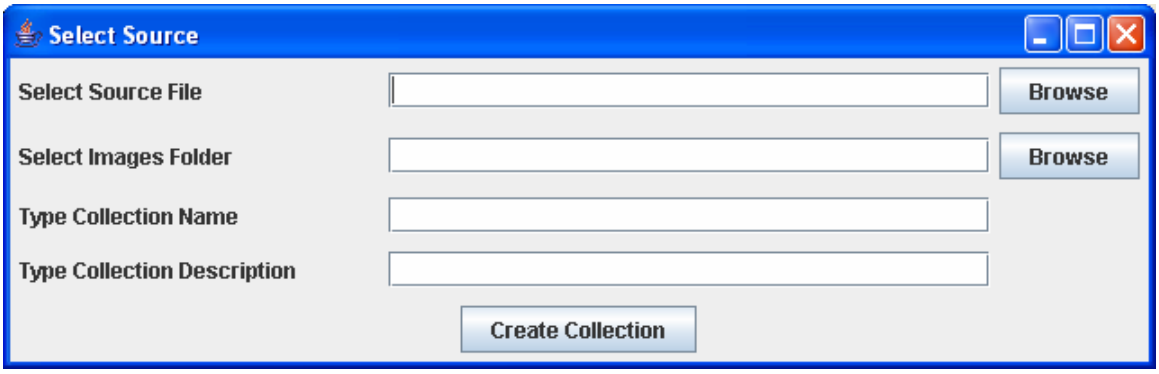
Name of the image file (note: *the image file name must match the image name given in the image's folder*), Title *(this will appear in both the Image List and the Image*



Information frames of the Toolkit), text about the image, Creator, Creator date, Measurements, Medium, Date.

Capitalized fields above correspond to metadata that appear in the Image Information box in the Toolkit. If some of this information is missing, the corresponding field will be empty. For a detailed explanation how to build this file see Appendix 1.

2.3. Resources



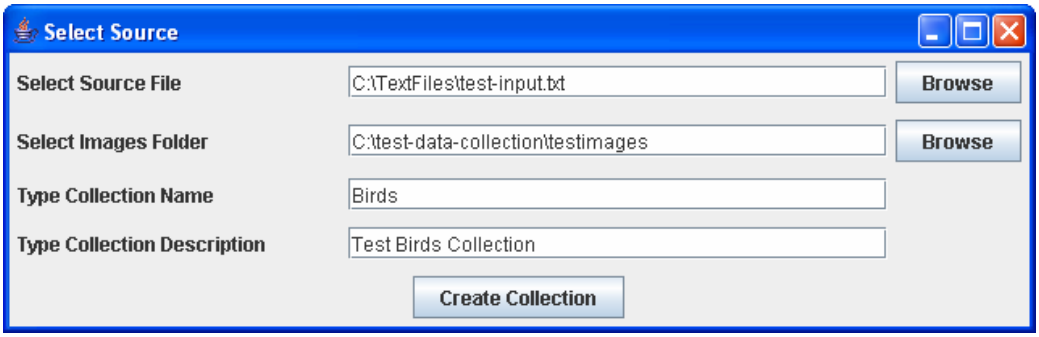
The CLiMB Toolkit does not include any pre-installed ontologies or thesaural resources. To use this feature, you must license the desired resources and call them remotely.

2.4. Activating the Collection

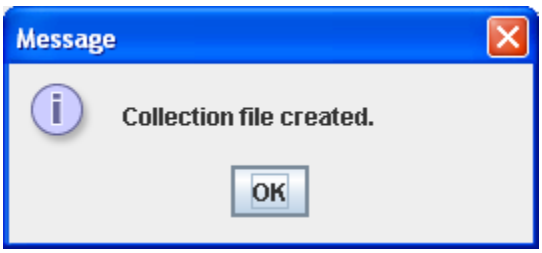
To build the collection click on the “create_collection.bat”, located in the folder where you extracted climb.zip. The configuration “Select Source” window will appear with four fields for entering the information about the collection.

You can type in the paths to the Source File (text) and the Image Folder or you can use the “Browse” buttons. When you press the “Browse” button the pop-up window appears which allows you to navigate to the desired resources.

When all information is entered, press “Create Collection” button.



A pop-up message will tell you when collection is created.



Press "OK". Note that the "Select Source" window will still be open and you can create another collection from here. Replace all the fields with the information for the next collection and press “Create Collection”. When you are finished adding collections, close the “Select Source” window.

The Collection File(s) (collection_name.jar) is placed in the folder where the CLiMB Toolkit jar file is located and this collection can now be opened in the Toolkit. The next time you start the Toolkit this collection will appear in the selection list.

3. Downloading the Toolkit

Visit the CLiMB homepage at http://www.umiacs.umd.edu/~climb/. Navigate to the **Technology and Prototype** page. You will see two buttons to download:

- a soft copy of this manual in its latest version
- the CLiMB Toolkit itself

In order to download the Toolkit, you will need to agree to the open source license. We also request that you register with your name and institution. This information is being collected to help us track users and uses. With your permission, we will also put you in touch with others in the CLiMB user group.

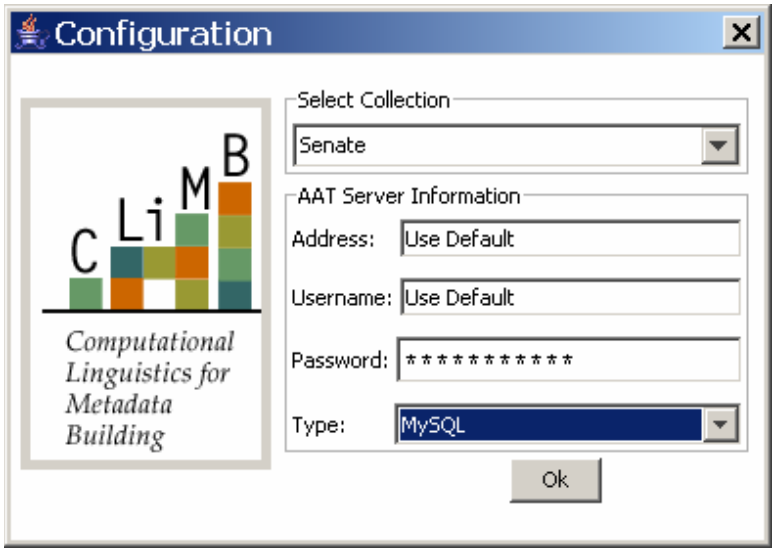
4. The Standard CLiMB Toolkit and Trial Collection

The Toolkit is distributed as a climb.zip file. Extract (unzip) this file and save to a folder, which will contain: CLiMB Toolkit.jar and Senate.jar. The Senate.jar contains images, metadata, and text from the Senate Catalogue of Fine Art. This pre-installed collection is included to help you get started and illustrate the way the system works.

To use the Toolkit with other collections, image, metadata, and text sets will need to be licensed and loaded into the system according to the instructions in Section 2. To fully benefit from the disambiguation and refinement of extracted metadata offered by CLiMB, you will also need to secure a license to the Getty Vocabularies or other authority files and thesauri.

5. Starting the Toolkit

To start the Toolkit, open the folder which contains it and any image collections, and double-click on *CLiMB Toolkit*. A “Configuration” box will pop up.



Click on the “Select Collection” drop-down box and choose from the list of already downloaded and created image collections. Enter the address and login information for your Thesauri databases under "AAT Server Information".

The first time a collection is opened, the Toolkit will extract the collection files to a folder with the collection name. Each subsequent time that collection is opened, the Toolkit accesses the files directly from the unzipped folder and will start faster.

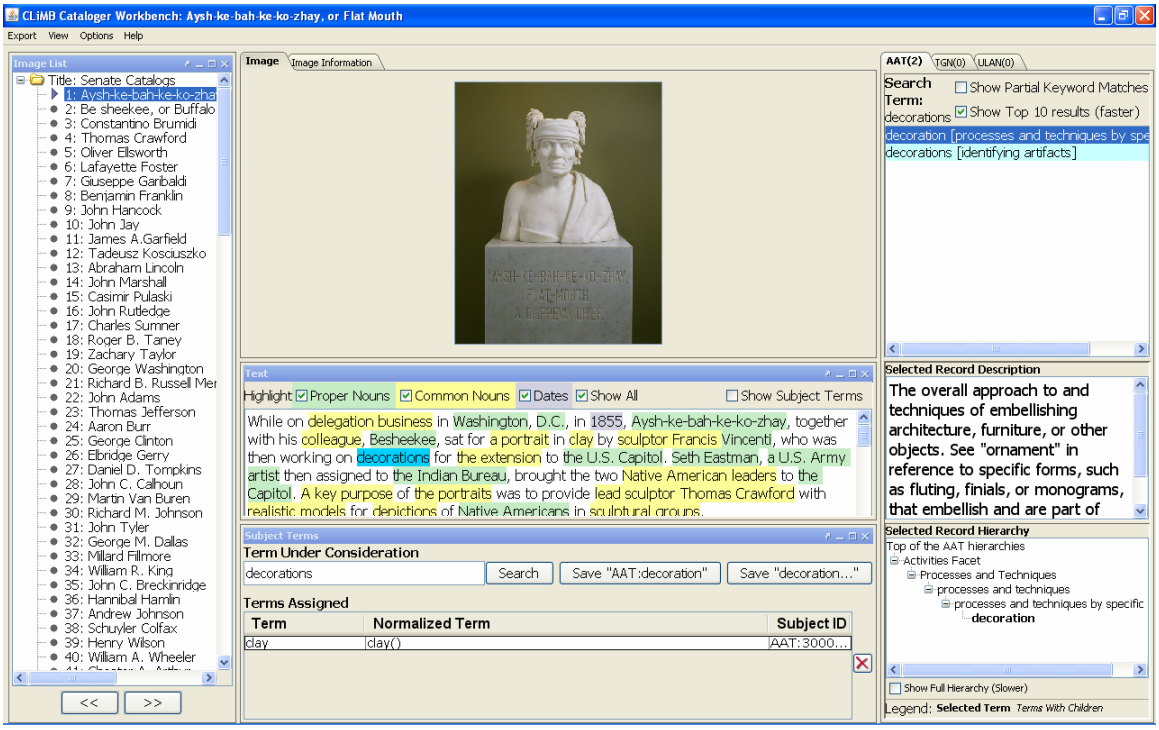
6. Default Layout

On the far left of the CLiMB Toolkit is the list of images included in the collection. Each collection within the CLiMB system includes a set of images with basic metadata and associated text. The image itself will load in the top center panel with the text underneath. The metadata can be viewed by clicking on the tab labeled "Image Information", directly above the image. To switch back to viewing the image in this panel, click on the "Image" tab. Instructions on opening images and their files appears in Section 8.

The panel on the far right contains the Getty Vocabularies, authority files for the description of works of art, architecture, and material culture:

- the **Art and Architecture Thesaurus** (AAT): a structured vocabulary for describing objects and their surrogates. The AAT's structure is comprised of seven major facets (Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, and Objects) from which multiple hierarchies descend.
- the **Thesaurus of Geographic Names** (TGN): an authority for place names, including place names as they appear in English as well as in other languages, historical names.
- the **Union List of Artist Names** (ULAN): a name authority that includes the given names of artists, as well as any known pseudonyms, variant spellings, and name changes (e.g., married names).

The *Subject Terms* panel appears bottom center, beneath the text. This panel is used for reviewing and assigning terms to the image record. Instructions for selecting terms from text and for generating your own terms appear in Sections 11 and 12.

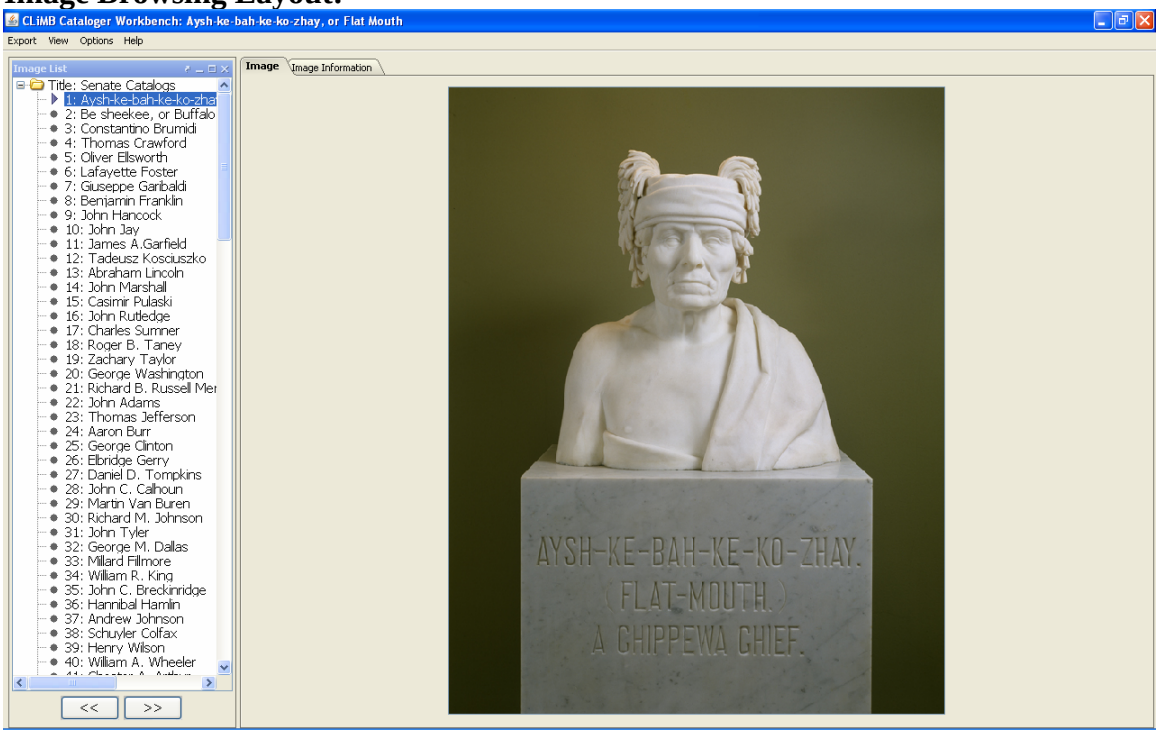


7. Navigating through the Frames

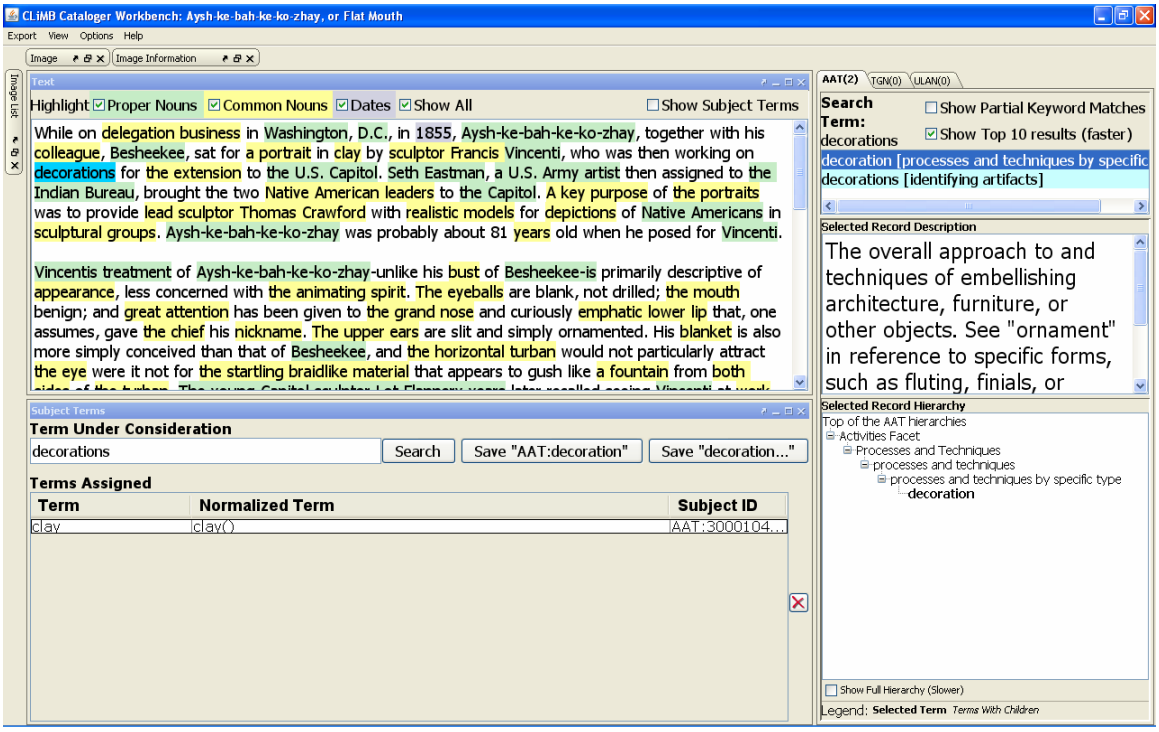
The Toolkit has a flexible user interface which allows the different frames to be moved around and collapsed or closed by using the options shown at the top right corner of each box. To return to the default layout, go to “view” on the top left corner of the workbench, go down to “layout” and then choose “default layout.” There are also two other preconfigured layouts available through this menu:

- "Image Browsing Layout": without Thesauri panel
- "Mapping Layout": with collapsed Image List.

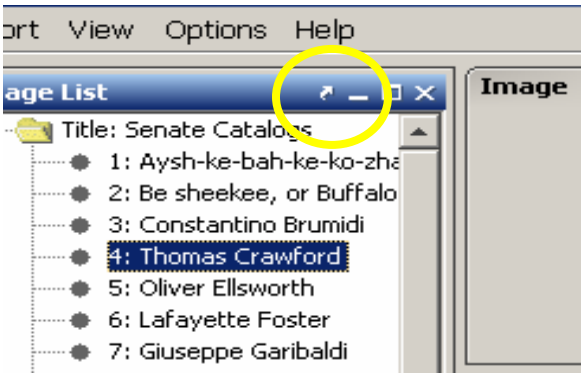
Image Browsing Layout:



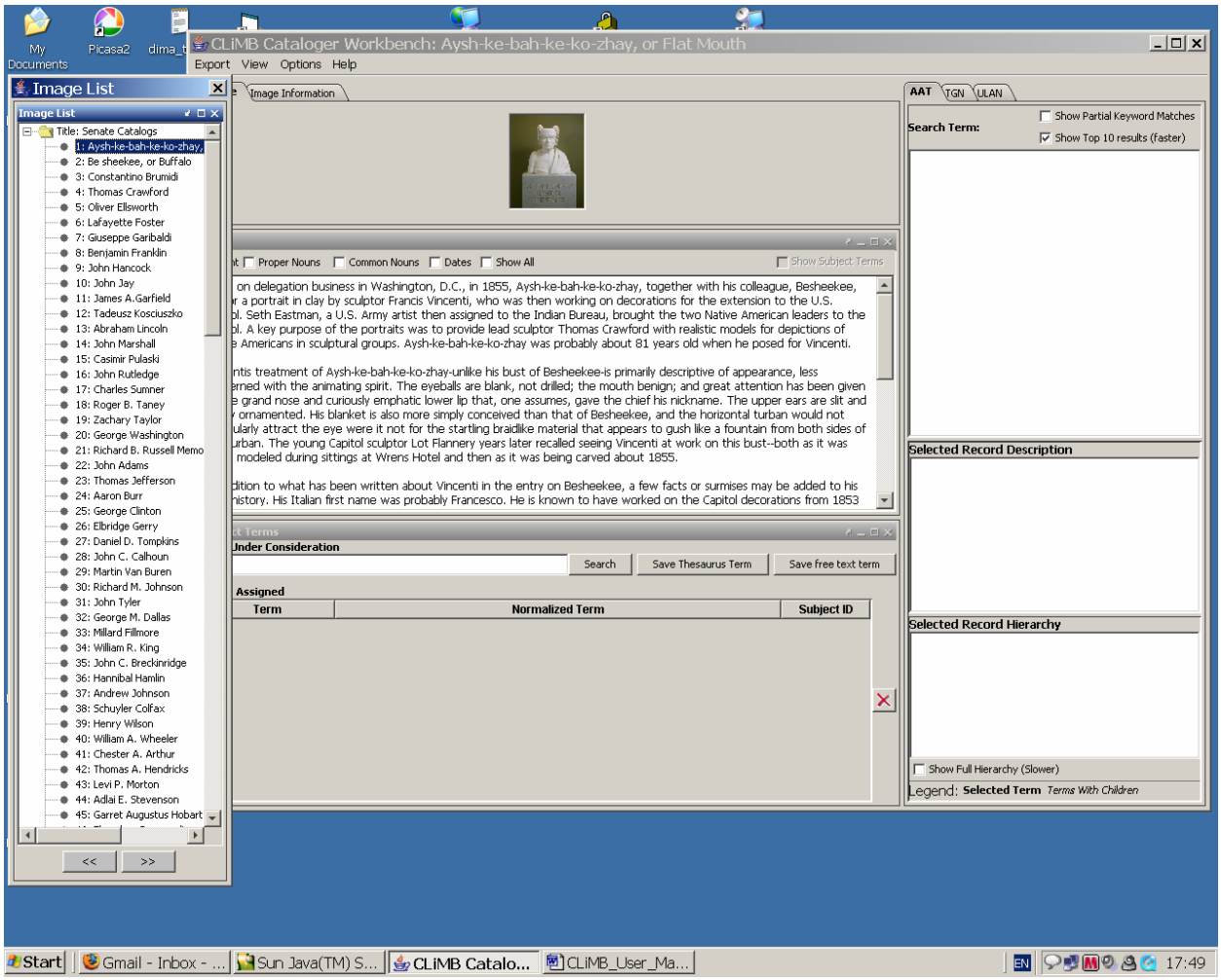
Mapping Layout:



One of the options at the top right corner of each box is a little arrow next to the "-" for docking/undocking.



Clicking on this arrow will "undock" the frame from the workbench, giving more space for other frames. For example, undocking the Image List looks like this:



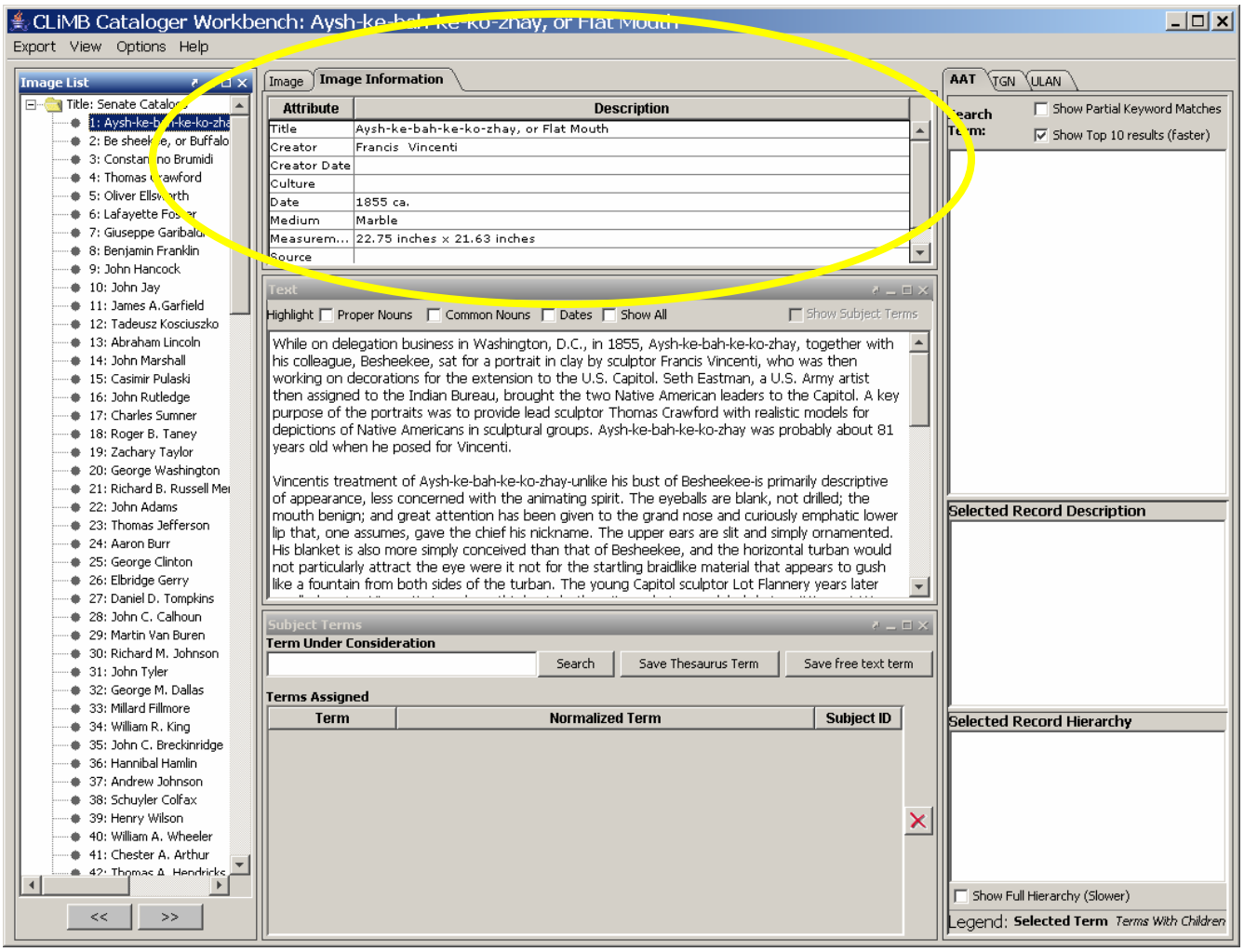
You can navigate through the main and undocked windows separately. To return undocked frame to the workbench, simply re-click the arrow icon.

8. Opening Images, Metadata, and Texts

To load an image, select the title from the far left panel labeled “Image List”. The image will appear in the top center panel of the workbench screen.

To resize the image, move the mouse over the image to enlarge it for a better view. You may turn this option on and off by clicking on "Auto Resize Image" under "Options" at the top left corner of workbench window.

Click on *Image Information* to view basic metadata regarding titles, authors, measurements, dates, etc of the selected image.



To return to the image view, click on the *Image* tab.

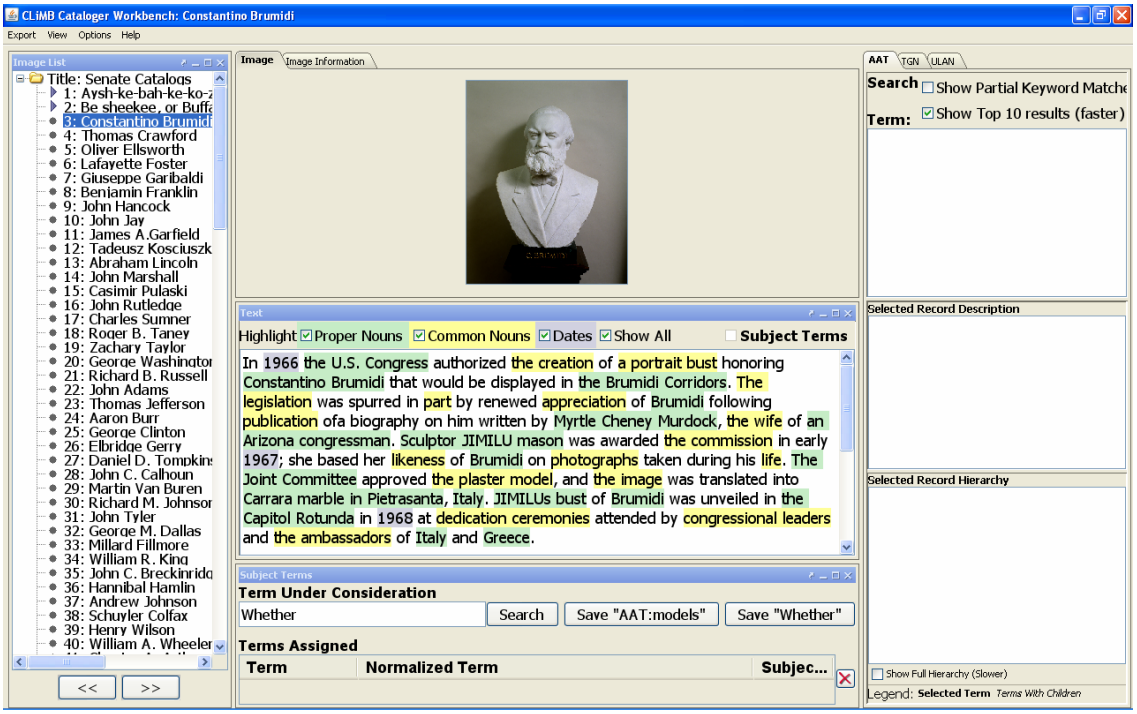
When the image opens, the associated text also loads in the Text panel, directly below the image. To adjust the font, right-click and select the desired font style and size. The keyboard shortcuts for enlarging or reducing font size are “ctrl +” and “ctrl –”, respectively.

9. Viewing Potential Terms

Check the box next to common nouns, proper nouns, or dates to highlight any or all within the text. Highlighting for each of these term types is color-coded for ease of review.

For example, consider the text associated with the bust of Constantino Brumidi in the Senate Collection. Clicking on the boxes next to common nouns and proper nouns would yield yellow highlighting of common nouns and green highlighting of proper nouns, with dates highlighted in purple:

In 1966 the U.S. Congress authorized the creation of a portrait bust honoring Constantino Brumidi that would be displayed in the Brumidi Corridors.

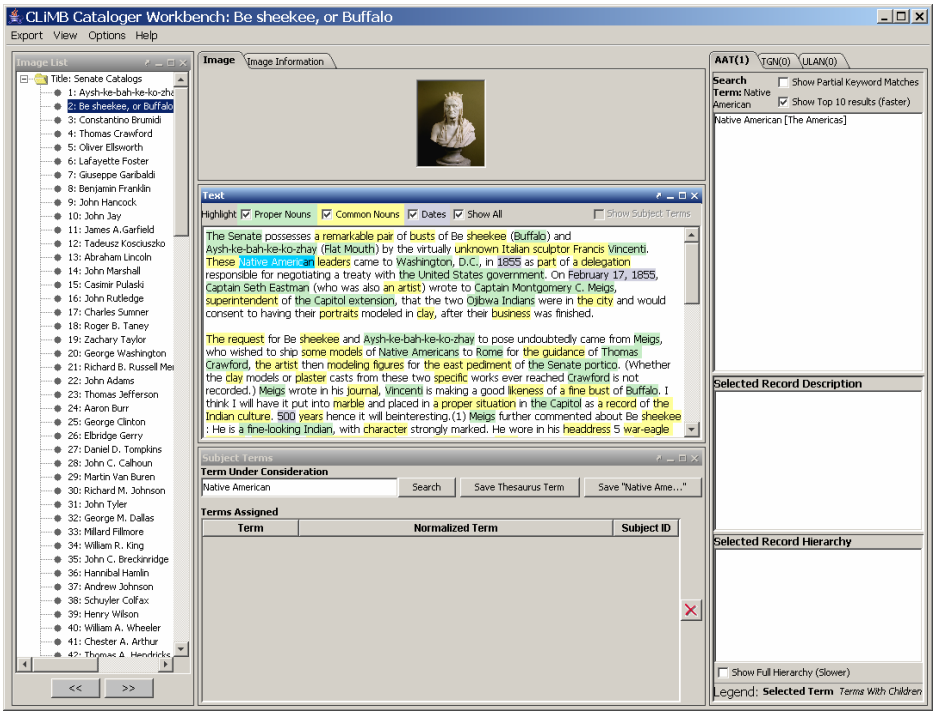


Note that if no dates are mentioned in the text, the “Dates” checkbox will be grayed out and will not be available for selection.

10. Looking up Terms in the Resources

Terms occurring in the text may have one or more matches in the Getty Vocabularies, displayed on the right hand side. To check for matches, click on the term you wish to look up. This term will appear in the “Terms Under Consideration” box and any matches will appear under the tabs marked AAT (Art and Architecture Thesaurus), TGN (Thesaurus for Geographic Names), and ULAN (Union List of Artist Names). To look up a term that does not occur in the text, type it directly into the “Terms Under Consideration” box and click the "Search" button to perform a look-up. Click on the different tabs to see each thesaurus’ entries for the selected word. For example, selecting "**heritage**" in the text below yields 7 results under tab AAT, 2 under TGN and 1 under ULAN and it looks like this:

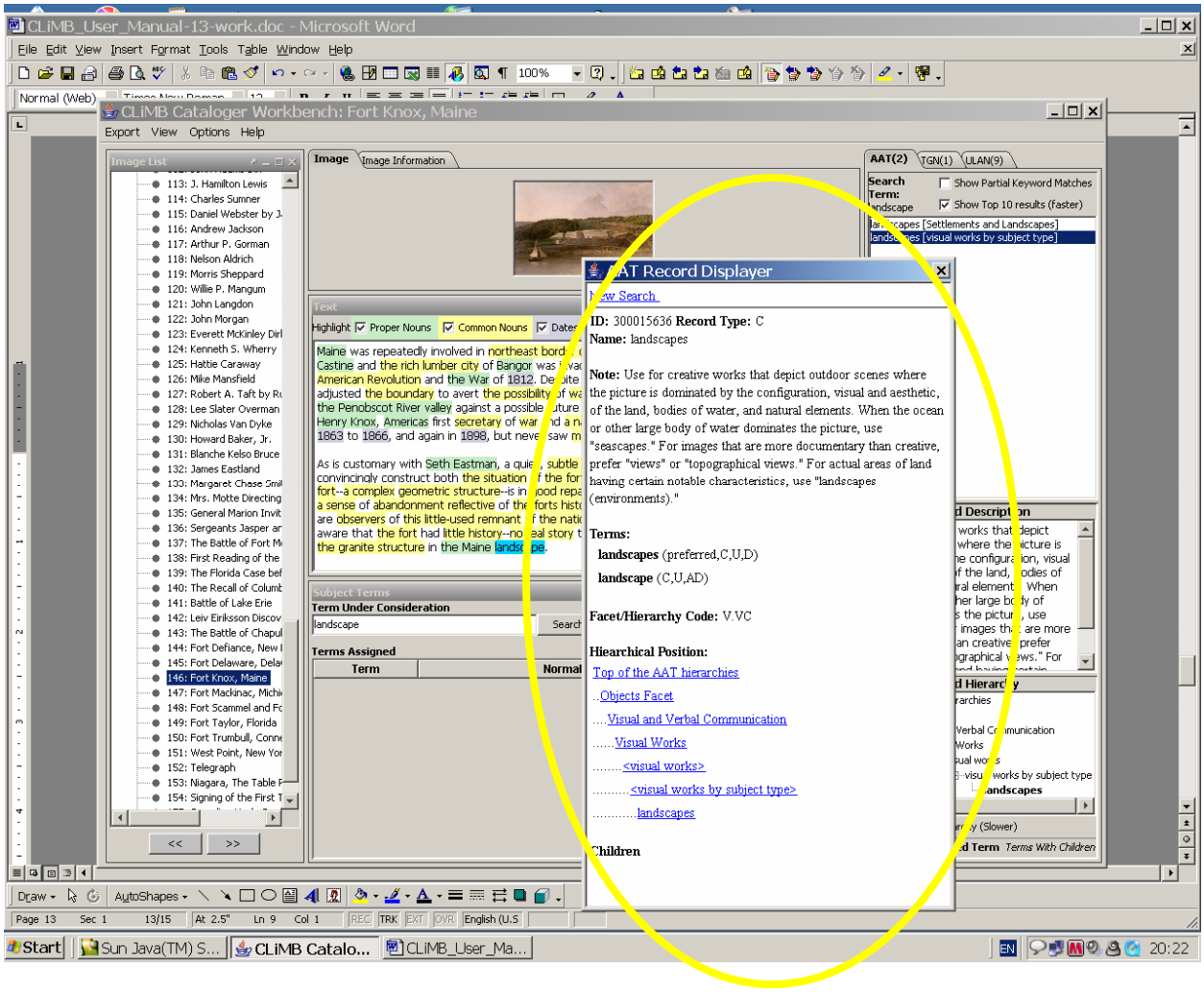
Note that you may select either single words or phrases*. For example, if you select "Native American" the Toolkit highlights both words, looks them up as a phrase in the Getty Vocabularies, and returns 1 result. Both words in this phrase appear in the "Term Under Consideration" box. The screen looks like this:



* ***Note about the Art and Architecture Thesaurus:*** The AAT is comprised of descriptors representing a single concept. Many of these descriptors are single-word terms. When looking up

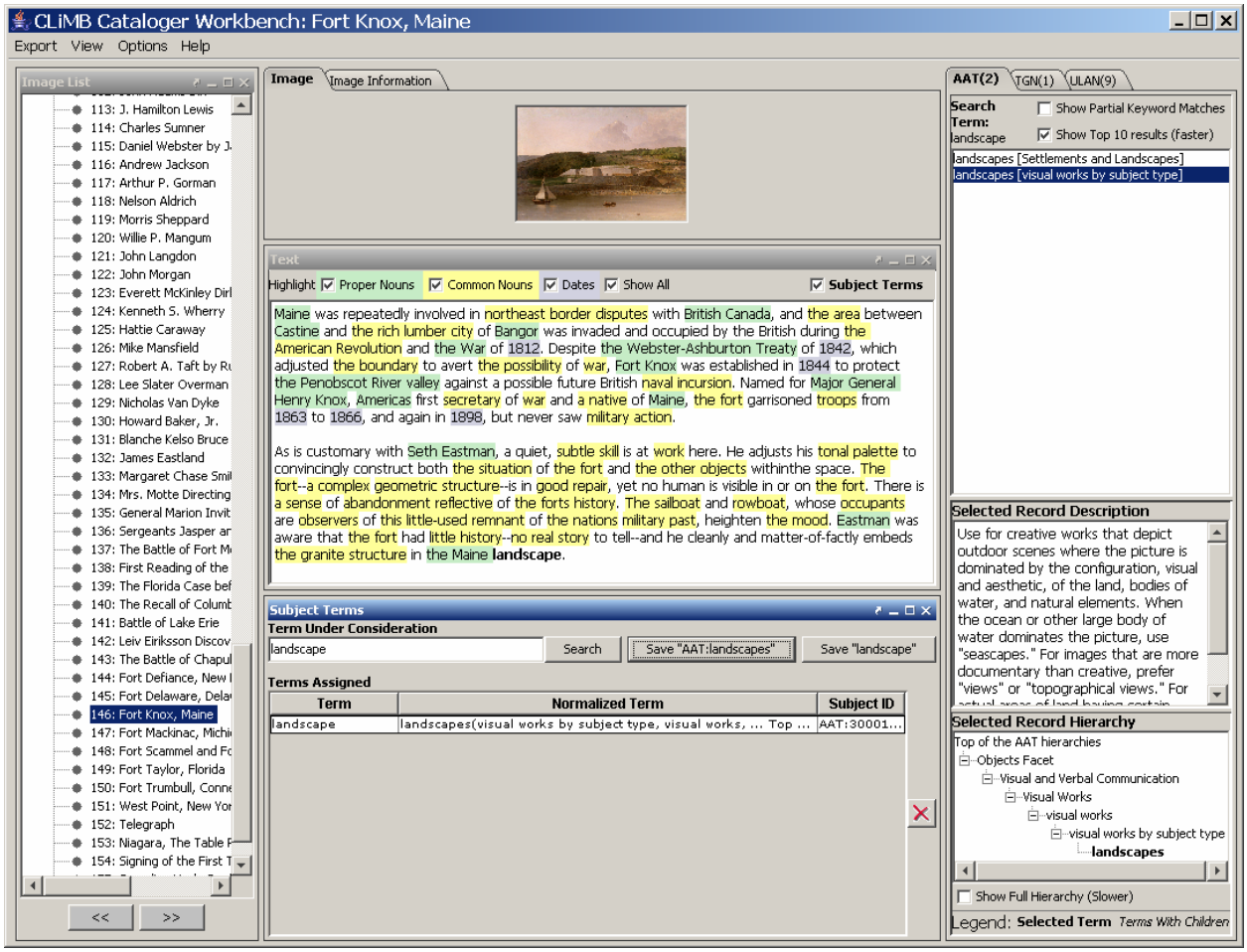
multi-word phrases in the AAT, it is important to remember that the CLiMB Toolkit is searching for an entry that includes all words in the phrase. For multi-word phrases, the user will likely need to perform multiple searches in the CLiMB AAT browser and then combine descriptors upon export. Indexing rules for combining AAT descriptors as *modified phrases* or *strings of descriptors* can be found in: Petersen, T. and Barnett, J.B., eds. Guide to Indexing and Cataloging with the Art & Architecture Thesaurus. The Getty Art History Information Program, Oxford University Press, New York. 1994. 38-46.

You can double click on one of the terms in the panel on the right to see the complete record for this term. It will appear in a separate window and will look like this:



11. Adding Subject Terms from the Text

After a term is selected in the text, click on the appropriate definition given by the thesaurus to see the record description and the hierarchy. To add the mapped term to the “Terms Assigned” list click on the middle button to the right of the "Term Under Consideration", which shows “Save ‘AAT’ :<your term>". To add terms with no matches in the thesauri, see the Section 12.

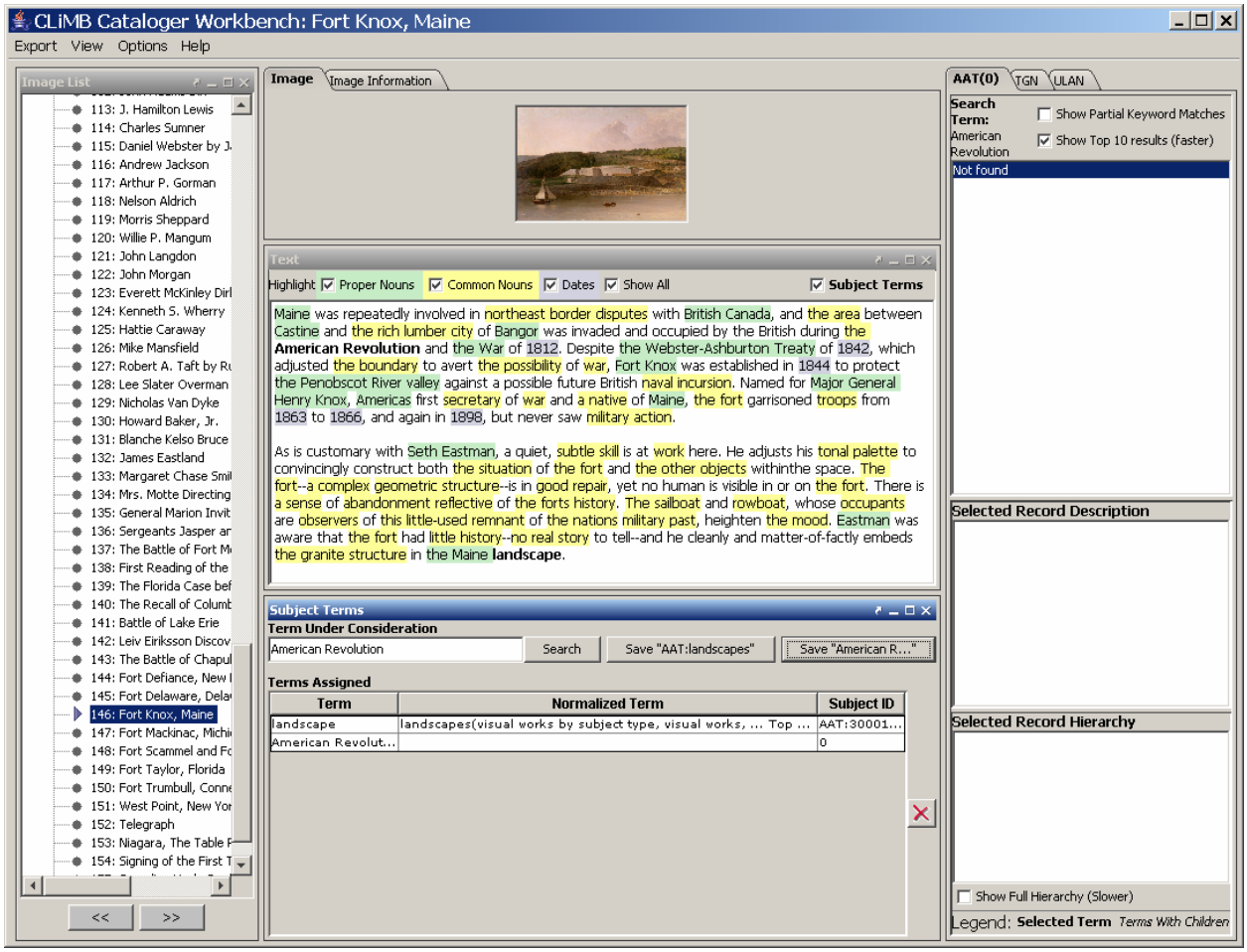


Once you have assigned some terms, the "Show Subject Terms" check box in the text frame becomes active. This feature allows you to view which terms have already been added from the text. When this checkbox is selected, Subject Terms will appear in bold.

12. Adding Your Own Subject Terms

You can add terms which you deem relevant to the image even if they do not occur in the text. Type the term you wish to add directly into the “Term Under Consideration” box.

In the example below, "American Revolution" was added for the image "Fort Knox, Maine". Pressing the "Search" button on the right maps the term to the potential thesaurus definitions. In this example, no match was found. If wish to add the unnormalized term, click on the third icon that says “Save ‘<your term>’". The term will be added to the “Terms Assigned” list with an empty "Normalized Term" field and "0" for Subject ID. See below:



13. Deleting Terms

To select a term for deletion, click on the term in the “Terms Assigned” window. Once the term is selected, click on the red “X” which appears next to this “Terms Assigned” box. To delete all terms, click on the “X” without selecting any terms. The system will ask you if you are sure you want to delete all terms.

14. Exporting Assigned Terms

To export the results, go to the upper left corner of the workbench, and under “Export,” choose XML, Text, VRA 3.0, or MARC.

15. Under Development

We have performed exploratory machine learning for semantic functional categories, research prototypes on disambiguation. For details on these advanced linguistic techniques, please see our final report and publications, available on the CLiMB website at www.umiacs.umd.edu/~climb.

User Manual: Appendix 1

Building an input text file for a collection.

Metadata and the text associated with these images must be in one flat file. Each line of this file corresponds to one image. Fields are separated by <tab> and must be entered in the following order:

Image file name (note: *the image file name must match the image name given in the image's folder*), Title *(this will appear in both the Image List and the Image Information frames of the Toolkit*), text about the image, Creator, Creator date, Measurements, Medium, Date.

Capitalized fields above correspond to metadata that appear in the Image Information box in the Toolkit. If some of this information is missing, the corresponding field will be empty and in the file it should be indicated by the one <tab> following another.

One way to create this file is to enter information in an Excel spreadsheet and save the file using "Save as type: Text (Tab delimited)(*.txt)".

The following page shows a sample collection file for three images from the Senate Catalogue of Fine Arts:

1. "George Washington" (number 20)
2. "George Clinton" (number 25)
3. "Fort Knox, Maine" (number 146).

The first row in the sample chart, which shows column names for the sake of clarity, would need to be deleted before saving the file.

Note: When populating the text column, all text for a given image should be entered into one cell. When saving the file as Text (tab delimited) Excel may enclose some content in quotation marks but this will be automatically resolved when the fill is input into the Toolkit. You can see these quotes around "text" fields in the sample chart and also in the sample tab delimited file.

Text must always be enclosed with both an opening paragraph tag ( ) and a closing paragraph tag ( ). This should be repeated for each paragraph within a body of text associated with the image. For example: text for paragraph 1 text for paragraph 2 . To conserve space, the sample chart does not display the full text; omitted text is indicated by […].

1	Name of the image file	Title	Text	Creator	Creator date	Measurements	Medium	Date
2	21_00020.jpg	George Washington	In April 1966, the Joint Committee on the Library accepted a marble bust of George Washington from Mary Frances Drinker of Jenkintown, Pennsylvania. In her initial letter to Senator Everett Jordan of North Carolina, then chairman of the committee, the donor stated that the Washington bust-in the [...]			31.25 inches x 22.5 inches		
3	22_00004.jpg	George Clinton	Sculptor Vittorio A. Ciani received the commission for a marble bust of Vice President George Clinton under authority of a Senate resolution approved in 1886. Ciani, an Italian citizen who [...]	Vittorio A. Ciani	(1858-1908)	29.25 inches x 27 inches	Marble	Modeled 1893, Carved 1894
4	33_00013.jpg	Fort Knox, Maine	Maine was repeatedly involved in northeast border disputes with British Canada, and the area between Castine and the rich lumber city of Bangor was invaded and occupied by the British during the American Revolution and the War of 1812. Despite [...]	Seth Eastman	(1808-1875)	24.25 inches x 35.5 inches	Oil on canvas	1870-1875

When saved as a tab delimited file, each row will be saved as one line. Below you can see the saved file (opened in Notepad and copy/pasted here) for these three images. Here, each text is shown in full. The markup is as follows:

- [Tab] = Tabs (field demarcations)
- [EOL] = End Of Line
- , = paragraph opening and closing tags.

Note that two of the images in this example have all fields populated, but the first one has nothing in the columns “Creator”, “Creator date”, “Medium”, “Date”. Recall that empty fields will be indicated by consecutive [Tab] tags.

21_00020.jpg[Tab] George Washington[Tab] " In April 1966, the Joint Committee on the Library accepted a marble bust of George Washington from Mary Frances Drinker of Jenkintown, Pennsylvania. In her initial letter to Senator Everett Jordan of North Carolina, then chairman of the committee, the donor stated that the Washington bust-in the style of noted French sculptor Jean-Antoine Houdon-had belonged to her father, Henry Middleton Fisher, and had been on loan to the Philadelphia Museum of Art for some years. In recommending the acquisition of the bust to the Joint Committee, Architect of the Capitol J. George Stewart noted Washingtons close association with the design of the Capitol, adding that the first president had in fact laid the cornerstone for the building. The artist of the Senate bust is unknown. While on loan to the Philadelphia Museum of Art from 1937 until its transfer to the Capitol in 1966, the bust was classified as French, early 19th century. The work resembles the 1785 study of Washington by Houdon. In that year, Houdon had traveled with three assistants to Washingtons home, Mount Vernon. There, he modeled a bust of the Revolutionary War hero inclay and made a life mask of Washingtons face. The sculptor and his assistants subsequently produced many versions of Washington, including busts, statuettes, and statues in plaster, bronze, and marble. Gustavus Eisen, in his study Portraits of Washington, identifies four styles of Washington busts sculpted by Houdon. The Senate bust most closely corresponds to the version Eisen terms chest covered with drapery in toga style. (1) The sculptor of this bust, however, departed from Houdons classic likeness both in the drape of the toga (which is not held by a button on the right shoulder, as is the toga of Houdons bust) and in the wavy treatment of the hair. "[Tab] [Tab] [Tab] 31.25 inches x 22.5 inches[Tab] [Tab] [EOL]
22_00004.jpg[Tab] George Clinton[Tab] " Sculptor Vittorio A. Ciani received the commission for a marble bust of Vice President George Clinton under authority of a Senate resolution approved in 1886. Ciani, an Italian citizen who maintained a studio in New York City, based his work on the best-known likeness of the vice president: an oil on canvas portrait by Ezra Ames. In December 1893 Ciani wrote Architect of the Capitol Edward Clark that he had completed the clay study model. The finished work, in marble,

was installed in the Senate Chamber in 1894 as part of the Vice Presidential Bust Collection. Ciani is also known for sculpted works in St. Bartholomews Church, Grace Church, and Columbia College--all in New York City--and at The Breakers, the Newport, Rhode Island, home of the Vanderbilts. In 1897 Ciani received the cross of the Crown of Italy for historic merit from King Humbert. "[Tab] Vittorio A. Ciani[Tab] (1858-1908) [Tab] 29.25 inches x 27 inches[Tab] Marble[Tab] "Modeled 1893, Carved 1894"[EOL]
33_00013.jpg[Tab] "Fort Knox, Maine"[Tab] " Maine was repeatedly involved in northeast border disputes with British Canada, and the area between Castine and the rich lumber city of Bangor was invaded and occupied by the British during the American Revolution and the War of 1812. Despite the Webster-Ashburton Treaty of 1842, which adjusted the boundary to avert the possibility of war, Fort Knox was established in 1844 to protect the Penobscot River valley against a possible future British naval incursion. Named for Major General Henry Knox, Americas first secretary of war and a native of Maine, the fort garrisoned troops from 1863 to 1866, and again in 1898, but never saw military action. As is customary with Seth Eastman, a quiet, subtle skill is at work here. He adjusts his tonal palette to convincingly construct both the situation of the fort and the other objects withinthe space. The fort--a complex geometric structure--is in good repair, yet no human is visible in or on the fort. There is a sense of abandonment reflective of the forts history. The sailboat and rowboat, whose occupants are observers of this little-used remnant of the nations military past, heighten the mood. Eastman was aware that the fort had little history--no real story to tell--and he cleanly and matter-of-factly embeds the granite structure in the Maine landscape. "[Tab] Seth Eastman[Tab] (1808-1875) [Tab] 24.25 inches x 35.5 inches[Tab] Oil on canvas[Tab] 1870-1875[EOL]

Appendix 2: List of CLiMB-2 Presentations

1. Judith Klavans, Carolyn Sheffield, Eileen Abels, Joan Beaudoin, Laura Jenemann, Tom Lipincott, Jimmy Lin, Rebecca Passonneau, Tandeep Sidhu, Dagobert Soergel, and Tae Yano. **Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access.** OntoImage 2008: 2nd International *"Language Resources for Content-Based Image Retrieval" Workshop.* 2008. Marrakech, Morocco.

2. Rebecca Passonneu, Tom Lippincott, Tae Yano, and Judith Klavans. 2008. **Relation between Agreement Measures on Human Labeling and Machine Learning Performance: Results from an Art History Domain.** Proceedings of the Sixth International Language Resources and Evaluation, LREC. 2008. Marrakech, Morocco.

3. Judith Klavans, Carolyn Sheffield, Jimmy Lin, and Tandeep Sidhu. 2008. **Computational Linguistics for Metadata Building (CLiMB) Demo Session.** Joint Conference on Digital Libraries, June 2008. Pittsburgh, Pennsylvania.

4. Rebecca J. Passonneau, Tae Yano, Tom Lippincott, and Judith Klavans. 2008. **Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning**. Workshop 3: International Workshop on Metadata Mining for Image Understanding; VISAPP International Conference on Computer Vision Theory and Applications. MMIU 2008. Funchal, Madeira - Portugal.

5. Judith Klavans, Tandeep Sidhu, Carolyn Sheffield, Dagobert Soergel, Jimmy Lin, Eileen Abels, Rebecca Passonneau. **Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata**. *VISAPP Workshop Metadata Mining for Image Understanding.* MMIU 2008. Funchal, Madeira - Portugal.

6. Judith Klavans, Carolyn Sheffield, Jimmy Lin, Rebecca Passonneau, Eileen Abels, Dagobert Soergel, Tandeep Sidhu. **CLiMB Poster Session.** Visual Resource Association, March 2008. San Diego, California.

7. Judith Klavans, Jimmy Lin, Rebecca Passonneau, Eileen Abels, Carolyn Sheffield, Dagobert Soergel, Tandeep Sidhu. **CLiMB Poster Session.** *Human Computer Interaction Lab Symbosium***,** 2007, University of Maryland, College Park.

8. Judith Klavans, Carolyn Sheffield, Tandeep Sidhu, **CLiMB.** *Visual Resources Association (VRA)***,** March 30, 2007, Kansas City, MO.

9. Rebecca Passonneau, Roberta Blitz, David Elson, Angela Giral, Judith Klavans. **CLiMB ToolKit: A case study of iterative evaluation in a multidisciplinary project.** *Proceedings of the Fifth International Conference*

on Language Resources and Evaluation (LREC) Genoa, Italy. May 24-26, 2006

10. Angela Giral, Marilyn D. White, Jimmy Lin, Judith Klavans.**CLiMB.** *Visual Resources Association (VRA)* Baltimore, MD. March 7, 2006

11. Judith Klavans, Marilyn White, Angela Giral. *Coalition for Networked Informaiton (CNI)* Washington, D.C. April 4-5, 2005.

Appendix 3: CLiMB-2 Publications (full text)

4.3.1 Judith Klavans, Carolyn Sheffield, Eileen Abels, Joan Beaudoin, Laura Jenemann, Tom Lipincott, Jimmy Lin, Rebecca Passonneau, Tandeep Sidhu, Dagobert Soergel, and Tae Yano. **Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access.** *OntoImage 2008: 2nd International "Language Resources for Content-Based Image Retrieval" Workshop.* 2008. Marrakech, Morocco.

Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access

Judith Klavans[1,2], Carolyn Sheffield[2], Eileen Abels[3], Joan Beaudoin[3], Laura Jenemann, Jimmy Lin[1,2], Tom Lippincott[4], Rebecca Passonneau[4], Tandeep Sidhu[1], Dagobert Soergel[2], Tae Yano[5]

[1]Laboratory for Computational Linguistics and Information Processing (CLIP) at the Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD [2]The iSchool at the University of Maryland, College Park, MD

[3]College of Information Science and Technology, Drexel University, Philadelphia PA

[4]Center for Computational Learning Systems, Columbia University, New York NY

[5]Department of Computer Science, Carnegie Mellon University, Pittsburgh PA

E-mail: jklavans@umd.edu, eileen.abels@ischool.drexel.edu, jeb56@drexel.edu, lj27@drexel.edu, jimmylin@umd.edu, tom@cs.columbia.edu, becky@ccls.columbia.edu, csheffie@umd.edu, tsidhu@umd.edu, dsoergel@umd.edu, taey@cs.cmu.edu

Abstract

We present a system which applies text mining using computational linguistic techniques to automatically extract, categorize, disambiguate and filter metadata for image access. Candidate subject terms are identified through standard approaches; novel semantic categorization using machine learning and disambiguation using both WordNet and a domain specific thesaurus are applied. The resulting metadata can be manually edited by image catalogers or filtered by semi-automatic rules. We describe the implementation of this workbench created for, and evaluated by, image catalogers. We discuss the system's current functionality, developed under the Computational Linguistics for Metadata Building (CLiMB) research project. The CLiMB Toolkit has been tested with several collections, including: Art Images for College Teaching (AICT), ARTStor, the National Gallery of Art (NGA), the Senate Museum, and from collaborative projects such as the Landscape Architecture Image Resource (LAIR) and the field guides of the Vernacular Architecture Group (VAG).

1. Project Goals

Creating access to ever-growing collections of digital images in scholarly environments has become increasingly difficult. Studies indicate that current cataloging practices are insufficient for accommodating this volume of visual materials, particularly for diverse user needs. The goal of the CLiMB project is to leverage text already written about images for automatically identifying, categorizing, filtering and selecting high quality descriptive metadata for image access.

Typically, in libraries and museums, cataloging is performed manually with minimal tombstone cataloging, i.e. the basic set of information (e.g. name of work, creator, date). However, what is usually lacking are rich descriptive terms (e.g. for Picasso's Guernica, "screaming horse", "the frozen women", "fauns" and "minotaurs")[1]. In addition, many legacy records lack subject entries altogether. The literature on end users' image searching practices, though sparse, indicates that this level of subject description may be insufficient for some user groups, including both general users and domain experts with knowledge of specialized vocabularies. Furthermore, the lack of subject-oriented description precludes searching and image analysis

across topic area (e.g. searching for works with “minotaurs” as a theme).

Our hypothesis is that automatic and semi-automatic techniques may help fill the existing metadata gap by facilitating the assignment of subject terms. In particular, we are interested in the impact of computational linguistic technologies in extracting relevant access points from pre-selected texts. The CLiMB Toolkit applies Natural Language Processing (NLP), categorization, and disambiguation techniques over texts about images to identify, filter, and normalize high-quality subject metadata

2. The CLiMB Toolkit

Figure 2 shows a screen shot of the CLiMB Toolkit user interface for an image and text from the National Gallery of Art online collection[2]. Note that the center top panel contains the image, so catalogers can examine items as they work. The center panel contains the input text, with proper and common nouns highlighted. Terms under consideration are displayed below the full text with thesaural information accessible in the right-hand panel. Under this is the term the user has selected for consideration. The right-hand panel gives thesaural information. For normalizing terms, we use the Getty Vocabularies[3]: the Art and Architecture Thesaurus (AAT), the Thesaurus for Geographic Names (TGN), and the Union List of Artist Names (ULAN). In this example, two senses for the word “landscape” are displayed on the right. Note that the top portion of the panel displays possible matches in the AAT, followed by the middle portion which shows the chosen definition for the selected term, and finally, the bottom panel in which the entire hierarchy is displayed for the user to view and used to identify any related terms.

To extract terms from these relevant segments, we use off-the-shelf software to perform traditional NLP techniques. In the current Toolkit, the Stanford tagger (Toutanova and Manning, 2000; Toutanova et al., 2003) is used since it is Java compliant and currently outperforms other taggers. We have used the open source Lucene toolbox to index. Internally developed noun phrase and proper noun identification rules have been applied. As part of categorization, we have applied a machine learning technique trained over text in the art and architecture domain to select a functional semantic category (Passonneau, 2008). Finally, we explore several disambiguation techniques, which we continue to refine and test with our user groups (Sidhu, 2007). Finally, candidate terms are proposed to catalogers for selection and export into an image database.

Currently, CLiMB focuses on nouns and noun phrases. Recent literature on image indexing indicates, however, that other parts of speech may be valuable in retrieving images. In a study of image professionals, (graphic designers, advertising staff, etc.), Jorgensen (2005) found that “while nouns account for the largest percentage of term type in image searches (just over 50%), adjectives account for 18% of the total term usage, verbs 10%, proper nouns 5%, concept 8%, byline 2%, visual content 2%, and date 1%. Of course, these results are highly dependent on the users and their image needs, but it does give some indication of the relative importances of the term types being searched.”

3. Related Research

Broad domain users (as opposed to specialists) require access using broader non-specialist terms. Choi and Rasmussen (2003) studied the image-searching behaviors of faculty and graduate students in the domain of American history and found that generalists submitted more subject-oriented queries than known author and title searches. Currently, much cataloging is geared towards the specialist. On the other end of the spectrum is pure indexing of textual material in the physical domain of an image, such as that done by google (Palmer n.d.). Although such approaches are valuable for initial image access, the resulting high recall can make for a frustrating browsing experience for the end user.

[1] Taken from the exhibition notes from the Picasso exhibit at the National Gallery of Victoria, published by www.thornton.com

[2] www.nga.gov

[3] http://www.getty.edu/research/conducting_research/ vocabularies/

On the other hand, the subjective nature of images inherently complicates the generation of accurate and thorough descriptions. Berinstein (1999) points out that even the guidelines provided by the Shatford-Panofsky matrix on what to describe are fluid and may be difficult to apply. Shatford (1994), building on Panofsky (1962), proposed a method for identifying image attributes, which includes analysis of the generic and specific events, objects, and names that a picture is "of" and the more abstract symbols and moods that a picture is "about". Panofsky describes the pre-iconographic, iconographic, and iconologic levels of meaning found in Renaissance art images. Shatford's generic and specific levels correspond to Panofsky's pre-iconographic and iconographic levels, respectively, and encompass the more objective and straightforward subject matter depicted in an image. The iconologic level (Shatford's about) addresses the more symbolic, interpretive, subjective meanings of an image. To aid user access, catalogers are encouraged to consider both general and specific terms for describing the objective content of an image as well as to include the more subjective iconologic, symbolic, or interpretive meanings. Iconologic terms may be the most difficult for catalogers to assign but occur often in texts describing images.

4. Current Cataloging Approaches

In the CLiMB workflow studies, we examined existing cataloging practices and gathered cataloger perspectives on current challenges in image indexing. Understanding the component processes in current practice has enabled the development of the CLiMB workbench to be easily integrated into existing standards, systems, and practices. Furthermore, by determining which challenges are general to the field and which arise in conjunction with specific collections, we were able to identify additional needs which our research may address. In architecture collections, for example, text may describe a building or architectural site as a whole while the corresponding image typically provides only a detailed view of the work. Part-whole relationships such as these present specific linguistic challenges for associating segments of text with one or more images. This research is not the topic of this paper, and will be described in a forthcoming article.

5. CLiMB Architecture: Systems and Methods

The CLiMB architecture is shown in Figure 1. The data flow for CLiMB starts at the upper left which shows the input to the system:

1. an image,
2. minimal metadata (e.g. image, name, creator)
3. text.

This input is pre-processed, using external technologies, to identify coherent segments of text and associate those segments with relevant images. Input texts are marked up using TEI lite (Text Encoding Initiative) to identify topical divisions (chapters, sections, etc.). These divisions, or segments, are then mapped to corresponding images through the identification of plate and figure numbers. For art historical survey texts, such as Jansen (2004) and Gardner (2001), the automation of text-image association produces reliable results. CLiMB has investigated the application of linguistic technologies to semi-automatically classify, or categorize, text segments according to their semantic relationship to the image(s) which they describe Passonneau, et al (2007).

Through our partnership with the Getty Research Institute, we have been given access to three resources:

- The Art & Architecture Thesaurus (AAT), a structured vocabulary for describing art objects, architecture, and other cultural or archival materials. The AAT's structure is comprised of seven major facets (Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, and Objects) from which multiple hierarchies descend. In total, AAT has 31,000 such

[4] http://wordnet.princeton.edu/
[5] http://www.iconclass.nl/
[6] http://www.loc.gov/rr/print/tgm1/
[7] http://www.loc.gov/rr/print/tgm2/

records. Within the AAT, there are 1,400 homonyms, i.e., terms that can lead to several AAT records that may have multiple meanings only one of which may apply in a given context.

• The Union List of Artist Names (ULAN), a name authority that includes the given names of artists, as well as any known pseudonyms, variant spellings, and name changes (e.g., married names). The structure of this resource is similar to the Agents facet of the AAT in that it contains Person and Corporate Body as its primary facets.

• The Thesaurus of Geographic Names (TGN), an authority for place names, including place names as they appear in English as well as in other languages, historical names, and names in natural order and inverted order.

These vocabularies are well-established and widely-used multi-faceted thesauri for the cataloging and indexing of art, architecture, artifactual, and archival materials. Each of these resources specifies which variation of a given concept or name is the preferred term, enabling consistent cataloging across collections. We have utilized these resources to link terms derived from testbed texts to standardized, controlled terms, thus helping users expand their information space. The Getty resources are used to select the particular homograph of a term.

5.1 Disambiguation

We have tested three approaches to disambiguation in our domain, using the AAT as our baseline thesaurus (Sidhu, 2007). However, it is clear that we need to utilize additional terminological resources since many common terms—and senses of ambiguous terms--are missing from the specialist thesaurus. The challenge of using domain-specific vocabularies combined with general vocabularies, and the impact on disambiguation, is a little-studied topic. We have observed that terms with many senses in the AAT may have just one sense in a general dictionary, and that some terms with many senses in a general resource are simply missing altogether in the AAT. The impact of these observations on disambiguation has yet to be established.

In order to test our disambiguation technique, we first annotated a text to use for evaluation. Following standard procedure in word sense disambiguation tasks (Palmer et al., 2006), two labelers manually mapped 601 subject terms to the AAT. Inter-annotator agreement for this task was encouragingly high, at 91%, providing a notional upper bound for automatic system performance (Gale et al., 1992). We have used SenseRelate (Banerjee and Pederson, 2003; Patwardhan et al., 2003) for disambiguating AAT senses. SenseRelate uses word sense definitions from WordNet 2.1, a large lexical database of English nouns, verbs, adjectives, and adverbs.4

Results from our evaluations (discussed in Sidhu et al, 2007) show that mapping to WordNet first and then to the AAT causes errors. As a general resource, WordNet is domain independent and thus offers wider, more comprehensive coverage. However, the lack of domain specificity also creates overhead as there are many irrelevant senses to choose from and the correct sense needed for art and architecture discourse may not be available. Similarly, Iyer and Keefe (2004) report on an exploratory study on the use of WordNet to clarify concepts for searching architectural visual resources. Twenty participants were shown images which they were asked to locate using natural language or WordNet terms. Although 70% of participants stated that WordNet clarified the terms or the images, 30% reported problems with conceptualizing the image, and 55% had terminology problems. To address these types of problems, we are exploring the option of re-implementing concepts behind SenseRelate to directly map terms to the AAT. Additionally, in Future Work we will test approaches for employing hybrid techniques (including machine learning) for disambiguation. This will enable us to explore the trade-off in precision between different configurations of resource calling.

5.1.1. Catalog Record Creation: Select

As shown in Figure 2, a cataloger is presented with the image to be cataloged, the text segment associated with the image, and a number of index terms suggested by the Toolkit. The user decides which of the terms proposed by the CLiMB system should be included in the image's record.

5.2 Testbed Collections

We are currently working with five image-text sets and one image collection for which we are conducting experiments with dispersed texts located online. Table 1 illustrates the relationship between the associated texts and the image collections which we use to test our system.

Feedback from catalogers indicates that one thesaural resource is insufficient for cataloging a range of art historical and architecture images. The Getty resources are extensive but, as with any resource, are not entirely comprehensive. Our goal is to expand our capabilities for disambiguating domain-specific terminology by cross-searching multiple, established thesauri in the art and architecture domain. Resources currently under consideration include Iconclass[5] and the Library of Congress' Thesaurus for Graphic Materials (TGM) I and II[6,7].

6. Conclusion and Future Work

The CLiMB project techniques exceed simple keyword extraction and indexing by:

- applying novel semantic categorization to text segments,
- identifying and filtering linguistically coherent phrases,
- associating terms with a thesaurus, and
- applying disambiguation algorithms to these terms.

Although each of these techniques has been used in other projects, they have not been combined and tested in the art and architecture domains for improving digital library access. Our future work will consist of three foci:

- Integration of functional semantic categorization with disambiguation
- Improvement of disambiguation
- Testing the system and its components with users to drive improvements

We also hope to incorporate the output of CLiMB text data mining with a social tagging approach to image labeling, such as that of steve.museum to examine terminological comparisons and their impact on image access.

Acknowledgements

We acknowledge input on the project from Dr. Murtha Baca of the Getty Vocabularies Institute. We also appreciate ongoing input from both our internal and external advisory boards, the members for which are listed on our web pages. Finally, many users whom we cannot name have helped in evaluations along the way. The project has been funded by the Andrew W. Mellon Foundation to the Center for Research on Information Access at Columbia University. Later support was provided to the College of Information Studies and the University of Maryland Institute for Advanced Computer Science at the University of Maryland, College Park, MD.

Image Collection	Text	Image/Text Relationship
National Gallery of Art (NGA) Online Collection	Narratives associated with images on the NGA website	Integrated
U.S. Senate Museum	U.S. Senate Catalogue of Fine Arts	Integrated
The Vernacular Architecture Forum (VAF)	VAF Field Guides:	Integrated
The Society of Architectural Historians (SAH): World Architectural Survey and the American Architectural Survey	*Buildings Across Time: An Introduction to World Architecture* by Marian Moffett, et al.	External
Landscape Architecture Image Resource (LAIR)	*Landscape Design: A Cultural and Architectural History* by Elizabeth Barlow Rogers	External
Art History Survey Collection (AHSC), ARTstor	Disparate texts located online	External

Table 1: Sources of Image and Testbed Text Collections

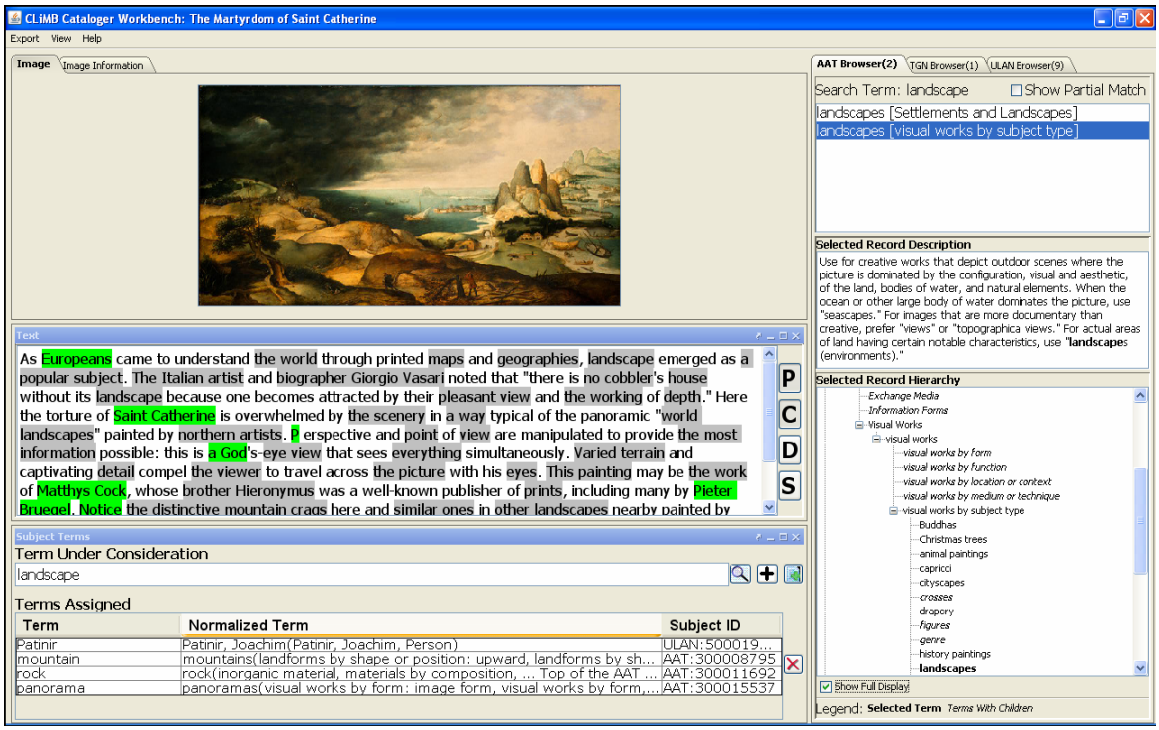


Figure 2: CLiMB User Interface for the term “landscape”

References

Banerjee, S., Pedersen, T. (2003) Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pp. 805–810.

Berinstein, P. (1999). Do you see what I see? Image indexing principles for the rest of us. Online, 23(2), 85-87.

Choi, Y., Rasmussen, E. M. (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, 54.6: 498-511.

Gale, W., K. W. Church, and D. Yarowsky. (1992) "Estimating upper and lower bounds on the performance of word-sense disambiguation programs." Proceedings of the 30th conference on Association for Computational Linguistics. pp. 249-256.

Gardner, Helen. (2001) Art through the Ages. 11th edition. Edited by Fred S. Kleiner, Christin J. Mamiya, Richard G. Tansey. Harcourt College Publishers, Fort Worth, Texas.

Iyer, H., & Keefe, J. M. (2004). WordNet and keyword searching in art and architectural image databases. VRA Bulletin, 30, pp. 22-27.

Janson, H. W. (2004) History of Art: the Western tradition. Revised 6th edition. Upper Saddle River, NJ: Pearson/Prentice-Hall.

Jorgensen Corinne, and Peter Jorgensen. (2005) Image querying by image professionals. Journal of the American Society for Information Science and Technology. Hoboken: Oct 2005. Vol. 56, Iss. 12; pg. 1346.

Klavans, Judith L. (2006) Computational Linguistics for Metadata Building (CLiMB). In Procedings of the OntoImage Workshop, G. Greffenstette, ed. Language Resources and Evaluation Conference (LREC), Genova, Italy.

Palmer, Justin (n.d.) Optimizing Your Images for Google Image Search - 4 Image SEO Tips. http://ezinearticles.com.

Palmer, M., Ng, H.T., Dang, H.T. (2006) Evaluation. In: Edmonds, P., Agirre, E. (eds.): Word Sense Disambiguation: Algorithms, Applications, and Trends. Text, Speech, and Language Technology Series, Kluwer Academic Publishers, Netherlands.

Panofsky, E. (1962). Studies in Iconology: Humanistic Themes in the Art of the Renaissance. Harper & Row, New York .

Passonneau Rebecca, Tae Yano, Judith Klavans, Rachael Bradley, Carolyn Sheffield, Eileen Abels, Laura Jenemann (2007) Selecting and Categorizing Textual Descriptions of Images in the Context of an Image Indexer's Toolkit. Technical Report. Columbia University.

Passonneau Rebecca J., Tae Yano, Tom Lippincott, and Judith Klavans. (2008). Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning. International Workshop on Metadata Mining for Image Understanding; VISAPP International Conference on Computer Vision Theory and Applications. MMIU 2008. Funchal, Madeira - Portugal.

Patwardhan, S., Banerjee, S., Pedersen, T. (2003) Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.

Shatford-Layne, S. (1994): "Some issues in the indexing of images." Journal of the American Society for Information Science 45.8. pp. 583-588.

Sidhu, T., Klavans, J.L., Lin, J. (2007) Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTech 2007), 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic

Toutanova, K., and C. D. Manning. (2000) Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. EMNLP/VLC 2000. pp. 63–70.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1. Edmonton, Canada: Association for Computational Linguistics. pp. 173-180.

4.3.2 Rebecca Passonneu, Tom Lippincott, Tae Yano, and Judith Klavans. 2008. **Relation between Agreement Measures on Human Labeling and Machine Learning Performance: Results from an Art History Domain**. *Proceedings of the Sixth International Language Resources and Evaluation*, LREC. 2008. Marrakech, Morocco.

Relation between Agreement Measures on Human Labeling and Machine Learning Performance: Results from an Art History Image Indexing Domain

Rebecca J. Passonneau[1], Tae Yano[2], Tom Lippincott[1], and Judith Klavans[3]
[1] Columbia University
(becky|tom)@cs.columbia.edu
[2] Carnegie Mellon University
tyyano@gmail.com
[3] University of Maryland
jklavans@umd.edu

Word count: 1,993

1 Introduction

We describe a series of studies aimed at identifying specifications for marking up textual input for an image indexer's toolkit. Given an image and a text extract that describes an art or architectural work depicted in an image, our goal is to identify the semantic function served by a span of text with respect to image description. We illustrate this below in Figure 1.

The domain of digital images and texts we focus on parallels the ARTstor *Art HistorySurvey Collection (AHSC)*, a Mellon funded collection of 4,000 images. The AHSC is based on thirteen standard art history survey texts, thus there is a strong correlation between the images and these texts. The AHSC images all have tombstone metadata (e.g., the name of the work, the artist, date, the location of the work), but few have subject matter metadata. We are currently using two of the texts from the AHSC concordance that we scanned and encoded in TEI-Lite (http://www.teic.org/Lite/teiu5 split en.html).

In consultation with domain experts, we developed a set of seven functional semantic categories to apply to paragraphs or sentences associated with specific images. Our categories are derived from what we observe in the texts, but have a loose correspondence with categories of information discussed in the image indexing literature [3, 7, 2]. Figure 1 in the next section illustrates three of our seven categories that we have done machine learning for.

Many studies of machine learning use annotated data, and report the levels of interannotator agreement, but do not directly address the question of how much agreement is enough for learnability. For example, in both [4] and [13], as in the present study, human labelers classify spans of text (sentences or sequences of sentences), and

machine learners are developed to perform the same classification. Here we report on a series of pilot studies where we developed functional semantic categories to label art history survey texts, to measure interannotator agreement under a variety of annotation task constraints, and to evaluate machine learning performance. We have relatively low agreement overall, but good performance on the categories we have sufficient data for.

Our goal with respect to the image indexer's toolkit is to find a set of one or more labels that is useful to image indexers, and that an automatic classifier can apply with high reliability to art history survey texts. Our research goals emerge from the observation that for many semantic tasks, there is relatively low interannotator agreement (see [1]). First, we wanted to understand why previous investigators have found such a wide range of interannotator agreement on similar tasks [5,9]. To address this, our pilot annotation studies included four experiments where we varied annotation parameters. Our second research goal is to investigate the relation between agreement and learnability and the nature of our annotation schema.

Figure 1. Simplified illustration of semantic classification of text extracts

Historical Context	Of the great projects built by Akhenaten hardly anything remains . . .
Through his choice of masters, he fostered a new style.	
Implementation	Known as the Amarna style, it can be seen at its best in
it can be seen at its best in	
Image Content	a sunk relief portrait of Akhenaten and his family. The intimate
domestic scene suggests	
Historical Context	that the relief was meant to serve as a shrine in a private household.

Enhanced XML representation:

```
<p><semcat type="historical">Of the great projects built by Akhenaten hardly anything remains</semcat>. . . . <semcat type="historical">Through his choice of masters, he fostered a new style.</semcat>
<semcat type="implementation">
Known as the Amarna style, it can be seen at its best in a sunk relief portrait of Akhenaten and his family.</semcat>
<semcat type="historical">The intimate domestic scene suggests that the relief was meant to serve as a shrine in a private household.</semcat> . . ..</p>
```

2 Brief Example

Within a paragraph about a given image, the descriptive information can be categorized into distinct types depending on the semantic function of the text with respect to the work depicted in the image. Figure 1 illustrates text from the first part of a few paragraphs associated with an image of a relief portrait of Akhenaten and his family. The image here is taken from the ARTstor Art Images for College Teaching collection (AICT): http://www.arthist.umn.edu/aict/html/ancient/EN/EN006.html. The text fragment is from one of the texts in the concordance to the ARTstor Art History Survey Collection. It has been separated into labeled text spans exemplifying the three categories we have performed learning experiments on. As illustrated, a single sentence can have subparts with distinct semantic functions. The sample of a provisional xml representation shows a

sentence-level assignment because we will not attempt to find subspans within sentences.

3 Human Labeling

We conducted four pilot studies on the labeling where we varied the number of labels that could be assigned to a single item (one, two or unrestricted), the size of the text fragment being labeled (paragraph or sentence), the number of annotators (two to seven), and the type of training for annotators (none, finished examples presented to trainees, true training examples with feedback). During the pilot studies, we developed a labeling interface, then re-implemented the interface for our ongoing, large scale data collection effort (see [8]).

To measure interannotator agreement, we use Krippendorff's *Alpha* [6] along with a set-based distance measure [10] to allow partial credit when the set of labels chosen by one annotator overlaps another's set. Our measures of interannotator agreement varied widely. These results are consistent with previous literature on interannotator consistency in the library and image cataloging domains [5,9], but in contrast to previous work, we can relate deltas in the amount of agreement to specific causes.

Annotation efforts typically aim for agreement measures above a threshold of 0.67, as suggested by Klaus Krippendorff [6]. We have previously argued that because agreement coefficients do not have a known probability distribution, and because they are applied to many kinds of data from many disciplines (see [1]), there is no single ideal threshold for all cases [10, 11]. Instead, we suggest that it is an empirical question that can be investigated in many ways, for example relating measures of tasks in which the annotations are used to the observed agreement levels. A related point has been made in [12], where Riedsma and Carletta report on simulations of learnability from data with different levels of agreement. They present evidence that performance of machine learners does not correlate directly with agreement levels.

Our results on the human labeling task indicate that reliability improves if annotators can select multiple labels, which is consistent with our previous results on a lexical semantic annotation task [11]. We also find that labeling can be done more consistently by senior domain experts, that sentence level labeling can be done more consistently than paragraph level labeling, and that the consistency of the labeling depends heavily on the image/text pair under consideration. In the first pilot dataset (10 images, 24 paragraphs, 159 sentences) where we collected both sentence and paragraph labeling, the overall interannotator agreement among seven annotators was 0.24 for paragraphs and 0.30 for sentences. In the first dataset from our large scale annotation effort (25 images, 45 paragraphs, 313 sentences) annotated by five individuals including two senior domain experts, overall agreement was 0.40 for paragraphs and 0.47 for sentences. When we computed agreement for all 2 to 5 combinations of annotators, the highs (for the two senior experts) were 0.56 (paragraphs) and 0.55 (sentences). Agreement among five annotators broken down by the image being described ranged from 0.70 to 0.16.

4 Machine Learning

Using data from our pilot studies of human labeling, augmented by an additional set of images labeled by one of the co-authors, we investigated the learnability of three categories: Image Content, Historical Context and Implementation. There were insufficient examples from the other categories. All learning was done using WEKA. We created three types of feature sets. Set A consisted of word features selected on the basis of significant chi-square tests. Set B consisted of hand-picked features in approximately half a dozen groups, such as words and phrases characteristic of the art history domain (e.g., *masterpiece*), and words and phrases referring to parts of the human body. Set C consisted of the union of Sets A and B. Among three types of learning algorithms we tested, naive Bayes, SVM and tree-based classifiers, naive Bayes performed best overall. On ten-fold cross-validation, the highest classification accuracy on Image Content relied on feature set C, and achieved 83% accuracy, compared with 63% for Historical Context and 53% for Implementation. The highest accuracy for Historical Context used feature set A: 70%. Using a random forest classifier for the Implementation class, we achieved an accuracy of 80%.

The most frequent label combination for both paragraphs and sentences was the single label Image Content out of 47 distinct combinations of labels for sentences, and 38 combinations for paragraphs. This partly accounts for the relatively high accuracy of learning for the Image Content classifier.

5 Conclusion

Because we are conducting interannotator agreement studies in tandem with machine learning, we can investigate the relationship between the two. We argued (as in [10] that this is an empirical question, given current knowledge. Our results bear out the simulation study presented in [12] that good learning performance can occur when agreement is less than the 0.67 threshold proposed by Krippendorff. This does not mean that good learning performance never requires higher levels of agreement (see [12]). Instead, it shows that in richly annotated datasets such as this one, where we have attempted to develop a set of fully covering categories, interannotator agreement and learnablity interact with the distributions of various categories in the labeled data. We discuss this issue in greater length in the full paper.

References

1. R. Artstein and M. Poesio. Kappa3 = alpha (or beta). Technical Report NLE Technote 2005- 01, University of Essex, Essex, 2005.
2. M. Baca. *Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information*. The Haworth Press, Inc., 2003. Available through Library Literature.
3. H. Chen. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, pages 260–273, 2001.

4. B. Hachey and C. Grover. Sentence classification experiments for legal text summarisation. In *Proceedings of the 17th Annual Conference on Legal Knowledge and Information Systems (Jurix)*, 2004.
5. A. Giral and A. Taylor. Indexing overlap and consistency between the Avery Index to Architectural Periodicals and the Architectural Periodicals Index. Library Resources and Technical Services 37(1):19-44, 1993.
6. K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, 1980.
7. S. S. Layne. Some issues in the indexing of images. Journal of the American Society for Information Science, pages 583–8, 1994.
8. T. Lippincott, T. Yano and R. Passonneau. Submitted. ANTArt: An Extensible Framework for CollectingAnnotations on Texts that Describe Art Images. Submitted to LREC 2008.
9. K. Markey. Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. Library and Information Science Research, pages 155–177, 1984.
10. R. Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.
11. R. Passonneau, N. Habash and O. Rambow. Inter-annotatator agreement on a multilingual semantic annotation task. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.
12. D. Riedsma and J. Carletta. Reliability measurement: there's no safe limit To appear in Computational Linguistics.
13. S. Teufel and M. Moens. Summarising scientific articles – experiments with relevance and rhetorical status. Computational Linguistics, pages 409–445, 2002.

4.3.3 Rebecca J. Passonneau, Tae Yano, Tom Lippincott, and Judith Klavans. 2008. **Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning**. *Workshop 3: International Workshop on Metadata Mining for Image Understanding*; VISAPP International Conference on Computer Vision Theory and Applications. MMIU 2008. Funchal, Madeira - Portugal.

Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning*

Rebecca J. Passonneau[1], Tae Yano[2], Tom Lippincott[3], and Judith Klavans[4]

[1] Center for Computational Learning Systems, Columbia University
becky@ccls.columbia.edu

[2] Department of Computer Science, Carnegie Mellon University
taey@cs.cmu.edu

[3] Department of Computer Science, Columbia University
tom@cs.columbia.edu

[4] College of Information Studies, University of Maryland
jklavans@umd.edu

Abstract. The CLiMB project investigates semi-automatic methods to extract descriptive metadata from texts for indexing digital image collections. We developed a set of functional semantic categories to classify text extracts that describe images. Each semantic category names a functional relation between an image depicting a work of art historical significance, and expository text associated with the image. This includes description of the image, discussion of the historical context in which the work was created, and so on. We present interannotator agreement results on human classification of text extracts, and accuracy results from initial machine learning experiments. In our pilot studies, human agreement varied widely, depending on the labeler's expertise, the image-text pair, the number of labels that could be assigned to one text, and the type of training, if any, we gave labelers. Initial machine learning results indicate the three most relevant categories are machine learnable. Based on our pilot work, we implemented a labeling interface that we are currently using to collect a large dataset of text that will be used in training and testing machine classifiers.

1 Introduction

The work presented here was developed in the context of the Computational Linguistics for Metadata Building (CLiMB) research project, which has been investigating methods for automated support to image catalogers and other image professionals [5]. The CLiMB project is developing a Toolkit for image catalogers that would allow them to access electronic versions of texts in order to harvest descriptive metadata. Here we address how to classify text extracted from electronic documents into functional semantic categories that are relevant for indexing images of artworks. The tagging would be included in the preprocessing phase prior to importing texts into the Toolkit.

* We thank the current and past members of the project members; numerous advisors who reviewed our labeling categories and tested the interface; and especially, volunteer labelers from University of Maryland, Columbia University, Drexel University, and Indiana University.

Figure 1 shows an image taken from the ARTstor Art Images for College Teaching collection (AICT): `http://www.arthist.umn.edu/aict/html/ancient/ EN/EN006.html`. It depicts a relief portrait of Akhenaten and his family. Also shown is an extract from an art history survey text describing an image of the same work. If the terms **Akhenaten** and **shrine** were used to index the image, it would not be clear whether the image depicts Akhenaten or a shrine or both. Consider that the word Akhenaten occurs in a sentence about Akhenaten's role in fostering the Amarna style, and in another sentence indicating that he is depicted in the work. The word shrine occurs in a sentence indicating how the depicted work was used. Our goal is to automatically tag sentences like these prior to semi-automatic or automatic extraction of the bold face terms, and for the extracted terms to be associated with tags corresponding to our semantic categories; see the right side of Figure 1. The tagging would permit terms to be filtered or prioritized during the term selection process. It could also facilitate image search; for a user who wants an image of a shrine, it would be possible to exclude cases where *shrine* does not come from text tagged as *Image Content*.

Fig. 1. Illustration of an image and associated descriptive text

Historical Context: Akhenaten
Image Content: Akhenaten
Historical Context: shrine

Of the great projects built by **Akhenaten** hardly anything remains . . . . Through his choice of masters, he fostered a new style. Known as the Amarna style, it can be seen at its best in a sunk relief portrait of **Akhenaten** and his family. The intimate domestic scene suggests that the relief was meant to serve as a **shrine** in a private household.

In consultation with experts, we developed a set of seven categories to apply to paragraphs or sentences extracted from art history survey texts, where the text extracts are about a specific image. A larger number of categories would lead to much sparser data; a smaller number would lead to categories that are less distinct. Two of the categories, for example, are **Image Content**, defined as text that describes the content of the image, and **Implementation**, text that describes how the work was created.

In our four pilot studies, interannotator agreement among humans varied widely, depending on the labeler's expertise, the image-text pair under consideration, the number of labels that could be assigned to one text, and the type of training, if any, we gave labelers. Human agreement improved if annotators could select multiple labels, which is consistent with our previous results on a lexical semantic annotation task [10]. We also found that agreement was higher among domain experts, and that the level of agreement depended heavily on the image/text pair under consideration.

Our seven semantic categories vary in relevance, frequency and distinguishability. Thus we do not anticipate attempting to apply machine learning to every category. Using texts labeled during our pilot studies, we have initial results on three of the classes. For example, using a separate naive Bayes classifier for each category, we have been able to achieve 80% accuracy. This indicates that large scale machine learning is feasible.

In section 2, we summarize related work on interannotator agreement among humans. Section 3 describes our datasets. Section 4 describes our pilot studies on human labeling, and our current large scale effort. We present preliminary learning results in section 5 and our conclusions in section 6.

2 Related Work

There are relatively few discussions of interannotator agreement or inter-indexer consistency for image indexing and classification tasks. Two works that address the topic deeply and broadly are [8] and [4]. In the twenty plus years since Markey's [8] analysis of forty years of inter-indexer consistency tests, no comparable review has appeared, and her observations still hold. Although her goal was to use the conclusions from previous work to sort through the issues involved in indexing visual material, all the tests referenced in her paper were on indexing of printed material. She notes significant variability, with accuracy (percent agreement) results ranging from 82% to a low of 4%.

Giral and Taylor [4] looked at indexing overlap and consistency on catalog records for the same items in architectural collections, including an analysis of subject descriptors. On large (≥ 1400) samples of records from the Avery Index to Architectural Periodicals and the Architectural Periodicals Index, they compare proportions of items in their samples that match according to a variety of criteria, and compute 90% confidence intervals. Only 7% of items match entirely, and they find some element of overlap in descriptors in only about 40% of the remaining cases (± 3%).

Markey noted two features of documents that affect inter-indexer consistency: document length, and the complexity of the document, which is difficult to quantify. Our image/text pairs, which correspond to Markey's documents, are quite short. We did not attempt to measure complexity, but we did find wide variation in labeling consistency depending on the image/text pair being labeled. This indicates that the image/text pairs have inherent properties that make them more or less difficult for humans to agree on.

Markey found no significant difference between the indexers with or without experience in using such schemes. She found no higher levels of inter-indexer consistency among subject specialists, as compared with non-specialists. This is in contrast to our results. In our pilot studies, the two developers of the categories (the first two co-authors) were the most familiar with them, and had the highest interannotator agreement. In our current large-scale labeling effort, the highest agrement is found among the most expert pairs of labelers. We also found that across studies, agreement increased when we provided more training. Markey found that using a standardized scheme led to higher inter-indexer consistency, ranging from 80% to 34% (in contrast to 4%; see above). This is roughly the range we find, using a different metric but a similar scale.

3 Texts

The domain of digital images and texts we focus on parallels the ARTstor *Art History Survey Collection (AHSC)*. ARTstor is a Mellon funded non-profit organization developing digital image collections and resources. The AHSC is a collection of 4,000 images. One of our motivations for focusing on the AHSC is that it is based on thirteen standard art history survey texts, thus there is a strong correlation between the images and texts that describe them. The AHSC images all have metadata providing the name of the work, the artist, date, and so on, but very few have subject matter metadata.

We are currently using two of the texts from the AHSC concordance of thirteen art history survey volumes. Both books cover a broad time range, from Neolithic art to late 20th century, and have a similar lineup of chapter topics. Each text contains roughly thirty chapters (approximately five megabytes), with twenty to forty color images each.

For research purposes, we created electronic versions of the two texts, encoded in TEI compliant xml. TEI is a widely used interdisciplinary standard of text representation. (See http://www.tei-c.org/Lite/teiu5 split en.html for more detail of this schema.) Chapters, subdivisions, and paragraphs (but not sentences) have distinctive xml tags.

To construct image/text pairs for our text labeling experiments, we started with a software module we developed for importing text into our image indexer's Toolkit. This software module relies primarily on the relative position of xml tags for image plates, major text divisions, and paragraph boundaries. It takes a chapter as input, and produces a list of all the plates in the chapter, with each plate number associated with

a sequential list of associated paragraph numbers. We manually correct the output before importing the data into our labeling interface. Using Google image search, we locate non-copyrighted images of the works depicted in the book plates.

4 Text Labeling Experiments

4.1 Semantic Category Labels

Our current guidelines give four pieces of information per semantic category: the category name, one or two questions the labeled text should answer, one or two paragraphs describing the category, and four image/text pairs that exemplify each category. For the Image Content category (or label), the questions are *Does the text describe what the art work looks like? What conventional use of symbols does the artist rely on?*

Over a period of four months, we developed a set of functional semantic categories for classifying paragraphs and sentences in our art history survey texts. Three criteria motivated the classification. Most important, we did not attempt to develop an independent set of categories based on existing image indexing work. We took the texts as our starting point. Second, the set of classes were designed to apply to all chapters regardless of time period, and to allow most extracted paragraphs or sentences to fall into a specific category (i.e., to constitute a ”covering” of the text), rather than to a default *Other* class. Finally, we worked with an image librarian at Columbia University and a metadata expert to arrive at a relevant set.

Table 1 summarizes our seven semantic categories. The column on the left indicates the name of the label, and the column on the right gives a highly abbreviated description of the type of textual content that should be assigned a given label. The labels appear here in the same order that they appear in the interface. It puts the most central category first (Image Content), and lists categories that have a similar focus together. Thus the first three categories are all about the depicted art work (form, meaning, manner); Biographic and Historical Context are both about the historical context.

Table 1. Seven Functional Semantic Categories for Labeling Text Extracts

Category Label	Description
Image Content	Text that mentions the depicted object, discusses the subject matter, and describes what the artwork looks like, or contains.
Interpretation	Text in which the author provides his or her interpretation of the work.
Implementation	Text that explains artistic methods used to create the work, including the style, any technical problems, new techniques or approaches, etc.
Comparison	Text that discusses the art object in reference to one or more other works to compare or contrast the imagery, technique, subject matter, materials, etc.
Biographic	Text that provides information about the artist, the patron, or other people involved in creating the work, or who have a direct and meaningful link to the work after it was created.
Historical Context	Text describing the social or historical context in which the depicted work was Created, including who commissioned it, or the impact of the image on the social or historical context of the time.
Significance	Text pointing to the specific art historical significance of the image.

During the first month, we arrived at a provisional set of six categories consisting of everything in Figure 1 apart from the italicized category, which now has the name *Implementation*, and developed our first set of guidelines. We added the seventh category after a month or so of pilot work. During the remaining three months we revised the category names and definitions.

4.2 Materials: Datasets, Annotation Constraints, Annotators, and other Task Parameters

We created three sets of image/text pairs for use in the experiments listed in Table 2. Set 1 consisted of thirteen images and 52 associated paragraphs. Set 2 consisted of nine images and 24 associated paragraphs. Set 3 consisted of ten images taken from two new chapters, and was used for sentence labeling (159 sentences) and paragraph labeling (24 paragraphs). Up to four paragraphs were associated with each image, but in most cases there were one or two paragraphs.

Labelers were recruited from the team of project researchers, their acquaintances, and colleagues at other institutions involved in image indexing.

The two parameters of most interest for comparing the experiments appear in columns five (Labels/Par) and six (Annotators). For the first two experiments, the first two coauthors were the annotators, and the number of labels that could be assigned to a single paragraph was unrestricted (any). In experiment 1, the maximum number of labels for a single paragraph was three; each annotator used three labels twice; 99% of the labelings consisted of one or two labels. In experiment 2, 71% of all labels from both annotators were one or two labels; the maximum of four labels occurred once per annotator.

Table 2. Annotation Task Parameters

Exp	Set	Images	Units	Label Set	Labels/Par	Annotators
1	1	13	52	6	any	2
2	2	9	24	7	any	2
3	2	9	24	7	two	5
4a	3	10	24	7	one	7
4b	3	10	159	7	one	7

Due to the relative infrequency of more than two labels in experiments 1 and 2, we added a restriction in experiment three that only two labels could be used. In experiment four, we restricted annotators to a single label per text extract, but expanded the task to include sentences.

For experiments 1 through 3, the labeling was done with pen and paper. For experiment 4, we implemented a browser-based labeling interface that included the guidelines, training materials, and labeling task. Since then, we have developed a much more flexible web-based labeling interface using the Django python environment.

In all our studies and data collection, labelers worked independently at remote sites, and could suspend and resume work at will. After experiment 3, labelers were required to go through a training sequence (approx. one hour). Paragraphs were presented one at a time along with the corresponding image. When we began using sentences as well as paragraphs, labelers would first select a paragraph label; then the labeler would be presented with the same paragraph in a sentence-by-sentence format. Labelers had an opportunity to review and revise their choices.

4.3 Evaluation Metrics

We report interannotator agreement using Krippendorff's α[6], which factors out chance agreement. It ranges from 1 for perfect agreement to values close to -1 for maximally non-random disagreement, with 0 representing no difference from chance distribution. An advantageous feature of αis that instead of treating agreement as a binary distinction, it permits the use of a distance metric to weight the degree of agreement from 0 to 1. Because annotators could make multiple selections, we used a distance metric we refer to as MASI [9]. It is intended for set-based annotations, and gives partial agreement credit when the annotators' sets overlap. Our experiments typically allowed annotators to assign multiple labels to the same text. If one annotator assigns the single label {**Image Content**} to the same text that another annotator labels {**Image Content**, **Implementation**}, a non-weighted agreement measure would assign a score of 0 for non-agreement. In contrast, MASI would assign a weighting of $\frac{1}{3}$ (see [9] for details).

4.4 Human Labeling Pilot Studies

Table 3. Interannotator consistency of paragraph labeling under multiple conditions

Exper.	Dataset	Label Set	#Choices	#Labelers	α_{MASI}
1	Set 1	6	any	2	0.76
2	Set 2	7	any	2	0.93
3	Set 2	7	two	5	0.46
4a	Set 3	7	one	7	0.24
4a’	Set 3	7	merge 4b	7	0.36
4b	Set 3	7	one	7	0.30

Results for the four pilot experiments appear in Table 3. Experiment 2, with the final labeling set of seven labels, the first two co-authors as the sole annotators, and any number of label choices, had the best results. It improved on experiment 1, which used an earlier, less well-defined set of labels. It had a larger set of units (52 rather than 24 paragraphs) from two texts, rather than from a single text.

Experiment 3 was the first attempt to use a larger set of annotators. We hypothesized that with each new annotator, the number of distinct combinations of labels would increase, with the result that a large number of annotators would result in a large set of distinct classes, and correspondingly sparser data. In order to guard against this possiblity, we restricted the number of labels that annotators could apply to two. The resulting αMASI score of 0.46 reflects the relative unfamiliarity of a majority of the five annotators with the labeling categories and domain. When we computed interannotator consistency for all combinations of annotators from the set of five, we found that the two experienced annotators had values on the three measures (0.88, 0.88, 0.90) that were consistent with the results of experiment 2.

We collected sentence labelings for the first time in experiment 4: 4a pertains to the paragraph labels, and 4b to the sentence labels. For experiment 4a’, we computed agreement on paragraphs based on merging the sentence labels. We created a relatively short label consisting of each distinct type of label applied to any sentence in the paragraph. If three sentences of a five-sentence paragraph were labeled Image Content and two were labeled Historical Context, the paragraph level label we compute is the multi-label consisting of these two labels.

Experiments 4a and 4b yielded the poorest results, which we attribute to the constraint that annotators could only apply one label. The seven labelers consisted of the first two co-authors, plus five new annotators. As in experiment 3, we computed inter-annotator agreement metrics for all combinations of annotators in experiment 4a. For all 21 pairs of annotators, agreement ranged from a low of 0.15 to a high of 0.32.

In addition to much variation across annotators, we find wide variation depending on the individual units consisting of a single image and all text asscoiated with the image. For the ten units, agreement ranged from 0.12 to 0.40.

4.5 Initial Results of Large Scale Human Labeling

A key feature of our new labeling interface is that labelers can work concurrently on distinct labeling tasks. We plan to collect data on between six and ten datasets. We have currently collected labelings from six annotators on the first dataset consisting of 25 images (45 paragraphs, 313 sentences).

In the new interface, annotators can choose any number of labels. We recruited four new labelers, and used one previous labeler (not a co-author). Results for the first dataset, which consists of 25 images and 48 associated paragraphs (313 sentences), are better than experiments 3 and 4 where we also used multiple annotators. We believe the improvement is due to the training provided in the interface, and the lack of constraint on the number of labels annotators could pick.

As in experiment 4, sentence labeling had a higher agreement than for paragraphs. For sentences the overall αmeasure was 0.45, compared with 0.40 for paragraphs. For all combinations of 2 to 5 coders, paragraph labeling agreement ranged from 0.56 to 0.27, and ranged from 0.55 to 0.33 for sentences. Again, the two coders who are experts in the area of image indexing had the highest interannotator agreement. As in the pilot studies, there was a significant variation in agreement, depending on the unit, ranging from a high of 0.70 to a low of 0.16.

The most frequent label combination for both paragraphs and sentences was the single label Image Content. There were 47 distinct combinations of labels for sentences, of which 34 were label pairs and five were triples; the remaining 8 unigram labels were the seven labels plus the default ”Other”. There were 38 combinations for paragraphs: 7 singletons, 20 pairs, 10 triples, and 1 combination of four labels.

5 Preliminary Machine Learning Results

Using data from our pilot studies of human labeling, augmented by an additional set of images labeled by one of the co-authors, we investigated the learnability of three categories: Image Content, Historical Context and Implementation. There were insufficient examples from the other categories. All learning was done using WEKA [13], a Java-based toolkit that implements a wide range of machine-learning algorithms using a standard input format.

A typical text classification task involves topic or subject matter identification. In contrast, our task involves functional categories that are more abstract than topic. The novel nature of our classes (and source texts) warranted an exploratory phase prior to conducting large scale machine learning experiments. Our pilot machine learning dataset permitted us to explore features for text representation, and to experiment with a range of learning techniques.

To establish a baseline, we started with a classic text classification approach: Naive Bayes classifiers using unigram word counts to represent the input texts, with separate classifiers for each semantic category. We used chi-square independence tests to prune the word vectors in order to reduce the noise in the training data, which smaller corpora such as ours are sensitive to. The Image Content classifier performed best, presumably due to the relatively greater amount of training data.

We experimented with two other types of text representation; One drew on hand-selected features from approximately half a dozen semantic domains. For example, we created sets of key words and phrases characteristic of the art history domain, (e.g., *masterpiece*), and other words and phrases deemed discriminative by our experts and annotators, such as terms denoting parts of the human body. Input texts were represented using raw counts for these hand-picked features. Our third text representation combined hand-picked features with the unigram word count representation.

We tested several learning methods in addition to Naive Bayes (NB), including support vector machines and tree-based learners. NB performed best overall. On train-and-test validation, the highest classification accuracy was on Image Content using the third type of text representation, with 83% accuracy. Using the same representation and learner, accuracy was 63% for Historical Context and 53% for Implementation. The best Historical Context classifier was NB with the first text representation, where accuracy was 70%. The NB classifiers performed poorly for our third category, Implementation. For this category, performance improved significantly using Random Forest, one of the tree-based learning methods we tried. In more recent work, we have achieved similar performance on the same dataset using a wider range of automatically derived features. For example, we now use part of speech tagging to identify present tense, which helps discriminate Image Content from the other two classes. We are currently experimenting with WordNet’s semantic distance metrics to weight terms with respect to various semantic domains, such as appearance, physical dimension, and so on.

6 Conclusions and Future Work

We have presented a detailed analysis of the development of a functional semantic labeling for art history texts, and have identified some of the problems that arise in achieving consistently high agreement scores

across multiple annotators. One issue, the variance across texts, is more difficult to address. The other key issue is that annotators with expertise are much more consistent with each other than non experts. As we continue collecting data, and updating our training with the expert consensus on previously labeled examples, we hope to learn something about training and experts. However, we have found that we can still achieve high accuracy with machine learning. As pointed out in [12], the relationship between interannotator agreement and learnability is not a predictable one.

We believe the initial learning results are quite promising. One difficulty for learning functional semantic categories is that many of the content words are not relevant features, since they will be different for descriptions of different images. In contrast, for topical text classification, content words are often sufficient for automatic classification, which is the intuition behind approaches such as latent semantic indexing. By using features such as verb tense, which distinguishes the **Image Content** class from others, we have achieved high results on relatively small datasets. On the other hand, since our categories are functional, they may transfer more easily to texts that are substantially different from our training and test materials.

As illustrated in the introduction, we anticipate that classifying text into functional semantic categories can provide more control over selection of metadata. Our categories have a rough correspondence with categories discussed in the image indexing literature [7, 3, 2]. As a result, it should be possible to map between our categories and the types of controlled vocabularies used in university visual resource centers. The external knowledge sources our project has examined include the three Getty resources (Art and Architecture Thesaurus, Thesaurus of Geographic Names, Union List of Artist Names), the Library of Congress Authorities and Library of Congress Thesauri for Graphic Materials, and ICONCLASS, a library classification for art and iconography.

References

1 R. Artstein and M. Poesio. Kappa^3 = alpha (or beta). Technical Report NLE Technote 200501, University of Essex, Essex, 2005.

2 M. Baca. *Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information*. The Haworth Press, Inc., 2003. Available through Library Literature, last accessed July 25, 2006.

3 H. Chen. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, pages 260–273, 2001.

4 A. Giral and A. Taylor. Indexing overlap and consistency between the Avery Index to Architectural Periodicals and the Architectural Periodicals Index. Library Resources and Technical Services 37(1):19-44, 1993.

5 J. Klavans. Using computational linguistic techniques and thesauri for enhancing metadata records in image search: The CLiMB project. Article in preparation.

6 K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, 1980.

7 S. S. Layne. Some issues in the indexing of images. Journal of the American Society for Information Science, pages 583–8, 1994.

8 K. Markey. Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. Library and Information Science Research, pages 155–177, 1984.

9 R. Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.

10 R. Passonneau, N. Habash and O. Rambow. Inter-annotatator agreement on a multilingual semantic annotation task. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.

11 R. J. Passonneau, D. Elson, R. Blitz, and J. Klavans. CLiMB Toolkit: A case study of iterative evaluation in a multidis ciplinary project. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.

12 D. Riedsma and J. Carletta. Reliability measurement: there’s no safe limit To appear in Computational Linguistics.

13 I. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann: San Fransisco, 2000.

4.3.4 Judith Klavans, Tandeep Sidhu, Carolyn Sheffield, Dagobert Soergel, Jimmy Lin, Eileen Abels, Rebecca Passonneau. **Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata**. *VISAPP Workshop Metadata Mining for Image Understanding*. MMIU 2008. Funchal, Madeira - Portugal.

Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata

Judith L. Klavans[1,2], Tandeep Sidhu[1], Carolyn Sheffield[1], Dagobert Soergel[1], Jimmy Lin[1,2], Eileen Abels [3], Rebecca Passonneau [4]

[1]College of Information Studies (CLIS)
[2]University of Maryland Institute for Advanced Computer Science (UMIACS)
University of Maryland, College Park, Maryland
[3]College of Information Science and Technology, Drexel University, Phila PA,
[4]Center for Computational Learning Systems, Columbia University, NY NY

Abstract. In this paper, we present a fully-implemented system using computational linguistic techniques to apply automatic text mining for the extraction of metadata for image access. We describe the implementation of a workbench created for, and evaluated by, image catalogers. We discuss the current functionality and future goals for this image catalogers' toolkit, developed in the Computational Linguistics for Metadata Building (CLiMB) research project.[6] Our primary user group for initial phases of the project is the cataloger expert; in future work we address applications for end users.

1 The Problem: Insufficient Subject Access to Images

The CLiMB 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, at best. In an initial observational study conducted with six image catalogers, we found that typically 1 – 8 subject terms are assigned, and many legacy records lack subject entries altogether.

The literature on end users' image searching indicates that this level of subject description may be insufficient for some user groups. In a study of the image-searching behaviors of faculty and graduate students in American history, Choi and Rasmussen [3], found that 92% of the 38 participants considered the textual information associated with the images in the Library of Congress' American Memory Collection inadequate. The number of subject descriptors assigned to an image in this collection is comparable to or exceeds those found in the exploratory CLiMB studies. Furthermore, these searchers submitted more subject-oriented queries than known artist and title searches. Similar results demonstrating the importance of subject retrieval have been reported in other studies, including Keister [6], Collins [4], and Chen [2].

2 Solutions

[6] This project was first funded by the Mellon Foundation to the Center for Research on Information Access at Columbia University.

The CLiMB project was initiated to address the subject metadata gap under the hypothesis that automatic and semi-automatic techniques may enable 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 technologies to identify, filter, and normalize high-quality subject descriptors. Like Pastra et al. [10] 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.

For this project, we use the standard Cataloging Cultural Objects (CCO) definition of subject metadata[7]. According to this definition, the subject element of an image catalog record should include terms which provide "an identification, description, or interpretation of what is depicted in and by a work or image." The CCO guidelines also incorporate instructions on analyzing images based on the work of Shatford-Layne (formerly Shatford). Shatford [12], building on Panofsky [8], proposed a method for identifying image attributes, which includes analysis of the generic and specific events, objects, and names that a picture is "of" and the more abstract symbols and moods that a picture is "about". Panofsky describes the pre-iconographic, iconographic, and iconologic levels of meaning found in Renaissance art images. Shatford's generic and specific levels correspond to Panofsky's pre-iconographic and iconographic levels, respectively, and encompass the more objective and straightforward subject matter depicted in an image. The iconologic level (Shatford's about) addresses the more symbolic, interpretive, subjective meanings of an image. To aid user access, catalogers are encouraged to consider both general and specific terms for describing the objective content of an image as well as to include the more subjective iconologic, symbolic, or interpretive meanings. Iconologic terms may be the most difficult for catalogers to assign but occur often in texts describing images.

3 Preparatory Studies of Cataloging

In order to get a better sense of the cataloging process and to inform our system design, we conducted studies on the process of subject term selection by image catalogers. Our goal was to collect data on the process as a whole in order to improve both our system function (either through rules or statistical methods) and our system functionality (i.e. how to incorporate our results into an existing workflow and how to perhaps replace a portion of the workflow with automatic techniques). In this section, we discuss two of these formative studies.

The first study was designed to identify the types of subject terms a cataloger may assign to a given image. Identifying these expert term assignments will help guide the development of heuristic rules for automatically identifying high-quality descriptor candidates and filtering out term types which are rarely assigned manually. Participants were given four stimuli:

1) a hypothetical query for an image;

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

2) an image;
3) another image—this time with associated text; and
4) an image paired with a list of CLiMB-extracted terms.

For the first two stimuli, catalogers were asked to generate subject terms. For the third and fourth stimuli, catalogers were asked to select terms from the associated text or list of terms. We selected four image/text pairs from the National Gallery of Art Collection. To control for varying textual content which may occur with different image types, we chose one landscape, one portrait, one still life, and one iconographic image; we employed a Latin Square design. Twenty image catalogers recruited through the Visual Resource Association participated in the study. Through a combined quantitative and qualitative approach, we analyzed

- the number of terms assigned per task,
- the types of terms assigned, and
- the level of agreement between catalogers in terms used for the same concept (to be discussed in a future publication).

Table 1: Distribution of term assignments by category.

Terms assigned for landscape image, Task 2	Category
Gauguin	artist name
pea green	Color
Orange	Color
black and white	Color
Cow(s) / dairy cows / cattle	Figures/Objects
stacks of hay / mounds of hay/ bales / hay	Figures/Objects
Crops	Figures/Objects
Herding	Figures/Objects
woman herding	Figures/Objects
capped woman	Figures/Objects
Poppies	Figures/Objects
Rocks	Figures/Objects
white dress	Figures/Objects
19th Century	Period
fields/vegetable field	Place
France	Place
Brittany	Place
Dutch landscape/Dutch countryside; paintings and landscapes / paintings and countrysides	Type
sea/seascape or canal	Type

In analyzing the types of terms catalogers assigned, we identified seven categories (in order of frequency): figure/object, place, artist names, period/date, type, style and color. Table 1 above shows a subset of results from catalogers completing just one of the tasks for the landscape image. Results for landscape art across all four tasks yielded 13 terms for figure/object, 9 for place, 8 for artist names, 7 for period/date, 6 for type, and 4 for style and color. This distribution will help guide the priorities placed on term selection.

For the second study, we took a broader look at the overall image-indexing workflow, including standards, local policies, and actual practices, to determine how the CLiMB Toolkit fits into the cataloging process as a whole. This study not only enabled us to define interface parameters and necessary functionality, it also confirmed the lack of subject access currently provided by human indexers. We examined the similarities and differences in image cataloging practices both within a single institution and across three separate institutions. By observing catalogers as they indexed images from their respective collections, we also investigated the number and types of subject terms added per catalog record. Within and across these academic visual resource centers, we found that general practices and workflow patterns varied little, and that the number of subject terms entered per catalog record varied but typically fell somewhere between one and eight. One of the primary differences across institutions was the use of different software and metadata schemas, some of which were locally developed. These results indicate that, with flexible export functionality built in to a generic workbench, the CLiMB Toolkit should integrate smoothly with existing practices and different work environments, with little or no tailoring required.

4 CLiMB Architecture and Interface

This section describes the techniques we have developed to semi-automatically identify terms which qualify as potential subject descriptors. Our techniques exceed simple keyword indexing by:

- applying advanced semantic categorization to text segments,
- identifying coherent phrases,
- associating terms with a thesaurus, and
- applying disambiguation algorithms to these terms.

CLiMB combines new and pre-existing technologies in a flexible, client-side architecture which has been implemented into a downloadable toolkit, and which can be tailored to the user's needs. Figure 1 shows the overall architecture of the CLiMB Toolkit. The upper left shows the input to the system, an image, minimal metadata (e.g. image, name, creator), and text. To date, we have input six testbed collections, described more fully in Section 5.

The first stage of CLiMB's processing pipeline associates portions of the input text with images. Note that this requires segmentation, and association of segmented text with the image being described. In clear cases, such as online image captions or in exhibition catalogs, association of image with text is a given. However, in cases where there is a more diffuse relationship between text and image (as in art history texts, for example), it is a computational challenge to ensure that text is accurately associated with the correct image, and not with an image in close proximity (which may or may not be described by the text). This logic creates associations between text and image based on explicit references in the text, rather than taking any text in proximity of an image.

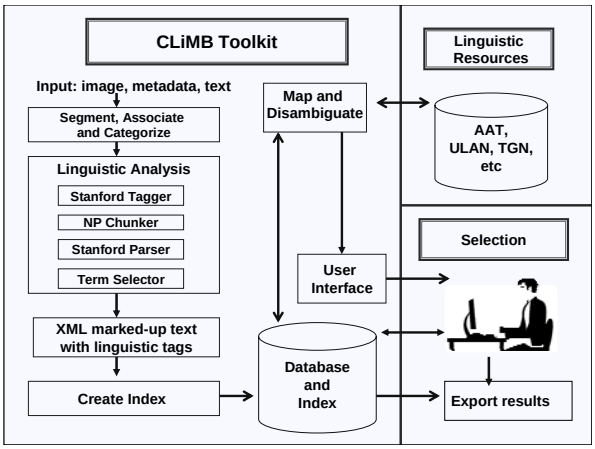


Figure 1: CLiMB Architecture.

In addition to segmentation, we are developing methods to categorize spans of text (e.g., sentences or paragraphs) as to their semantic function in the text. 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. [9].

The next phase, Linguistic Analysis, consists of several subprocesses. After sentence segmentation, a part-of-speech (POS) tagger labels (i.e. tags) the function of each word in a text, e.g., noun, verb, preposition, etc. Complete noun phrases can then be identified by the NP chunker based on tag patterns. For example, a determiner, followed by any number of adjectives, followed by any number of nouns, is one such pattern that identifies a noun phrase, as in "the impressive still life drawing". The tagger used for CLiMB, the Stanford tagger[8] provides sentential analysis of syntactic constructions, e.g., verb phrases, relative clauses. The output of Linguistic Analysis consists of XML-tagged words which contain substantial part of speech and syntactic parsed labels. Lucene is used to create an index for these tagged words.[9]

At this point, the noun phrases stored in the index are input to the disambiguation algorithm, which then enables sense mapping, so that the proper descriptor can be selected from a controlled vocabulary. Words and phrases often have multiple meanings which correspond to different descriptors. The ability to select one sense from many is referred to as lexical disambiguation. Currently, we map to descriptors from the Getty Art and Architecture Thesaurus (AAT), the Getty Union List of Artist Names (ULAN), and the Getty Thesaurus of Geographic Names (TGN).[10] The AAT is a well-established and widely-used multi-faceted thesaurus of terms for the cataloging and indexing of art,

[8] Both the tagger and parser are available at: http://nlp.stanford.edu/software.
[9] Lucene is a search engine library: http://lucene.apache.org.
[10] Getty resources can be accessed at: getty.edu/research/conducting_research/vocabularies/aat

architecture, artifactual, and archival materials. AAT has 31,000 such records and among those, there are 1,400 homonyms, i.e., records with same preferred name. For example, the term "wings" has five senses in the AAT. Each sense falls under distinct but separate subdomains of art and architecture, ranging from building divisions and theater spaces to costume accessories, furniture components, and visual works components.

Table 2 shows the breakdown of the AAT vocabulary by number of senses with a sample lexical item for each frequency. As with most dictionaries and thesauri, most items have two to three senses, and only a few have more.

Table 2: Scope of the disambiguation problem in the AAT Thesaurus.

# of Senses	# of Terms	Example
1	29,576	Scaglioni
2	1097	Bells
3	215	Painting
4	50	Alabaster
5	39	Wings
6 -7	14	Boards
8+	9	Emerald Carmine

First, we use all modifiers that are in the noun phrase to find the correct AAT record (Lookup Modifier). We search for the modifiers in the record description, variant names, and the parent hierarchy names of all the matching AAT senses. If this technique narrowed down the record set to one, then we found our correct record. For example, consider the term "ceiling coffers." For this term we found two records: "coffers" (coffered ceiling components) and "coffers" (chests). The first record has the modifier "ceiling" in its record description, so we were able to determine that this was the correct record. Next, we use SenseRelate to help select the correct WordNet sense of the noun phrase (or its head noun). Using that sense definition from WordNet, we next examined which of the AAT senses best matches with the WordNet sense definition. For this, we used a word overlapping technique which takes senses of WordNet for each polysemous term in AAT and selects the highest value of word overlaps. If none of the AAT records received any positive score (above a threshold), then this technique could not find the best match. Other techniques, Best Record Match and Most Common Sense, are presented in Sidhu et al. [13].

For evaluation of the disambiguation model, we followed standard procedure in word sense disambiguation tasks (Palmer et al. [7]). Two labelers manually mapped 601 subject terms to a controlled vocabulary. Inter-annotator agreement for this task was 91%, providing a notional upper bound for automatic system performance (Gale et al. [5]) and a dataset for evaluation. We used SenseRelate (Banerjee and Pederson [1], Patwardhan et al. [11]) for disambiguating AAT senses. SenseRelate uses word sense definitions from

WordNet 2.1, a large lexical database of English words.[11] The impact of using a general vocabulary such as WordNet compared to specialist vocabularies is an empirical issue which we are examining in current research.

Table 3 shows results of running different techniques on this data. Row 1 shows how few terms were mapped by the lookup modifier technique; only one was mapped for the Training Set. Rows 2 and 3 show that the SenseRelate technique was most successful in labeling terms, followed by the Best Record Match technique. The Most Common Sense technique (Row 4) was also poor. An analysis of results and errors shows that our overall accuracy is between 50-55% compared to 70% common in general disambiguation. In future work, we will explore re-implementing concepts behind SenseRelate to directly map terms to the AAT and additional approaches using hybrid techniques (including machine learning) for disambiguation. Currently, we are awaiting results from manual disambiguation tests with human catalogers before refining and integrating the module. Our plan is to use results to rank and select a sense for mapping that the user will confirm; once we collect enough feedback, we can apply learning to eliminate senses with greater confidence than at present.

Table 3: Breakdown of AAT mappings by Disambiguation Technique.

	Technique Name	Training (n=128)	Test (n=96)
1	Lookup Modifier	1	3
2	SenseRelate	108	63
3	Best Record Match	14	12
4	Most Common Sense	5	18

Figure 2 shows a screen shot of the CLiMB user interface, after having performed a search over the National Gallery of Art collection, and having run the text through the Toolkit.[12] Note that the center top panel contains the image, so the user can look at the item to be described. The center panel contains the input text, with proper and common nouns highlighted. Under this is the term the user has selected to enter. The right-hand panel gives the thesaural information. At the top of the right are the two senses for the word "landscape" with an indication of where they occur in the AAT hierarchy. Next is the text description of the sense selected. Finally, the entire hierarchy is displayed, bottom right, for the user to view and identify any related terms.

As part of the evaluation of the CLiMB approach, 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 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, the same term may refer to a distinct section of a wall, within a border or frame.

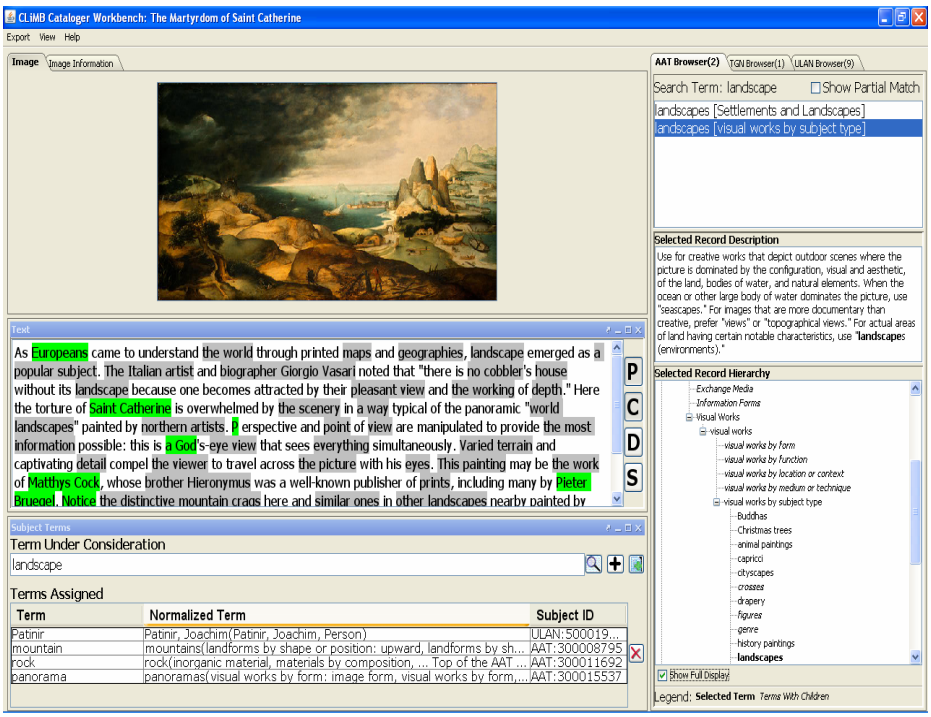
[11] http://wordnet.princeton.edu/

[12] In the interest of space, we have included a full screen shot, accompanied by text explanations. If reviewers prefer, this can be enlarged or split into two Figures.

We are currently working with five image-text sets and one image collection for which we are conducting experiments with dispersed digital texts. These six collections will be used for different phases of evaluation, discussed under Future Work. The texts and images for two of the collections, the National Gallery of Art (NGA) Online Collection and the U.S. Senate Catalogue of Fine Arts, can be found online and are in the public domain. For three of the other image collections, The Vernacular Architecture Forum (VAF)[13], The Society of Architectural Historians (SAH)[14], and The Landscape Architecture Image Resource (LAIR)[15], we have secured digital copies of relevant texts along with permissions for use in our testing. The final collection is the Art History Survey Collection, made available to us through ARTstor[16].

5 Conclusions and Future Work

We are working in a challenging domain with a highly specialized vocabulary. Currently we depend on the external program SenseRelate to perform much of the disambiguation. Furthermore, SenseRelate maps terms to WordNet and we then map the WordNet sense to an AAT sense. This extra step is overhead, and it causes errors in our algorithm. We are looking to incorporate additional domain-specific vocabularies for future testing, rather than more general resources which add noise. Sources under consideration include ICONCLASS and the Library of Congress' Thesaurus for Graphic Materials I & II.



[13] http://www.vernaculararchitectureforum.org/
[14] www.sah.org/
[15] www.lair.umd.edu/
[16] www.artstor.org

Figure 2: **CLiMB user interface for term "landscape".**

For future work, we have also designed a series of studies to test the toolkit in situ. We have partners from several museums and libraries, mentioned in Evaluation, that will test CLiMB with their cataloging staff, and then work with us to design evaluations of Toolkit success in three areas:

1) staff perception on Toolkit ease of use for cataloging within their collections;
2) end user satisfaction with these enhanced records; and
3) several component evaluations, including the named entity recognizer, the noun phrase selector, and the disambiguation component.

The proverbial tradeoff between precision and recall may vary for different sectors of the image community; we believe our research in different venues will provide insights on this critical issue. Finally, we intend to explore new directions for integrating CLiMB with current social networking technologies, including social tagging, trust-based ranking of tags, and recommender systems. These technologies address end user needs and offer CLiMB the potential to achieve more personalized results.

6 Acknowledgements

We acknowledge: the Mellon Foundation; Dr. Murtha Baca, director of the Getty Vocabulary Program and Digital Resource Management, Getty Research Institute for providing us with research access to resources; collections partners, including Jeff Cohen, Bryn Mawr College and University of Pennsylvania; Jack Sullivan, University of Maryland; the Senate Museum and Library and ARTStor.

References

1. Banerjee, S., Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, (2003) 805–810 [7]
2. Chen, H.: An Analysis of Image Retrieval Tasks in the Field of Art History. Information Processing & Management, Vol. 37, No. 5 (2001) 701-720
3. Choi, Y., Rasmussen, E. 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 (2003) 498-511
4. Collins, K.: Providing Subject Access to Images: A Study of User Queries. The American Archivist, Vol. 61 (1998) 36-55
5. Gale, W., Church, K., Yarowsky, D.: A Method for Disambiguation Word Senses in a Large Corpus. Computers and Humanities, Vol. 26 (1993) 415-439.
6. Keister, L.H.: User Types and Queries: Impact on Image Access Systems. In: Fidel, R., Hahn, T.B., Rasmussen, E., Smith, P.J. (eds.): Challenges in Indexing Electronic Text and Images. Learned Information for the American Society of Information Science, Medford (1994) 7-22
7. Palmer, M., Ng, H.T., Dang, H.T.: Evaluation. In: Edmonds, P., Agirre, E. (eds.): Word Sense Disambiguation: Algorithms, Applications, and Trends. Text, Speech, and Language Technology Series, Kluwer Academic Publishers (2006)
8. Panofsky, E. Studies in Iconology: Humanistic Themes in the Art of the Renaissance. Harper & Rowe, New York (1962)
9. Passonneau, R., Yano, T., Lippincott, T., Klavans, J. 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 (2008)
10. Pastra, K., Saggion, H., Wilks, Y.: Intelligent Indexing of Crime-Scene Photographs. In: IEEE Intelligent Systems: Special Issue on Advances in Natural Language and Processing, Vol. 18, Iss. 1. (2003) 55-61.
11. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2003)

12. Shatford, S.: Analyzing the Subject of a Picture: A Theoretical Approach. Cataloging & Classification Quarterly, Vol. 6, Iss. 3 (1986) 39-62
13. Sidhu, T., Klavans, J.L., Lin, J.: Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTech 2007), 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic (2007)

4.3.5 Rebecca Passonneau, Roberta Blitz, David Elson, Angela Giral, Judith Klavans. CLiMB ToolKit: A case study of iterative evaluation in a multidisciplinary project. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC) Genoa, Italy. May 24-26, 2006

CLiMB ToolKit: A Case Study of Iterative Evaluation in a Multidisciplinary Project

Rebecca Passonneau, † Roberta Blitz, † David Elson, † Angela Giral, † Judith Klavans‡

Columbia University†
New York, NY, U.S.A.
(becky|delson)@cs.columbia.edu, (rlb179|giral)@columbia.edu

University of Maryland‡
College Park, MD, U.S.A.
klavans@casl.umd.edu

Abstract
Digital image collections in libraries and other curatorial institutions grow too rapidly to create new descriptive metadata for subject matter search or browsing. CLiMB (Computational Linguistics for Metadata Building) was a project designed to address this dilemma that involved computer scientists, linguists, librarians, and art librarians. The CLiMB project followed an iterative evaluation model: each next phase of the project emerged from the results of an evaluation. After assembling a suite of text processing tools to be used in extracting metadata, we conducted a formative evaluation with thirteen participants, using a survey in which we varied the order and type of four conditions under which respondents would propose or select image search terms. Results of the formative evaluation led us to conclude that a CLiMB ToolKit would work best if its main function was to propose terms for users to review. After implementing a prototype ToolKit using a browser interface, we conducted an evaluation with ten experts. Users found the ToolKit very habitable, remained consistently satisfied throughout a lengthy evaluation, and selected a large number of terms per image.

1. Introduction

Digital image collections in libraries and other curatorial institutions grow too rapidly to create new descriptive metadata for subject matter search or browsing (Blitz et al., 2004). CLiMB (Computational Linguistics for Metadata Building) was a project designed to address this dilemma that involved computer scientists, linguists, librarians, and art librarians. The goal was to explore the use of computational linguistic techniques to automatically extract descriptive metadata from electronic or scanned texts (Klavans et al., 2004) (Klavans, In submission).

The CLiMB project followed an iterative evaluation model, with each next phase of the project emerging from the results of an evaluation. Here we describe the two phases that immediately preceded and followed the development of a prototype CLiMB ToolKit. After testing and selecting a suite of computational linguistic tools, such as noun phrase chunkers, and knowledge sources such as external vocabularies, a formative evaluation was conducted to determine how best to present such tools to image professionals with cataloging needs. The evaluation results suggested that the most useful tool would be one that would allow catalogers to interactively process texts so as to select metadata from automatically proposed text elements.

Following the formative evaluation, we developed the CLiMB ToolKit, a browser interface to the text processing tools and knowledge sources, and presented a live demonstration as part of a CLiMB presentation at the March 2004 meeting of the Visual Resources Association. Largely through this presentation, we were able to recruit a balanced group of ten librarians and image professionals to participate in an April evaluation of the ToolKit. On measures of task success and user satisfaction, the ToolKit rated very highly. In addition, there was a high degree of group ratification of proposed metadata. Figure 1 shows the fourteen top ranked terms from a total of 27 completely distinct terms proposed by the ten participants for Jan Jansz den Uyl's painting, the *Banquet Piece*. This set of terms provides a good

indication of the subject matter of the painting, both explicit (*burned down candle, empty glass*) and implicit (*vanitas, five senses*). They also have a high degree of overlap with the Getty Art & Architecture Thesaurus (AAT), [1] a controlled vocabulary that users could get to from the ToolKit.

[1]Columbia Libraries had a site license for the AAT.

Our experience in designing, implementing and evaluating the CLiMB ToolKit demonstrates the benefits of conducting evaluations before and during the design process. It also illustrates the richness of metadata thatcombines ordinary language phrases (*burned down candle, empty glass, white linen tablecloth*) with terminology from a controlled vocabulary.

Rank Term Freq
1 *vanitas (image)* 9
2 *(Dutch) still life (painting)* 8
3 *(burned down) candle* 7
4 *glass* 6
5 *pewter* 6
6 *Dutch* 5
7 *empty glass* 5
8 *((white) linen) tablecloth* 5
9 *lute* 5
10 *five senses* 4
11 *(luxury) tablewares* 4
12 *owl* 4
13 *painting* 4
14 *seventeenth century* 4

Figure 1. Top fifty percent of ranked subject terms proposed for Jan Jansz den Uyl's *Banquet Piece.*

Section two provides the motivation for addressing the problem of supplying descriptive metadata for images, and for the image collections and texts used in our evaluations. Section three gives a brief overview of the tools and knowledge sources experimented with prior to the formative evaluation, which is described in section four. Section five gives a brief overview of the ToolKit, and we describe the evaluation of the ToolKit in section six. We conclude with a brief summary of results.

2. Motivation

When librarians create descriptive metadata for texts by analyzing the subject matter, they face complex issues from a knowledge representation point of view, and significant expense. Subject matter analysis of cultural artifacts faces similar complexities (Bacca, 2002). Recent estimates in the Columbia Libraries, for example, give a minimum cost for descriptive metadata of $15.00 per volume (Blitz et. al, 2004). Using this as an estimate, adding subject matter metadata to Columbia's Image Bank of 32,000 images (1/10th the size of ArtStor http://www.artstor.org/info/), would cost about $500,000. The premise of CLiMB is that descriptive content for certain digital collections exists implicitly in scholarly monographs and other publications that discuss the images or the cultural artifacts they depict. From this premise, the goal emerged of finding a method to make the descriptive content explicit, link it to appropriate images, and allow image librarians and visual resource professionals make use of it in creating metadata.

Empirical research (Fawcett, 1979; Mandel, 1985; Taylor, 1995) has shown that even in academic contexts, many searches for images involve requests for content, such as images of "crowds" or "a balloon frame house." Defining what type of image collection would be a good starting point, and determining criteria for texts to associate with an image collection, turned out to be two of the key problems addressed in the earliest phases of the CLiMB project. Catalogers of digital image collections can face quite distinct problems in selecting descriptive metadata, depending on the structure, content and cohesiveness of the collection. Computational linguists similarly face distinct engineering and research challenges, depending on text characteristics such as the size of the vocabulary, the complexity of the sentence structure, and the principles that organize the text(s).

To limit the complexities CLiMB would face, two decisions were made regarding the image collections and texts. Because computational linguistic tools generally need to be tuned to different domains and text types, it was decided that a small number of carefully selected texts with the potential to relate to a large number of images would maximize the opportunity to automatically or semi-automatically identify subject metadata. The second decision was to narrow the subject areas to two: architecture and art.

3. Tools and Knowledge Sources for Processing CLiMB Texts

To provide an independent method of indexing the images and cultural artifacts that CLiMB would find subject metadata for, the project defined the need for a list of Target Object Identifiers (TOIs), or authoritative names that would be uniquely indexed within the context of the tools. Software written expressly for the project included a tool for automatically generating variants of authoritative names (Davis et al., 2003), referred to as the TOI finder, and a text segmenter based on frequency of hits from the TOI finder. In addition, LTChunk (Finch & Mikheev, 1997) is used to tag noun phrases. External word lists, such as back of book indices from the texts associated with collections, and vocabularies in the art and architectural domain, form part of the CLiMB knowledge sources.

4. Formative Evaluation

The formative evaluation took place after we had implemented a research version of CLiMB tools. It was conducted concurrently with a meeting with an External Advisory Board of experts from library science, art history, computer science and computational linguistics. The meeting served as a vehicle for the CLiMB team to become informed about issues that others with greater expertise could better anticipate, and for us to educate the experts regarding the outcome of our earlier investigations of the collection/text issues. It included our evaluation survey in which respondents were asked to propose terms for images from two testbed image collections, under four conditions. Thirteen participants completed the survey.

The survey included images from the Greene & Greene Collection of Architectural Records and Papers, Avery Architectural and Fine Arts Library (G&G); and the Anne S. Goodrich Collection of Chinese Paper Gods, C.V. Starr East Asian Library (CPG). The survey questions were distributed across four tasks that were presented in different orders to three different groups of respondents. Two of the tasks (free text, checklist) required associated texts. The text associated with (G&G) was Edward R. Bosley's book *Greene & Greene* (London: Phaidon, 2000). The text associated with CPG was Anne S. Goodrich's book, *Peking Paper Gods: A Look at Home Worship* (Nettetal: Steyler Verlag, 1991).

1) **User Scenario**: In this task, the survey item contained the following hypothetical user scenarios. Respondents were asked to list keywords and phrases that could be used "to search for relevant images in an image database."
1. I am writing a paper on domestic architecture in Southern California in the early part of the 20th century. I was told that there are homes with exteriors clad in a type of concrete or cement. How can I locate images?
2. I am trying to locate an image of the Buddhist goddess of compassion. I can't remember the name but I know this deity, widely worshipped by women in China, originated as a male figure in India. She is often portrayed wearing a headdress, attended by other figures, and often some type of plant is depicted. Can you help me find a picture?

2) **Image**: This survey item contained an image. Respondents were given the following instructions: "*Please write keywords and phrases that you would use to find this image in a database. You may write as many as you wish*."

3) Free Text: This task contained a passage from one of the texts associated with G&G or CPG. Respondents were asked to
"*Suppose there is a collection of related images that needs metadata keywords and phrases. Please select the words and phrases in this text that you feel would be good metadata for the images.*
1. Please circle 10 words or phrases as your top choices.
2. Please underline 10 as your second tierchoices."

4) **CLiMB Checklist**: Respondents were given a long list of words and phrases (194 G&G entries; 117 CPG entries) that had been extracted by CLiMB tools from the same texts presented in Task 3. Instructions were:
"*Please check off the words and phrases that you feel would be suitable metadata for the images in the collection*."
Fewer terms were returned for the user scenario survey item than for items two through four. This reflects the relative lack information from which to construct a query. Three of the respondents referred directly or indirectly to a hypothetical "reference interview" that would be conducted as part of the "user scenario," which we took as an indication that the scenario alone provided insufficient context to generate image search terms.

For the image condition, respondents were instructed to list words and phrases for a keyword search. The survey layout had seven bulleted lines where respondents could list their responses. To some degree, this predisposes respondents to try to produce something on the order of half a dozen terms. However, the instructions and layout were interpreted differently by different respondents. Some provided a list of individual terms, the implication being that any combination of these could be used in a Boolean search. Some provided lists of strings of terms, occasionally with minor variations within the list, the implication being that other possible combinations of the same terms were deliberately omitted. On average, respondents found 10 terms for G& and 7.6 terms for CPG.

The next two conditions generated many more terms than the user scenario or image conditions, and they were more specific. For the free text and checklist conditions, we found 82 distinct terms used for G&G, and 56 for CPG. Most terms were suggested by at most a single respondent, but there were fourteen suggested by multiple respondents. Importantly, there was a significant overlap of terms selected many humans, and terms with high weights assigned by CLiMB tools. Examples of terms from the first two conditions are "home", "exterior," "brick" and "driveway." Examples of terms from the text and checklist conditions include "garden pergola," "dark green tile," "plaster frieze," "ridge beams."

The smaller number and lower specificity of terms provided for the image and free text items showed how difficult users found generation of items de novo. In other words, the value of using text already describing an image was shown. We found that experts selected terms differently from non-experts, which indicated to us that the tools should be used by catalogers and image experts. We concluded that a CLiMB ToolKit would work best if its main function was to propose terms for users to review.

5. ToolKit

The ToolKit was implemented in a browser in order to use functionality that most users would already be familiar with, and to avoid interoperability issues. Users were given a high degree of control without having to understand CLiMB rules. Two prerequisites for using the CLiMB ToolKit were to have already constructed a TOI list, and to have associated texts in electronic format, either scanned text or text in an XML format such as TEI Lite. However, at the time of the Prototype Evaluation, the ability to handle TEI markup and integrate it with the rest of the tools had not yet been completed.

Figure 2. Text loading page of the CLiMB ToolKit

In the evaluation version of the ToolKit (0.9), the following functions had been implemented:
1. Loading and initialization of raw (ASCII) text was available;
2. After initialization, text could be processed by a noun phrase chunker (termed "chunking") that locates the beginnings and ends of noun phrases (such as the underlined expressions in this sentence from a *North Carolina Museum of Art: Handbook of the Collections* passage about Jan Jansz den Uyl's painting entitled *Banquet Piece*: "At the same time, symbolically charged elements such as the empty glass, burned-down candle, and lute at the far left hint at deeper meanings.");
3. A TOI list could be loaded, or TOIs could be manually created;
4. The TOI Finder could be run to locate references to TOIs in the loaded texts;
5. Texts that had been processed by the TOI Finder could also be sectioned into associational contexts correlated with specific TOIs;

6. Lists of Controlled Vocabulary could be loaded—included in this feature users were provided access to the Getty Art & Architecture Thesaurus (AAT), and the capability of selecting specific subsets from the AAT;
7. A Noun Phrase detail frame was available, e.g., to illustrate intersections of text phrases with AAT.
In sum, participants in the evaluation could load and process texts and TOIs, could view text in a variety of ways in order to locate relevant sections (for instance, text a specific image or TOI), and isolate noun phrases likely to be good candidates for metadata terms.

Text of question:
3.2 Finding another access point to the "project" help text was:
1. Completely obvious, . . . , **5.** Pretty difficult

6.6 How long did it take you to figure out how to close the project?
1. No time, . . . , **5.** Too much time

7.7 I _____ what happens when I click on a table heading.
1. Really like, . . . , **5.** Don't understand

8.8 So far, the concept of a CLiMB TOI is:
1. Very clear, . . . , **5.** Pretty confusing

9.9 Entering a new TOI:
1. Was easy, . . . , **5.** Was difficult

9.10 So far, my opinion of the look and feel of the CLiMB Toolkit is:
1. Great, . . . , **5.** Not so good

11.12 Figuring out how to view the text:
1. Was easy, . . . , **5.** Was difficult

11.13 Changing the text display options:
1. Was easy, . . . , **5.** Was difficult

11.15 So far, my opinion of the look and feel of the CLiMB Toolkit is:
1. Great, . . . , **5.** Not so good

15.18 Understanding the notion of a CLiMB "project" is:
1. Very easy, . . . , **5.** Confusing

16.20 I was able to follow the above steps to get my new project to this point:
1. Very easily, . . . , **5.** With difficulty

16.21 I find it _____ to understand why the TOI-Finder applies to the whole project, and why the sectioner applies to an individual text in a project.
1. Easy, . . . , **5.** Difficult or impossible

17.22 I found these 4 substeps for finding AAT terms in the project texts:
1. Very easy, . . . , **5.** Difficult or impossible

20.32 So far, my opinion of the look and feel of the CLiMB Toolkit is:
1. Great, . . . , **5.** Not so good

20.33 I was _____ the process of selecting descriptive metadata.
1. very pleased with, . . . , **5.** very displeased with

Table 1. Text of the sixteen scaled questions

6. Evaluation of Prototype Toolkit

Ten librarians, image professionals and metadata professionals participated in a ToolKit evaluation using samples of images from two collections, the web version of the North Carolina Museum of Art

http://ncartmuseum.org/, and the Greene & Greene Collection of Architectural Images, Avery

[2]NCMA images were licensed by Columbia from Saskia, Ltd.
[3] http://www.columbia.edu/cu/libraries/inside/projects/CLiMB

Archictectural and Fine Arts Library, Columbia University.[2] A questionnaire[3] directed participants through the same series of steps that would be required in actual use. These included: loading and processing the texts we provided, browsing images in image galleries we provided, selecting subtrees from the AAT, running the CLiMB tools, and finally, selecting subject terms to associate with sample images. The questionnaire interleaved user actions with general questions about the user's satisfaction with the ToolKit, and with specific questions regarding satisfaction with individual steps in the process. There were 41 questions interleaved among 22 task steps; 18 of the questions asked users to select items from a 5-point scale from positive (1) to negative (5) assessments.

Our results include task success on the process of selecting metadata for the images, and user satisfaction measures (cf. Paradise model (Walker et al., 1997)). Regarding task success, all participants found descriptive metadata. For example, 96 distinct terms were selected for an image of a painting by Jan Jansz den Uyl called *Banquet Piece*, including half a dozen terms selected by seven or more of the ten participants.

All participants completed Parts I and II of the questionnaire. In Part I, users were given an introductory overview of the Toolkit, exemplified by a preloaded project. In Part II, users were asked to create their own metadata selection project using texts and TOIs for the *North Carolina Museum of Art: Handbook of the Collection* (*NCMA Handbook*). There was an optional Part III pertaining to the Greene & Greene Collection at Avery Architectural and Fine Arts Library, Columbia University; few users completed this part, and it is not discussed further.

Question AVG SD
3.2 1.8 1.09
6.6 1 0.00
7.7 1.9 0.60
8.8 2.1 1.05
9.9 2.2 1.32
9.10 2 0.76
11.12 1.2 0.33
11.13 1.2 0.33
11.15 2 0.78
15.18 1.9 0.60
16.20 3.3 1.07
16.21 2.8 1.56
17.22 2 0.78
20.32 2.1 0.38
20.33 1.4 0.55

Table 2. Sums of responses to scaled questions

6 .1. Overview of quantitative results

We give an overview of the quantitative results by examining three specific measures. 1) We present the average score on the 5-point scale questions, which provides a single, summary metric of user satisfaction with the Toolkit. 2) We compare the questions that had the highest and lowest scores, which gives a view of the range of responses, and also specifics on what features were most and least satisfactory. 3) We compare the responses to a repeated question about overall user satisfaction that was presented at three points during the evaluation (9.10, 11.15, 20.32; in boldface in Table 1 and Table 2).

Analysis of the questions in which responses fall on a 5-point scale, where 1 is the most positive, yields a quantitative view of the respondent's evaluations. We received answers from all participants on sixteen of these questions; however, respondents treated one of these questions as eliciting multiple responses, so we

quantify only the remaining fifteen questions. The average (AVG) for all responses on all scaled questions was 2.0 (the standard deviation, SD, a measure of variability around the average, is 1.8). This indicates that overall, people were satisfied with the experience of using the Toolkit.

Of greater interest than the average of all scaled responses is to compare the questions receiving the lowest and the highest scores. The question receiving the least positive score was 16.20, where the average answer was 3.3 (standard deviation=1.1, median=3), or just below average satisfaction (note that the standard deviation for this question was lower than average, meaning there was more consistency among evaluators on this question). At Step 16, respondents were asked, in essence, to create an entirely new Toolkit project: they were directed to an image gallery, to texts, and to a TOI list. They were told to load the TOI list (Step 16d), and to load, initialize, chunk, and section the text (Step 16e). Prior to Step 16, they had been shown examples of TOIs and texts in a sample project, but had never carried out the procedures in Step 16d. The question following step 16 (16.20) asked evaluators whether they were able to initiate a new project on their own. It thus addressed the most time consuming and the most difficult step in the entire evaluation. As the most difficult question, it should naturally have received the lowest score. We would not have been surprised had the score been between 4 and 5; in sum, it is a very positive sign that respondents found this step close to manageable (3 on the Scaled scale) after so little exposure to the Toolkit.

Having seen that the least positive average score was for the question following the most complex and timeconsuming step, it should be no surprise that the questions receiving the most positive scores followed relatively easy steps, with one exception. There were four questions receiving average responses at the most positive end of the scale, or above 1.5: one at 1.4 (Q/A 20.33); two at 1.2 (Q/A 11.12 and 11.13); and one at 1 (Q/A 6.6). Of these, the most informative for the evaluation was Q/A 20.33, for it addresses the core functionality of the Toolkit. This question came at the end of a series of steps in which participants reviewed texts to find terms describing three images from the *NCMA Handbook*. They were essentially asked to rate their satisfaction with the entire process of extracting metadata terms from texts. The fact that the average response ranks among the most positive, and ranks nearly the same as the question following what is arguably the easiest step (Step 6, to close the current project and reopen it), reflects significant overall satisfaction and usability.

As a final means of summarizing the scaled responses, we note that three questions were identical queries about overall user satisfaction with the Toolkit. We repeated this question in order to gauge whether continued exposure to the Toolkit had a negative impact on user satisfaction. Note that there is no change in user satisfaction between items 9.10 and 11.15; the average response is 2 ("Pretty good"). At step 20.32, the average response is only modestly less positive (2.13 on a 5-point scale). In sum, user satisfaction starts out relatively high and remains stable throughout the process of creating a new project, and the subtasks of selecting metadata for three images.

6.2. Metadata Selection Task

Three questions (18.25, 19.28, 20.31) required the ten evaluators to perform a metadata selection task. This task provided the "task success" portion of the evaluation. Participants were directed to an image gallery containing forty-two images from the North Carolina Museum of Art with instructions to view three images in particular—Jan Jansz den Uyl's *Banquet Piece*, Jan Brueghel the Elder's *Harbor Scene with St. Paul's Departure from Caesarea*, and a statue of *Neptune* attributed to Benvenuto Cellini.

Participants were then asked to use the Toolkit to locate relevant passages in the *North Carolina Museum of Art Handbook*; to choose text display options; finally to select descriptive terms for a catalog record. The passage lengths were 225, 254, and 304 words, respectively. Because all participants completed question 18.25, we use it to illustrate the richness of terms it elicited. Ninetysix selections were made, yielding a total of twenty seven term or variant terms sets, the top ranked of which are shown in Figure 1. Parentheses are used to indicate terms and term variants, thus "*((white) linen) tablecloth*" represents three variants.

7. Conclusion

CLiMB addresses a critical need in the context of a rapidly evolving set of practices. Cataloging standards

for images are evolving (e.g.,VRA Core (VRA, 2004) at the same time that new software for managing digital image collections is emerging. It is difficult to design software for applications that are under development, so the high ratings given to the ToolKit speak to the success of our approach. Users found the ToolKit very habitable, remained consistently satisfied throughout a lengthy evaluation (an all day meeting with two evaluation sessions), and selected a large number of terms per image.

Two intrinsic indicators point to the high quality of the resulting metadata: a large degree of overlap among the ten participants in terms selected; multiple intersections of terms with AAT terminology. Extrinsic validation awaits a study of the impact of CLiMB metadata on image search. A second phase of CLiMB at the University of Maryland will address this and other issues pertaining to creating a useful and useable cataloger's tool.

Acknowledgments

This project was supported by the Mellon Foundation. The authors wish to thank each other, the many people who participated in the two evaluations, and an especial thanks to our team member, Stephen Paul Davis, who played a key role in the development of the ToolKit.

References

Bacca, Murtha, Ed. (2002). Art Image Access. GettyResearch Institute.

Blitz, Roberta; Giral, Angela; Passonneau, Rebecca. (2004). Project CLiMB: Using Computational Linguistic Techniques to Harvest Image Descriptors. Visual Resources Association Bulletin 31(1):53-68.

Davis, P.; Elson, D.; Klavans, J. (2003). Methods for precise named entity matching in digital collections. *Third ACM/IEEE Joint Conference on DigitalLibraries (JCDL)*.

Day, D.; Aberdeen, J.; Hirschman, L.; Kozierok, R.; Robinson, P.; Vilain, M. (1997). Mixed-Initiative Development of Language Processing Systems. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*. Washington, D.C., 1997.

Fawcett, T. (1979). Subject indexing in the visual arts. *Art Libraries Journal* (Spring 1979):5-30.

Finch, S. and Mikheev, A. (1997). A Workbench for Finding Structure in Texts. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*. Washington, D.C., 1997.

Klavans, J. (In submission). CLiMB: Computational Linguistics for Metadata Building.

Klavans, J.; Blitz, R.; Davis, P.; Davis, S.; Elson, D.; Giral, A.; Heinrich, A.; Horvath, V.; Magier, D.; Passonneau, R.; Renfro, P.; Scott, R.; Weber, R.; Wolven, R. (2004). CLiMB. Progress Report, Second Year Cumulative. http://www.columbia.edu/cu/libraries/inside/projects/CLiMB/.

Mandel, C. A. (1985). Enriching the library catalog record for subject access. Library Resources and Technical Services (January/March 1985):5-15.

Taylor, A. (1995). On the subject of subjects. The Journal of Academic Librarianship (November 1995):484-91.

Visual Resources Association (VRA). (2004). Chapter 6: Subject," in *Cataloguing Cultural Objects: a Guide to Describing Cultural Works and Their Images*, May 2004 draft version http://www.vraweb.org/CCOweb/.

Walker, M. A.; Litman, D.; Kamm, C; Abella, A. (1997). PARADISE: A Framework for evaluation spoken dialogue agents." In Proceedings of the ACL.

Appendix 4: Brochures

CLiMB

Research Goals

The overarching goal of the CLiMB project is to improve access to digital images. To achieve this, we are looking at ways different user groups search for images and make relevance judgments. We also investigate how catalogers describe images to determine how their efforts correlate with user needs for subject access.

We are gathering data from CLiMB on what types of information catalogers find valuable for describing images. Cataloger feedback will enable us to develop a tool capable of identifying valuable index terms.

In addition to examining catalogers' term preferences, the CLiMB team will look at the overall cataloging workflow. Knowledge of contemporary practices will enable us to create a Toolkit that corresponds to cataloging processes and existing standards – one that integrates smoothly with cataloger practices.

People

Principal Investigator
Judith L. Klavans, University of Maryland

Senior Personnel
Eileen Abels, Drexel University
Angela Giral, Columbia University
Jimmy Lin, University of Maryland
Rebecca Passonneau, Columbia University
Dagobert Soergel, University of Maryland

Collections and User Studies
Carolyn Sheffield, University of Maryland
Laura Jenemann, Drexel University

Programmer
Tandeep Sidhu, University of Maryland

Editor
Lisbeth Herer, University of Maryland

Project Coordinator
Deborah Wallace, University of Maryland

Masters and Doctoral Students
Tae Yano, Columbia University
Joan Beaudoin, Drexel University
Rachael Bradley, University of Maryland

Collections Partners
Jeff Cohen, Bryn Mawr and University of Pennsylvania
Jack Sullivan, University of Maryland

College of Information Studies
University of Maryland
College Park
College Park, MD 20742
Phone: 301-405-7913
Fax: 301-314-9145
http://www.umiacs.umd.edu/~climb/

CLiMB

Computational Linguistics for Metadata Building

Computational Linguistics for Metadata Building

Enhancing image access through computational linguistic techniques

Computational Linguistics **CLiMB** **for Metadata Building**

About CLiMB

Under the direction of Principal Investigator Judith Klavans, the Computational Linguistics for Metadata Building (CLiMB) research project is currently based at the College of Information Studies at the University of Maryland. Dr. Klavans initiated this project in 2002 while at Columbia University. It continues to break new ground by enhancing subject-oriented access for image collections.

How it works

The CLiMB Toolkit functions on computational linguistic techniques applied over selected electronic texts. Based on CLiMB algorithms, the Toolkit extracts terms and phrases that refer to a specified image in that text. It then maps those terms to a standardized thesaurus (pilot-tested using the Getty's Art and Architecture Thesaurus (AAT)), thus providing a list of suggested terms that both complies with a controlled vocabulary and enables users to broaden or narrow their searches through the hierarchical structure of a standardized thesaurus. The integration of a relevant thesaurus also facilitates the disambiguation of homonyms and hyper-nyms. The development of new and sophisticated disambiguation strategies is one of the key research goals of the project. More information on our research goals is listed on the back of this brochure.

System Architecture

Summary of the CLiMB process

1. Texts, images, and metadata imported
2. Text analyzed
3. Text stored
4. Cataloger selects terms from stored text
5. Terms mapped to thesaural resources
6. Terms exported into catalog record

C Li M B

Computational Linguistics for Metadata Building

Example

CLiMB suggested terms are highlighted

John Hoppner
The Frankland Sisters, 1795
Andrew W. Mellon Collection
1937.1.111
Image and text from the National Gallery of Art

The Frankland Sisters, another **Hoppner group portrait**, represents **descendants** of **Oliver Cromwell**. As **a faithful spaniel** sleeps at their **feet**, **Marianne** hugs ***Amelia***, who holds **a crayon** and **portfolio of sketches**. During **the Royal Academy**'s **1795 exhibition**, **a newspaper critic** noted that **the picture** "does **the Artist great credit: the Group** is natural and graceful; **the heads** are sweetly painted."

CLiMB

Research Goals

The overarching goal of the CLiMB project is to improve access to digital images by investigating various methods of automated support for image professionals. Though an increasing number of art history and architecture image collections are available online, their corresponding subject entries remain largely inadequate. To inform the development of an effective image indexing tool, we explore how catalogers describe images and how computational linguistic technologies can enhance the process.

Testing was conducted both to determine the types of subject terms a cataloger may assign to a given image and to establish how well the CLiMB Toolkit would fit into existing cataloging practices.

Future goals include investigating the integration of the program with social networking technologies, and creating partnerships with museums and libraries whose staff will further test CLiMB and propose new applications for our research. If you are interested in participating in live testing of the CLiMB software at your institution, please contact Carolyn Sheffield at csheffie@umd.edu.

People

Principal Investigator
Judith L. Klavans, University of Maryland

Researchers
Eileen Abels, Drexel University
Angela Giral, Columbia University
Laura Jenemann, Drexel University
Jimmy Lin, University of Maryland
Rebecca Passonneau, Columbia University
Carolyn Sheffield, University of Maryland
Dagobert Soergel, University of Maryland

Programmer
Tandeep Sidhu, University of Maryland

Editor
Lori Hoglund, University of Maryland

Masters and Doctoral Students
Tae Yano, Columbia University
Joan Beaudoin, Drexel University
Rachael Bradley, University of Maryland
Xiaoli Huang, University of Maryland
Tom Lippincott, Columbia University

Collections Partners
Jeff Cohen, Bryn Mawr & U. Penn
Jack Sullivan, University of Maryland

University of Maryland
College Park
College Park, MD 20742
Phone: 301-405-7913
Fax: 301-314-9145
http://www.umiacs.umd.edu/~dlmb/

CLiMB

Computational Linguistics for Metadata Building

Computational Linguistics for Metadata Building

Enhancing image access through computational linguistic techniques

Computational Linguistics **CLiMB** **for Metadata Building**

About CLiMB

Under the direction of Principal Investigator Judith Klavans, the Computational Linguistics for Metadata Building (CLiMB) research project has been based at the University of Maryland and formerly Columbia University through support of the Scholarly Communications program at the Andrew W. Mellon Foundation. It continues to break new ground by enhancing subject-oriented access for image collections.

The CLiMB Toolkit uses computational linguistic techniques to process, categorize, and disambiguate text associated with an image. Categorizing text segments by their semantic function enables associations between text and relevant images. For each sentence, a part-of-speech labeler marks the function of individual words. Since each term may have multiple meanings or senses, CLiMB's lexical disambiguation algorithm facilitates sense mapping to get a more context-specific matching of sense to a term.

The Toolkit maps terms from text to the Getty Vocabularies to facilitate the differentiation of homonyms. The development and refinement of novel disambiguation technologies are central to CLiMB's goals. We explore applications of semantic categorization for enriching this process. CLiMB was designed to become a cataloger's workbench to enable concept-based access and to go beyond keyword search.

Language Analysis With Thesaural Resources

"Donatello's *David* was the first free-standing bronze of the Renaissance period."

bronze (metal)
(copper alloy, <copper and copper alloy>, ...

bronzes (sculptures)

(<sculpture by material>, sculpture (visual work), ... Visual and Verbal Communication) [300047333] **sculpture, bronze sculpture**

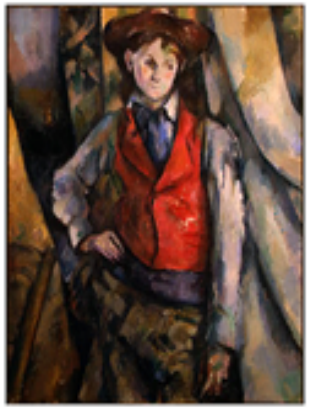
Hierarchical Position:
Objects Facet
Visual and Verbal Communication
Visual Works (Hierarchy Name)
Visual Works by medium or technique
Sculpture (visual work)
Sculpture by material
Bronzes (sculptures)

Summary of the CLiMB process

1. Texts, images, and metadata imported
2. Segments associated & categorized
3. Texts marked up with linguistic tags
4. Cataloger selects from candidate terms
5. Terms disambiguated
6. Terms mapped to thesaural resources
7. Terms exported into catalog record

Example

CLiMB suggested terms are highlighted

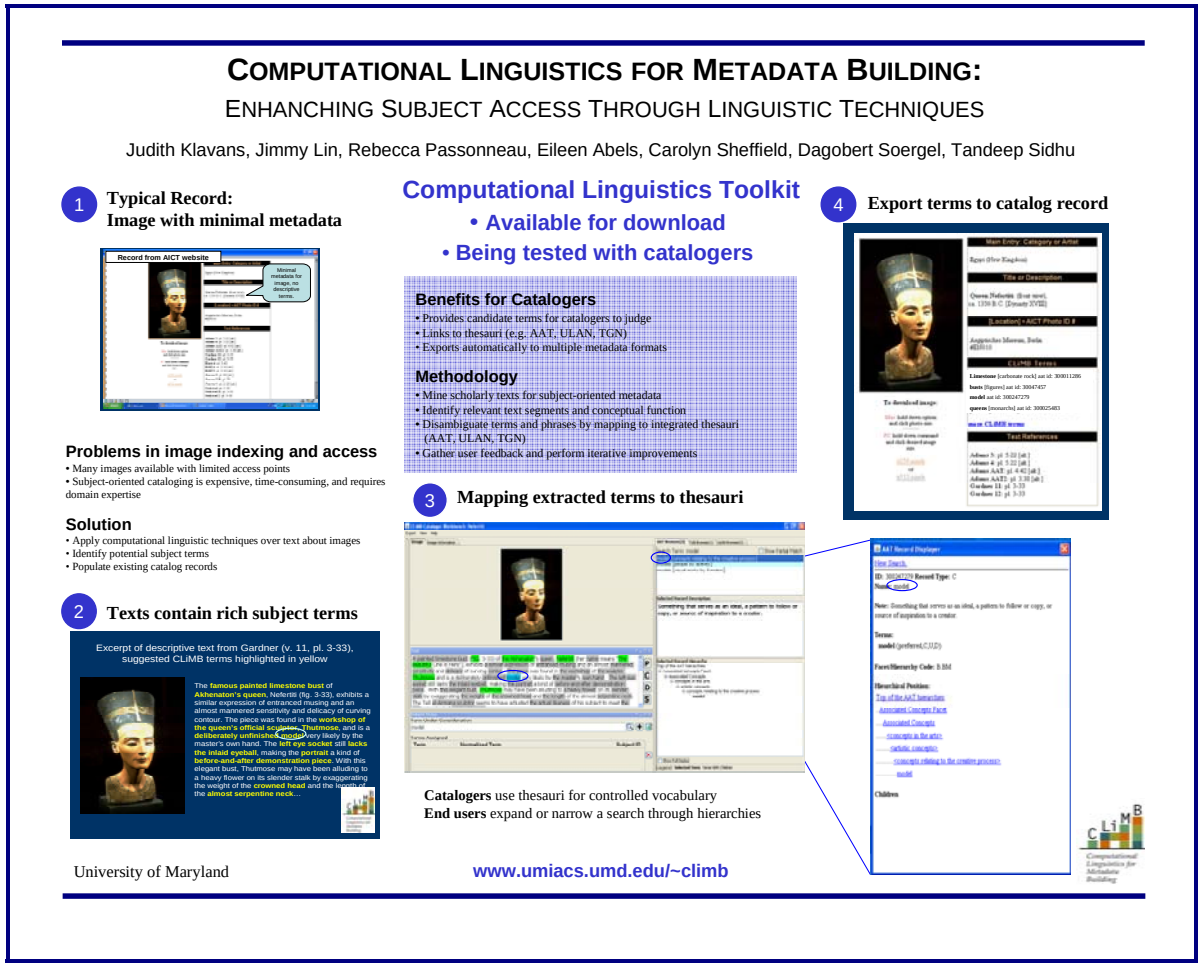


Paul Cézanne
Boy in a Red Waistcoat, 1888-1890
Collection of Mr. and Mrs. Paul Mellon
Image and text from the National Gallery of Art

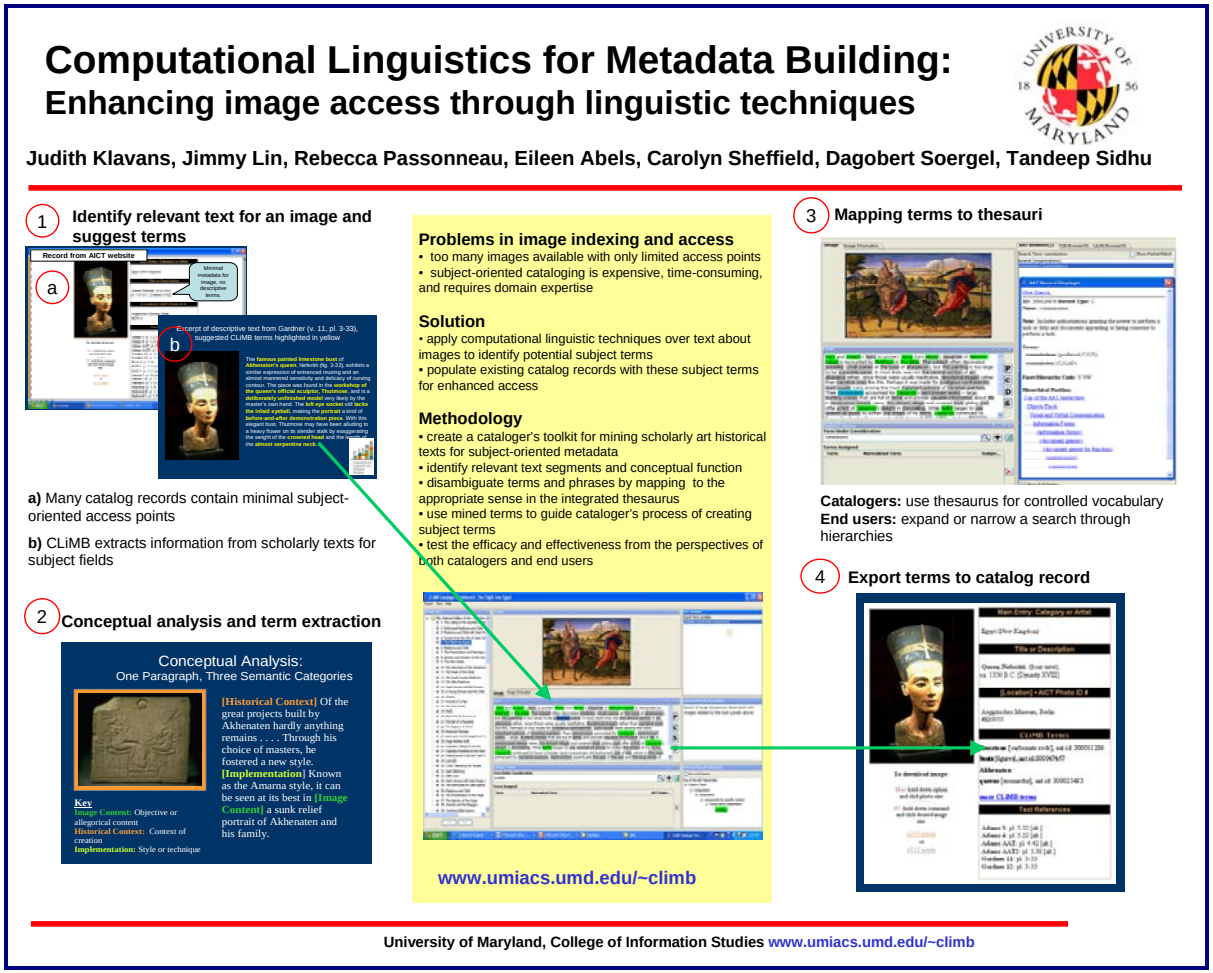
The boy's pose is that of **an academic life study**, and for some it has recalled **the languid elegance** of **sixteenth-century portraiture**. As **a young man** in **Paris**, **Cézanne** had learned not only from his **impressionist colleagues**, but also by studying **old masters** in the **Louvre**.

Appendix 5: CLiMB-2 Posters

Judith Klavans, Carolyn Sheffield, Jimmy Lin, Rebecca Passonneau, Eileen Abels, Dagobert Soergel, Tandeep Sidhu. **CLiMB Poster Session.** Visual Resource Association, March 2008. San Diego, California.



Judith Klavans, Jimmy Lin, Rebecca Passonneau, Eileen Abels, Carolyn Sheffield, Dagobert Soergel, Tandeep Sidhu. **CLiMB Poster Session.** *Human Computer Interaction Lab Symbosium***,** 2007, University of Maryland, College Park.



Appendix 6: Annotated Bibliography

Literature Review Summary

For image cataloging, the literature covers current practices, related challenges, and possible solutions. Cataloging manuals, such as Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images (Baca, et al. 2006), break down strategies for description into step by step procedures. In addition, the Panofsky and Shatford methods of analyzing the various levels of an image's content and meaning are discussed throughout the literature as a framework for thorough indexing. The staff time and expertise required to analyze these various levels of meaning, and to research their context, are two of the greatest challenges to sufficient image indexing discussed in the literature. Proposed solutions include automated indexing, most notably in terms of image recognition software. To our knowledge, literature regarding automation in terms of data mining and linguistic analysis, as CLiMB intends to apply them to image indexing, is minimal at best.

For image searching, the literature discusses access in terms of both expert and naïve users. In particular, user studies have uncovered which attributes are of interest to users and user preferences in refining searches. Other articles cover specific criteria applied in user relevance assessments, both visual and textual.

Cataloger Bibliography

General

Anderson, J. D., & Perez-Carballo, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management, 37*, 231-254.

In the first part of this two-part essay, Anderson and Perez-Carballo review the literature on human analysis for indexing. The authors begin with a brief overview on research comparing human and automatic indexing, focusing on variables such as extent, exhaustivity, etc. Explanations of the predominantly cognitive process performed by human indexers are then described, including the work of Chan, Soergel, Wilson, and Hjørland, among others. The authors also discuss arguments against cognitive processes that have been proposed, such as proponents of social construction, and arguments against the step-by-step indexing guidelines, which have been argued to leave out important details such as quality, "social context," and end users.

Relation to CLiMB: This article provides a good overview of the different ways the cognitive process of human indexing has been discussed by indexing experts, and compares human versus automatic indexing for variables in IR systems. This comparison lacks references to different studies, perhaps because "variables such as these have typically not been separately analyzed." The CLiMB Toolkit provides an opportunity to add to research on comparing human and automatic indexing.

Charbonneau, M. D. (2005). Production benchmarks for catalogers in academic libraries: Are we there yet? *Library Resources & Technical Services, 49*(1), 40-48.

Methods for appraising the performance of catalogers in academic libraries is a contentious issue. In this paper, the pros and cons of production benchmarks for catalogers are presented from the perspective of different library personnel. A key difficulty in appraising cataloging is because cataloging is more of an "intellectual enterprise" based on rules than a quantifiable good (p. 47).

Morris, D. E. (1992). Staff time and costs for cataloging. *Library Resources & Technical*

Services, *36*, 79-95.

This study reports findings of a staffing costs study conducted at the Technical Services Division at the Iowa State University Library. From 1987-1990, staff estimated time spent on technical services tasks for a one week period at various times over the three years. The study found that ordering comprised 25.72% of all technical services time, more than any other task. At 16.65%, cataloging comprised the next largest portion of technical services time. 35.5% of all cataloging time was spent on copy cataloging. However, per title, original cataloging averaged $34.13 per hour, while copy cataloging cost $8.18 per hour. The authors found that since the task of original cataloging was comprised of both minimal and full level cataloging, it could be made narrower, and more cost efficient, by eliminating minimal cataloging from this task. The study also found that the amount of cataloging time for authority work leapt from 13.9% to 20.6% during the three year period studied, which the authors attributed to inefficient automated authority control systems. The use of similar research methods across institutions in order to make cost studies in libraries more valuable to the library profession is recommended.

Relation to CLiMB: This article does not demonstrate the expense of image cataloging in particular. Literature specifically on the costs of special, art or visual material would be helpful, and perhaps some of the ARL studies have covered this topic as well. The amount of time and money spent on original cataloging and authority work is probably transferable to non-book materials. Also, the variables measured in this study, such as copy cataloging, searching, and authority work, are transferable to other studies of cataloging.

Morris, D. E. (2000). Cataloging staff costs revisited. *Library Resources & Technical Services, 44*, 70-83.

An update of Morris' 1992 report above finds that cataloging costs have decreased in the period between 1990 and 1997/1998. Acquisitions continuously comprised the largest amount of technical services hours, followed by cataloging. Between 1990/1991, average Cataloging Center hours dropped by 30%, while average titles cataloging increased by 41%. However, actual cataloging itself (copy, recataloging, and original) grew. In addition, cost of cataloging a title dropped 34%. Morris attributes increased cataloging productivity to the automation of tasks, which allows them to be accomplished more quickly. Collaborative cataloging initiatives are also highlighted as a time-saver. However, Morris does not note possible intervening variables between the relationship observed between decreased cataloging costs and greater automation.

Relation to CLiMB: Automating aspects of cataloging in order to save money is part of the research rationale for CLiMB's.

Image Indexing

Baxter, G. & Anderson, F. D. (1996). Image indexing and retrieval: some problems and proposed solutions. *Internet Research, 6*(4), 67-76.

The goal of this literature review is to present the problems in image indexing and examine their proposed solutions. Image retrieval is problematic because of the multidisciplinary nature of images, as well as images' inability to "tell us what they are about." One proposed solution to these problems has been thesaurus-based indexing systems, such as AAT, ICONCLASS, and TELCLASS. Another solution is visual thesauri systems. Picture description languages, are an alternative to manual indexing solutions, though like content-based image retrieval systems, these languages are of limited utility. The authors conclude that the proliferation of new, Internet accessible image retrieval systems, highlights the need for interoperable systems. Moreover, there is a need for multidisciplinary research, complicated by the decentralized nature of the Internet; the authors cite the InterNIC project as a hoped for exemplar for image indexing research.

Berinstein, P. (1999, March/April). Do you see what I see? Image indexing principles for the rest of us. *Online, 23*(2), 85-87.

Berinstein provides a concise overview of images' subjective and therefore, problematic nature; Panofsky's levels of meaning; and Shatford's (1986) approach to indexing images. Shatford's guidelines on *what* to describe are fluid and may be difficult to apply. Thus, Berinstein recommends considering the audience and letting the users browse. In addition, installing helper files is recommended, especially a thesaurus file, which has the benefit of automating the retrieval of synonyms. If thesauri are also hierarchical and browsable by end users, general and

specific indexing can be automated, and users can specify the level of indexing desired. Cataloger consistency too may become less of a concern.

Relation to CLiMB: The CLiMB Toolkit integrates Berinstein's recommendations for hierarchical, browsable thesauri.

Bierbaum, E. G. (1988). Records and access: Museum registration and library cataloging. *Cataloging & Classification Quarterly, 9*(1), 97-111.

While both libraries and museums collect and represent information-bearing objects, their differing philosophies have resulted in differing approaches to the management of their collections. Libraries focus on "carriers of language-based information," (p. 108), whereas museums are concerned with non-linguistic, "singular objects" (p. 109). By taking a coin silver teaspoon through the museum process of accession, registration, and cataloging, Bierbaum illustrates the differences between the museum and library approach to cataloging. One point highlighted during this process is the Importance of a museum's mission as the basis for museums' classification schema. As each museum has a different mission, museums have developed different classification schema to serve their needs. The uniqueness of each museum's mission also explains why externally created subject access systems are not implemented in museums. Moreover, since museums facilitate public access through exhibition and instruction, rather than the direct access favored by libraries, museum catalogs and library catalogs are intended for different audiences. Bierbaum sees the differences between libraries and museums as an opportunity for sharing expertise which will benefit professionals in both institutions.

Relation to CLiMB: This article serves as a reminder of the challenges in applying controlled vocabulary to unique objects, and also shows the historical differences between museums and libraries. Controlled vocabulary is challenging in museum settings not only because their collections contain unique objects, but because these objects are highly contextualized by their larger organizational and historical surroundings. Museum catalogs have often been developed in a pre-networked environment where consistency and standardization is not of concern. Museums are not alone in these characteristics: they are shared by archives and special libraries, too.

Brown, P., Hidderley, R., Griffin, H., & Rollason, S. (1996). The democratic indexing of images. *The New Review of Hypermedia and Multimedia, 2,* 107-120.

Democratic indexing allows untrained users to assign terms to images. In this study, the practicality of democratically indexing the subject matter of images is explored. The variables measured include: the range of terms used in comparison to a framework of image categories (depth), the number of unique terms used (breadth), and the amount of overlap in use of terms between indexers (agreement). The study found that when individuals created their own terms, the range of terms increased and the level of agreement decreased. In addition to their study, the authors discuss a planned thesaurus which would provide a means for reconciling the individually contributed terms for the same image. This reconciliation process would result in a publicly viewable term if enough indexers selected the term. However, the authors note that the low levels of agreement found in their study would require a low threshold for agreement in order for any term to become public.

Relation to CLiMB: It should be noted that this study's participants did not like the 'levels of meaning' categories (Biographical, Structural Contents, etc.). Also, working with the 'levels of meaning' table lengthened the amount of time it took to complete the indexing task. Have the variables of agreement, breadth, and depth been part of the CLiMB2 end user surveys and studies?

Chen, H.-L., & Rasmussen, E. M. (1999). Intellectual access to images. *Library Trends, 48*(2), 291-302.

This articles serves as another rich review of the literature up to that point.

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science*, *45*(8), 572-576.

In this article, Fidel prioritizes indexing documents according to their potential use, rather than primarily through the representation of their contents or "aboutness." Fidel explains that the representation of a document's contents, called the document-oriented approach, proceeds in a two-step process of first analyzing a document's contents, then translating the analysis into indexing language (p. 573). Instead of this approach, Fidel recommends a checklist technique of indexing

where indexers still first analyze a documents' contents, but then ask whether or not users of a particular descriptor would seek the document (p. 574). This method could be facilitated through automatic indexing, though qualitative factors involved in judging a document's usefulness call for an understanding of human, information seeking. Thus, Fidel concludes by calling for further study of information behavior, in order to create both automatic and human user-centered indexing.

Relation to CLiMB: Since the publication of Fidel's article in 1994, users can now index documents directly, making the indexing process even more user-centered than Fidel's proposed method. The CLiMB toolkit works within the user-centered framework by allowing catalogers, rather than the automatic indexing technique, to make the final determination of which subject indexing terms should be applied.

Graham, M., & Mahurter, S. (2003, August). Getting the picture: The cataloguing and indexing of images: Report of a one day seminar held at CILIP, Friday 6 June. *Catalogue & Index*, (149), 6-7.

Report of a one day seminar on the indexing of images, organized by Margaret Graham from the Institute for Image Data Research (IIDR), Northumbria University. The seminar was divided into two parts: a discussion of four image indexing case studies, and an overview of image indexing and classification tools. Graham concluded the seminar by presenting preliminary results of a survey of "ARLIS and BAPLA members on the cataloging and indexing of images" (p. 6).

Relation to CLiMB: The report states, "Chris Porter, Global Director for Search Data, provided a fascinating insight into the development of the in-house thesaurus used to index images added to the extensive Getty databases" (p. 6). This work sounds as if it may be similar to CLiMB's, and therefore may be worth following up on. Are seminars one of the planned deliverables? If so, Graham's seminar provides an example of what might be included in one.

Harris, K. (1986, Winter). Indexing a special visual image collection. *Catalogue & Index*, *83*, 6-8.

This article describes the implementation of Dickens House Classification for a system created to access images at the Dickens House Museum. The structure of the index entries was guided by the concept that images have either esoteric or manifest meaning. Representing the syntagmatic relationships - relationships between subjects comprised of compound terms – is important for subject access to visual materials because they are not always 'manifest.' This perspective resulted in entries structured as relationships between manifest and esoteric elements. While the article explains citation order of the index entries, it is unclear how the faceted schema is accomplished, as the article lacks multiple examples for the same headings. Still, the conclusion that this system both contextualizes each entry and reduces scatter of related concepts (p. 8) reveals the benefits of faceted classification schema for image retrieval.

Relation to CLiMB: This article highlights how faceted schema can provide more information than non-faceted schema, which is particularly useful when trying to describe images. However, the faceted aspects of the schema must be made visible to the user in some way by showing how two terms are related. How CLiMB displays the faceted structure of AAT to users may affect the Toolkit's usage.

Ho, J. (2004). Cataloging practices and access methods for videos at ARL and public libraries in the United States. *Library Resources & Technical Services, 48,* 107-121.

This study reports findings from a questionnaire on video cataloging practices at 44 U.S. public and 41 ARL academic libraries. Examined in light of *ACRL's Guidelines for Media Resources in Academic Libraries,* topics studies were: libraries' use of online catalogs for providing access to videos; the degree to which catalogers view screen credits for obtaining bibliographic information for videos; and the level of fullness at which videos are cataloged.

The study found that all respondents used their online catalogs to provide access to their entire video collections, supporting the *Guidelines'* recommendation that media resources be made accessible through the same retrieval mechanisms as other library materials. The study also found that ARL libraries were more likely to view title credits for obtaining bibliographic material. ARL libraries were also more likely to provide full level cataloging (again, supported by the *Guidelines'* recommendations). This finding may be because ARL libraries have both more access to viewing

equipment (recommended by the *Guidelines*), and more full-time original cataloging staff than public libraries. Finally, in their investigation of the application of LCRI statement 21.29D for creating added entries for significant video personnel, the authors found that these rules were applied inconsistently, due in part to the ambiguity in defining what "significant" is. Based on their study, the authors recommend a reexamination of LCRI 21.29 to provide more guidance on when added entries for persons involved in a video work's creation should be applied.

It should be noted that this study does not mention the use of the Second Edition of *Archival Moving Image Materials: A Cataloging Manual (AMIM)* published in 2000. *AMIM* works with AACR2 rules and addresses some of the same issues found in Ho's study. Perhaps LCRI 21.29 may have been found to be clarified in libraries where *AMIM* has been adopted.

Relation to CLiMB: Ho's study draws attention to the need for a meta-standard to make the bibliographic control of moving images mesh with libraries' other materials. *LCRI* but especially *AMIM* essentially functions as such, standardizing the moving image format's mixed bag of library, archival, and unique thesauri and description practices for use in an AACR2 setting. In a way, the CLiMB toolkit too is intended to function in this meta-standard function – facilitating cataloging based on the uniqueness of images, rather than the uniqueness of each catalog.

Zorich, D. M. (2007). *Webwise 07: Stewardship in the digital age: managing museum and library collections for preservation and use.* Washington, DC. Institute of Museum and Library Services.

Digital preservation was the theme of this WebWise conference on Libraries and Museums in the Digital World. These proceedings summarize the key points made be presenters of the conference. *Workshop #2: Sharing images and data: Making access to collections easier and better* (p.15- discussed the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), and focused on harvesting and sharing metadata, a topic which is outside of CLiMB's scope. However, some of the comments by K.Hamma and M. Baca are relevant to CLiMB. Hamma noted that "while projects such as AMICO (www.amico.org) and ARTstor (www.artstor.org) demonstrated the usefulness of having aggregated databases of collections images and information, contributing to these databases was an extremely onerous process, and the quality of metadata and images being collected was inconsistent" (p. 15). Aggregated image databases did not address the need for easy-to-use, consistent metadata. This is a gap which CLiMB's user studies hope to address by creating an end result that is, at least, highly usable. CLiMB's mining and thesaural matching directly addresses the needs expressed in Baca's presentation. In the presentation *Shareable Metadata for Nonbibliographic Materials: Implications for Libraries and Archives,* Baca states that "Service providers and aggregators could help by using controlled vocabularies and thesauri as search assistants. They also could add value to their offerings by providing services such as vocabulary mapping, query expansion, vocabulary-assisted searching, user-added metadata, metadata enhancement, etc" (p. 19).

Jörgensen, C. (1998). Attributes of images in describing tasks. *Information Processing & Management, 34*(2/3), 161-174.

This exploratory study found that when people described images, the terms used to describe images were grouped into classes of terms relating to objects, people, color, and location. 107 M.L.S. students participated in three tasks: describing images, describing how they would request the images for known item searching, and describing the images five weeks after seeing them. Using the constant comparative technique, the most commonly occurring attributes of images were categorized as body part, location-specific, and activity. These attributes differ from more formalistic descriptors pertaining to composition, artistic technique. Attributes fell into broader classes which were similar for each of the tasks. The larger classes of objects, people, color, and location occurred most frequently. In comparing the terms resulting from the describing tasks to the terms used in LCTGM and Thesaurus Iconographiques (TI), Jorgensen found that the coverage of classes in the unconstrained tasks was of a wider ranger than those found in LTGM and TI. The results of this study point to a need to analyze and extend image indexing systems.

Relation to CLiMB: Attributes in the Objects class were the most frequently occurring terms. This is encouraging for CLiMB, as Objects are one facet of the AAT. Various labeling studies offer an opportunity to extend Jorgensen's research.

Jörgensen, C. (1999). Access to pictorial material: A review of current research and future

prospects. *Computers and the Humanities, 33,* 293-318.

This review of literature presents a current research applicable to image retrieval systems development. A summary of image-oriented classification and indexing systems, major research studies in image retrieval issues, and a research agenda for the future is presented. The advantages and disadvantages of various classification systems, including *AAT*, *LCTGM*, and content-based retrieval are first discussed. Jörgensen then describes research studies conducted. These studies have focused on the necessary attributes for image retrieval, queries, user studies, tasks, and searches. Included in this discussion is Jörgensen's proposal for a framework of image categories consisting of data images (raw data), informative images (visual material for communication organized by humans), and expressional images (subject to multiple interpretations). The multidisciplinary nature of images indicates that future research will likewise need to be multidisciplinary, pursuing solutions from a variety of areas such as information retrieval, computer science, and cognitive science. In her proposed research agenda, Jörgensen recommends analyzing current classification systems for their most useful aspects, then combining relevant thesauri and professional expertise in order to develop a system for a particular user group. In addition to this research in current tools, research in new representational techniques, image typologies, and defining visual similarities are recommended for future research. Jörgensen points out that one rarely pursued research question is into "the function of the *process* of searching for an image", rather than the end product of the search itself (p.315). Jörgensen ends with the recommendation that future image retrieval research should pursue defining, not just solving, the problems of access to visual materials.

Jörgensen, C. (2003). Organizing and providing access to images. In *Image Retrieval: Theory and Research* (pp. 69-138). Lanham, MD: Scarecrow Press.

This chapter of Jörgensen's book focuses on "intellectual control of and access to" images (p. 70). Specifically, an overview of systems of description, a discussion of image classification theory, and some of the empirical research conducted on the use of images is provided. First, the tools comprising image description systems are summarized. Data value tools such as subject headings and thesauri are discussed; in particular, Jörgensen's previous studies on the *Library of Congress Thesaurus of Graphical Materials: Topical Terms for Subject Access (TGM I)* and the *Art and Architecture Thesaurus (AAT)*. The application of these data value tools is guided by data content tools, such as *AACR2*. Major metadata initiatives, which Jörgenson calls data structure tools and include MARC and the VRA Core, work to organize the descriptors of data value tools. Following the overview of the three description tools is a discussion of the theoretical basis for these tools. Here Jörgensen discusses Erwin Panofsky's (1955) three levels of meaning, and Sara Shatford's extension (1986) of Panofsky. The third section of the review, "Image Users, Queries, and Searches", provides an overview of empirical studies focusing on the "users of image collections, their queries, the types of searches they conduct, and their searching behavior" (p. 120). Following a brief overview of current indexing practices, Jörgensen summarizes the variety of considerations involved in implementing systems for access to images: collection-based, user-based, vocabulary based, etc. The review of research on users' searches suggests that searches can be classified as unique (specific item) searches, prototypical (representative instances of persons, items, etc.), or abstract (items conveying feelings or concepts). Attributes in searches can range from being highly specified (children eating) to highly unspecified (sad) (p. 134) The chapter ends with the recommendation for further empirical study of both the information behavior of users and of image attributes in order to facilitate development of future image retrieval systems.

Relation to CLiMB: The benefits and detriments of the various indexing and cataloging tools used to describe images are helpful to keep in mind during the development of the toolkit, as are the variety considerations involved in implementing image retrieval systems.

Layne, S. S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science, 45*, 583-588.

The goals of image indexing guide this discussion of what should be indexed in an image. The goals of image indexing are to provide access to images based on their attributes and access based on how the images might be usefully grouped. Attributes for access to images are: biographical, subject, exemplified, and relationship. Access to useful groupings of images helps users make comparisons between images, discover specific images from a group of images, and provides for quick scanning. Groupings can be decided by asking a series of questions, such as "What level of detail is needed as the basis for groupings?" However, different users will require different attributes. Layne concludes by recommending studies that test the usefulness of attributes.

Relation to CLiMB: CLiMB has furthered Layne's research by testing inter-labeler consistency of text associated with images.

Matusiak, K. K. (2006). Towards user-centered indexing in digital image collections. *OCLC Systems & Services, 22*(4), 283-298.

Matusiak reviews the image indexing literature relevant to user-created indexing (such as social classification) and applies an overview of "social classification in relation to images." The literature reflects two approaches to indexing images: a concept-based (human indexer) and a content-based (automatic) approach. Also well noted are the inherent difficulties of transcribing visual images into textual descriptions. These difficulties lead to various problems for users. Yet, as multiple authors have noted, "Few studies evaluate the effectiveness of image indexing from the user perspective" and "few studies mention user participation in the indexing process or engage users in describing images as part of an evaluation of indexing systems" (p. 286).

To illustrate the pro's and con's of social tagging and traditional indexing, Matusiak compares two sets of images. Each set contains one image from Flickr, and one from the "Cities Around the World" digital image collection at the University of Wisconsin-Milwaukee Libraries. Flickr uses a less detailed description system than the that of the University, which employs multiple controlled vocabularies. Searching further on Flickr reveals a wider variety in content, number, and language of tags and number of tags than in the University catalog. Matusiak concludes that social tagging is seemingly subjective, "emerges organically," and "not an artificial construct" while traditional indexing is more consistent, provides description in a "hierarchical, structured manner," yet static. Social classification should therefore be construed as an enhancement, not a replacement, for traditional indexing. The ability for users to enhance pre-existing metadata, a system that allows feedback to indexer-assigned terms as well as eventual controlled vocabulary development (citing Merholz, 2004b), multilingual abilities, and user-generated tags' currency could all help facilitate "collaborative knowledge construction" which would eventually lead to more user-oriented indexing.

Relation to CLiMB: This article provides a concise, clearly written overview of the difficulties of image indexing. While user-generated social tags are beyond the scope of CLiMB, even just allowing for user generated feedback is an area which CLiMB could certainly explore.

Panofsky, E. (1962). *Studies in iconology: Humanistic themes in the art of the Renaissance.* New York: Harper & Rowe.

Rasmussen, E. M. (1997). Indexing images. *Annual Review of Information Science and Technology, 32,* 169-196.

This 1997 article provides a broad overview of image indexing and the associated challenges and developments. Rasmussen provides an extensive literature review throughout, pointing to relevant sources for user needs, image attributes, and indexing techniques. For user needs, he refers to several articles that explore what aspects of images different user groups are interested in. In particular, he cites a study conducted at the National Film Board of Canada (NFBC) that found the "of" aspects (i.e., the identifiable objects, people, events portrayed in the films) to be the most important for responding to their users' needs, but recognized the significance of the "about" aspects for searching art related collections. For example, Rasmussen also cited the information needs of art historians as outlined in Tibbo (1994), namely: creator, meaning (about), imagery, social and cultural contexts (about). A greater discussion of the "of" and "about" aspects is available in the section that covers the Panofsky and Shatford models of image attributes. For indexing techniques, he references Constantopoulos and Doerr's techniques, including the use of captions, keywords from controlled vocabulary, iconology, and hypermedia annotation (link to text).

Relation to CLiMB: The image indexing theories and techniques outlined in this article may suggest different functionalities of the toolkit, including ones that may not be readily apparent from our studies alone. In addition, this article addresses the need for more user studies, which is important as CLiMB2 will contribute to the field by conducting both cataloger and user studies.

Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, *6*(3), 39-62.

The process of analyzing the subject of a picture is the focus of this seminal article. Panofsky's (1962) three levels of meaning provide a framework for subject analysis through the analysis of the generic and specific events, objects, and names that a picture is *Of*, and the more abstract symbols and moods that a picture is *About*. Analysis of the "who," "what," "when," and "where" of a picture provides a means for classifying their subjects. The collection to which a picture belongs, as well its intended use and other circumstances, ultimately determine how a picture will be examined and described (p. 59).

Relation to CLiMB: By building a utility for cataloging the subject of images, CLiMB supports the use of Shatford's framework, as well as a shared interest in the relationship of meaning in pictures to meaning in language (p. 46).

Taylor, A. G. Subject analysis. (2004). In *The Organization of Information* (2nd ed., pp. 241-259). Westport, CT: Libraries Unlimited.

The purpose, challenges, and processes involved in subject analysis are described in this concise overview. For Taylor, subject analysis is the process of "aboutness." The first step in the process is conceptual analysis, where what an item "is" and what the item's content is "about" is determined (p. 242-3). The "aboutness" is then translated into the language of an indexing or classification schema. Regarding interindexing consistency, Taylor notes that the failure to determine the same "aboutness" of documents is not the same as the failure of controlled vocabulary. While Taylor touches briefly on nontextual subject analysis, she highlights the importance of consulting the object itself for nontextual and art information.

Relation to CLiMB: Taylor extends the "of-ness" and "about-ness" issues found in the image indexing articles (most notably those of Shatford) into a succinct explanation of subject analysis in general. This seems to be the framework of subject cataloging that CLiMB is situated in. Additionally, Taylor makes two points which are particularly relevant to CLiMB. First, CLiMB works within both ends of the continuum of exhaustivity by using AAT's depth indexing to retrieve words so that catalogers can more easily conceptually summarize documents for retrieval. CLiMB shares Taylor's principle that determining what a subject is about must ultimately involve human mediation, because "Computers can determine what words are used in a document but cannot determine meaning" (p. 241).

Svenonius, E. (1994). Access to nonbook materials: The limits of subject indexing for visual and aural languages. *Journal of the American Society for Information Science*, *45*, 600-606.

This article, through its discussion of the challenges of subject indexing for images, provides a sort of "statement of the problem" relevant to the CLiMB project. Svenonius discusses the inherent flaw of trying to describe the visual with the textual, stressing that when "the medium is the message" something is inevitably lost. She points out that the syntax of music and art are governed by different laws than language and translation between formats is not always possible. Furthermore, she discounts the supposition that realistically depicted subject matter allows for ease of indexing. Iconography, for example, involves real world objects but the meaning is not literal. In addition, Svenonius stresses that a given object or symbol may have multiple meanings (ex: water in Christianity). Nonrepresentational art presents additional problems in terms of description, as there is frequently no literal description to accommodate the visual content. Svenonius concludes by encouraging research into non-subject attributes that aid users in accessing images and other nonbook materials.

Relation to CLiMB – Subject indexing is challenging, and comprehensive indexing may be no more possible than a comprehensive monograph for the content and context of every artwork, and for every artist's intentions. However, subject based queries remain among the most frequent types of queries submitted to image databases and these challenges should not be seen as a reason to stop trying to improve subject access. Rather, awareness of these challenges should inform the development of new cataloging strategies tools to resolve the currently insufficient or non-existent subject access for many visual collections.

Winget, M. (2004, March). Intellectual access to digital art objects: Image attributes and art historical knowledge representation. Paper presented at the Visual Resources Association Conference, Portland, OR. Please note that this document is currently incomplete. Retrieved September 3, 2007 http://www.ischool.utexas.edu/~megan/research/Winget_Img-Rep.pdf

While Panofsky's theories of iconology have been formative to the process of image cataloging, alternative theories are here woven into the Shatford-Layne (1994) approach. These theories are summarized, most prominently those of David Summers. Notably, Winget argues that instead of "beating themselves up about 'about-ness'" (p.9), user studies point to a need for image catalogers to focus on providing access through increased description of images' salient details (p. 10).

Classification Systems

History of the AAT. (1994). In Petersen, T. (Director), *Art & architecture thesaurus* (2nd ed., Vol. 1, pp. 3-26). New York: Oxford University Press.

This is a history of the *Art and Architecture Thesaurus's* (AAT) development up to the publication of its second edition in 1994. The roots of the AAT can be traced to 1979, growing out of architectural historian Dora Crouch's desire for a thesaurus for art and architecture. This history shows how computer technology and information science, along with an awareness of humanities scholars' information needs, has been part of the AAT project throughout its lifespan. The article concludes by noting that "the AAT is a communication tool", representing both a perspective of art and architecture that is both "structured" and flexible enough to allow for growth (p. 26).

Relation to CLiMB: This history shows how the AAT evolved into to its final form. The article mentions that classification specialists, including Jean Aitchison, contributed the idea of organizing the AAT in a faceted structure.

OCLC Service adds Getty Vocabularies. (2006 December). *Advanced Technology Library, (35)* 2, 9.

This is an announcement that Getty Vocabularies (AAT, TGN, and ULAN) have been added to the OCLC Terminologies Services. OCLC Terminologies Services include other thesauri, such as LCTGM I & II.

Relation to CLiMB: According to http://www.oclc.org/research/projects/_termservices/resources/tspilot-services.htm (consulted May 16, 2007), "the service would provide one-stop access to terminology resources and would be accessible through any web-based metadata editor". It might be worthwhile to discover how accurately the service searches the terminologies. Analyzing the source's interface might be useful as well, because like CLiMB, OCLC Terminologies Services retrieves text from multiple sources, and collocates the results in one interface.

Baca, M. (2004). Fear of authority? Authority control and thesaurus building for art and material culture information. *Cataloging & Classification Quarterly, 38*(3/4), 143-151.

This essay provides an overview of controlled vocabulary resources available for museums and offers recommendations for customizing them. Using herself as an example, Baca describes how the concept of authority control is foreign to the museum world: both because art historians consider themselves to be the authorities and because the museum world lacks appreciation for the benefits of controlled vocabularies. To show how useful controlled vocabulary can be, Baca describes the AMICO project's lack of authority file. When trying to find information in an online environment – artist name variants (e.g., Gogh, Vincent van or Gogh, Vincent Willem Van) especially problematize information retrieval. After providing a brief explanation of how the AAT and ICONCLASS function as retrieval tools, Baca explains the ways these controlled vocabularies can be enhanced. Using an authority file to run searches against can aid retrieval. Additionally, enhancing controlled vocabulary terms with locally used terms and additional variants could aid end-users.

Relation to CLiMB: A few of Baca's points are particularly relevant to CLiMB. First, Baca states that the philosophy of the Getty Vocabulary Program is that no one name in an authority file cluster are "better" than the other. Thus, the Getty Vocabularies are open to the variability of natural language inherent in search language and in the creative language of art history texts. Second, while authority files can be used to match user terms to document terms, it is ultimately up to humans to determine whether or not the match is correct. CLiMB follows this premise by incorporating human catalogers as the ultimate judges of correct usage.

Baca, M. (2003). Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. *Cataloging and Classification Quarterly, 36*(3/4), 47-55.

Baca outlines basic, practical considerations for the development and maintenance of information retrieval systems for images and other cultural heritage information. In particular, she addresses the need for choosing metadata standards, controlled vocabularies, and metadata crosswalks appropriate to the collections being cataloged. She describes some of the metadata standards available and the types of collections they match best. From these descriptions, it would appear that the VRA Core (Visual Resources Association Core) is the most relevant for CLiMB as it was designed for describing surrogates (i.e. digital images) of art objects. The CDWA would also be of interest. Metadata crosswalks, when properly chosen and implemented, will enable searching across different information retrieval systems. Baca also stresses the importance of choosing classification systems capable of supporting cross references and including terminology pertinent to the items being indexed. Again, she provides a brief summary of systems available with a caveat that the best strategy may be to choose one or more and build on to it/them to meet the needs of the collection in question.

Relation to CLiMB – This article raises several questions concerning the conceptual foundation of the CLiMB toolkit. Is it compatible to all the standards and schemas that Baca addresses? In particular, can the categories fed to the segmenter be mapped to CDWA, VRA, or another metadata standard? If not, can it be easily resolved through existing crosswalks? Will CLiMB terms be modified to meet the rules of controlled vocabularies?

Soergel, D. (1995). The Art and Architecture Thesaurus (AAT): A critical appraisal. *Visual Resources*, *10*(4), 369-400.

In this critical appraisal, Soergel discusses the AAT in terms of usability, intuitiveness, and structure. For usability, Soergel looks at the formatting of the hierarchical and alphabetical displays and offers suggestions to improve readability. These suggestions will be useful in designing the interface for CLiMB2. For intuitiveness, Soergel looks at how easily catalogers and searchers can locate the appropriate term, for the appropriate context, through the hierarchical and alphabetical displays. He points out that, although the Introductory Material claims all descriptors are elemental concepts, many are actually precombined descriptors. This precombination has led to related terms being inconsistently dispersed throughout various facets. According to Soergel, the mono-hierarchical structure further hinders the meaningful and consistent arrangement of descriptors, since each descriptor can fit into one and only one hierarchy. The lack of cross references (except RT in alphabetical display, and then only inconsistently) limits descriptors to narrow contexts which are often not entirely appropriate or accurate for the information needs of catalogers and end users.

Relation to CLiMB: Soergel's article critically evaluates the structure and conceptual organization of the AAT, a resource that will be integrated into the CLiMB toolkit. Understanding the overall structure of this thesaurus can help facilitate the mapping of CLiMB terms to the appropriate facets and hierarchies. Familiarity with the AAT's inherent strengths and weaknesses can also inform the organization of the toolkit. A short discussion of what the AAT does not include is also provided (p3) and may be used to inform the types of terms the toolkit will seek.

Zeng, M. L. (1999). Metadata elements for object description and representation: A case report from a digitized historical fashion collection project. *Journal of the American Society for Information Science*, *50*(13), 1193-1208.

The author presents a project at Kent State University Museum to select an appropriate metadata schema selection for a digitized costume collection. Catalog records were created for eighty American costumes and some additional items in the museum's collections. The records were created in USMARC, VRA Core, and Dublin Core, and then analyzed by the author and museum curators. VRA Core, with modifications, was deemed the best suited for the collection. OhioLINK's recommendation for VRA, also contributed to Kent State's decision.

In the process of selecting the best metadata schema for the costume collection, the author developed three categories of descriptive metadata needed for the collection: registration information, descriptive information, and subject/topic information. These categories capture both the large amount of biographical information that a museum object's accession record contains, as well as details regarding fabric, pattern, etc. The author notes the importance of "ofness" over "aboutness" in subject indexing for costumes (p.1202).

Relation to CLiMB: This article may not directly support or further the work of CLiMB. It does, though, show an additional area where the toolkit would prove useful: in mining accession records for three dimensional materials and digital images of them.

Inter-indexer and Indexer-User Comparisons

Jörgensen, C. L. (1999). Image indexing: An analysis of selected classification systems in relation to image attributes named by naïve users. *Annual Review of OCLC Research*. Retrieved, July 19, 2007, from http://digitalarchive.oclc.org/da/ ViewObjectMain.jsp?fileid=0000002655:000000059275&reqid=190

Markey, K. (1984 April-June). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research*, *6*(2), 155-177.

Saracevic, T. (1991). Individual differences in organizing, searching and retrieving information. *Proceedings of the American Society for Information Science, 28,* 82-86.

Trant, J. (with the participants in the steve.museum project). (2006). Exploring the potential for social tagging and folksonomy in art museums: Proof of concept. *New Review of Hypermedia and Multimedia, 12,* 83-105. Data from study available at http://steve.museum/index.php?option=com_weblinks&catid=33&Itemid=58.

The steve.museum project conducted proof of concept tests to validate the project's proposition that social tagging enhances existing museum documentation of works of art (p.87). The goal of the tests was to determine if non-specialists contributed useful additional access points when describing what an image depicts. The first two tests conducted compared library professionals and five library assistants/associates assignment of terms to five works of art. Library assistants were found to provide both more terms and more distinct terms than library professionals.

The third and fourth tests were conducted on a computer using an automated desktop term collection tool. (Test 3 results are not available.) Thirty-nine participants comprised of museum administration staff and volunteers entered keywords for thirty single images. These images were presented three different ways: with identifying captions, without captions, and without captions but with category prompts.

The tests found that a high number of new, useful, and accurate terms were supplied by non-professional catalogers. The fourth test resulted in 6679 terms. Of the 3780 distinct terms collected, 3343 terms (88%) were not present in the existing art documentation. 77% of these terms were validated by the Metropolitan Museum of Art Subject Cataloging Committee as appropriately describing their corresponding works of art. Validation methods, however, were inconsistent. Valid terms generally outnumbered invalid terms, excepting an Islamic manuscript and an African sculpture. The Islamic manuscript was also skipped by 17% of the participants.

The study also found that the presence of a caption did not appear to determine either the number of terms applied, inter-indexer consistency of the five and three most common terms, or the validity of terms - the percentage of these terms applied, the consistency, and the validity of new terms were roughly equal between images presented with and without captions. However, the image with category presentation yielded only 22% of total terms assigned.

Inter-indexer consistency was significant for this test. Between the thirty participants, the five most common terms ranged in co-occurrence from 22.3% to 47.4%; the average consistency of the three most common terms was 41.8% for images with captions. The authors conclude that the study shows that non-specialists can supply new and useful access points which serve to enhance art museums' professional descriptions. Moreover, social tagging of museums' art works is both engaging, productive, and reflective of communities that museums seek to serve.

Relation to CLiMB: Based on the similarity between Trant and CLiMB's user study methodologies, it would be interesting to compare CLiMB results to Trant's.

Turner, J.M. (1995). Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images: Research results, *Proceedings of the American Society for*

Information Science, #, 9-12.

End User Bibliography

General

Choi, Y., & Rasmussen, E. M. (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, 54*, 498-511.

"The characteristics of users' visual information needs and queries, and categories of indexed terms to represent images in American history" are examined in this study (p. 499). Using the Library of Congress' American Memory collection, eighteen faculty and twenty graduate students were asked to describe their visual information needs in natural language and to search the collection using keywords. The authors found that most of the searches fell into the category of general/nameable need (Batley, 1988). Known author and title searches were rare, and subject and format terms were popular. Another finding of this study was that the average number of search terms used for image retrieval (4.87) was higher than that of textual retrieval (3.74 according to Goodrum & Spink, 2001). The authors state that subject oriented descriptors (representing about-ness) were important in users' judgment. 92% of respondents indicated that they were not satisfied with the textual information provided for the images. The discovery that "the most frequently occurring queries incorporated specific visual elements from the images themselves" supports previous image studies (Keister, 1994). Unlike Markey (1984), this study found that there are aspects of textual descriptions that are important to users.

Relation to CLiMB: The study emphasizes the need to enrich textual indicators of image content.

Collins, K. (1998). Providing subject access to images: A study of user queries. *The American Archivist, 61*(1), 36-55.

Cunningham, S. J., & Masoodian, M. (2006). Looking for a picture: an analysis of everyday image information searching. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 198-199.

This study is a response to the lack of research on everyday life image information needs. 31 students in an HCI course created 64 descriptions of image-related needs. The students each wrote a description of a recent information need and how they tried to satisfy it, then interviewed a friend to ask them the same information. The information needs were then categorized according to Batley's (1988) framework as adopted by Choi & Rasmussen (2003). Of the four categories of need - specific; general, nameable; general, abstract; and subjective - 70% (45) of the needs were specific. The 12.5% (8) subjective needs were "satisfied primarily by browsing." In addition, 20% (13) of information needs were satisfied by browsing, rather than searching. 50% (32) of the searches resulted in selecting a single image. Additionally, the authors observed that instead of comparing a found document to an information need as is done with text searching, in image searching, users compared the images found to each other.

Relation to CLiMB: This study demonstrates that image seeking and retrieval differs from text searching. Users fulfill their image needs through flexible browsing and comparison of images, rather than through text searching alone.

Goodrum, A., & Spink, A. (1999). Visual information seeking: A study of image queries on the World Wide Web. *Proceedings of the Annual Meeting of the American Society for Information Science, 36,* 665-674.

The authors analyzed approximately 33,000 image related queries taken from a set of over 1,000,000 Excite web search queries. At the time of the study, images were retrieved by query words that matched words found in filenames or the web pages themselves. The authors found that an average of 3.74 terms were used per query. Sexually explicit terms comprised one quarter of the one hundred most frequently occurring terms. However, these terms comprised only a small percentage of the total terms. More than half of the search terms were used only once. The meaning of the terms were ambiguous outside of the context of their queries, and some seemed to be used as

"refiners" (Enser, 1995). The authors acknowledge that the study's lack of relevance judgment made it difficult to conclude if multiple queries indicated changes to the initial query, or separate searches. While some comparison is made to other web query analysis work (Jansen et al., 1998), further comparison between image and text searching needs will be helpful to demonstrate how these queries differ.

Relation to CLiMB: Goodrum & Spink and the CLiMB Toolkit have a shared interest in the use of individual terms for image information retrieval. Goodrum & Spink stress the need for qualitative data, such as observations, in order to contextualize findings on the semantic level.

Goodrum, A. A. (2005). I can't tell you what I want, but I'll know it when I see it: Terminological disconnects in digital image reference. *Reference & User Services Quarterly*, *45*, 46-7 & 51-3.

This study sought to gain a better understanding of the digital reference librarian's role as mediator between image requests and their descriptions on the web. 590 image related requests from the Virtual Reference Desk (VRD) and AskEric archives were analyzed using "Categorization of Image Types" (Enser & Sandstrom, 2003) and visual refinement categories (Enser, 1995). Goodrum found that 77.3% of requests fell into the Documentary - General Purpose category (photos or paintings meant to faithfully represent reality). 55% of requests were categorized as containing Non-unique Refined visual refinements (a query for any object, such as a person, refined by time, action, etc.). When comparing terms in reference requests to terms in answers, and to image terms extracted from web-pages, this study found that fewer than 30% of websites matched three or more terms in users' requests. Moreover, the terms did not occur in file names, captions, or titles. Additionally, in using human-mediated, digital reference services, an average of 21 terms per request were used (rather than the 3.74 terms per Web search engine query found in Goodrum & Spink, 2001). This study confirms the existence of a semantic gap between the terms users employ to describe their image needs, and the terms that systems extract from text surrounding images. Overall, "the text does not describe the image" for documents on the web.

Relation to CLiMB: The gap between the words used in image information retrieval systems and those used in users' requests is one that the CLiMB Toolkit can fill.

Hastings, S. K. (1995). An exploratory study of intellectual access to digitized art images. In M. Williams (Ed.), *Proceedings of the 16th National Online Meeting* (pp. 177-185). Medford, NJ: Learned Information.

Hastings studied how "variations in the retrieval parameters and access points" effect art historians' queries of an art image database. Data collection involved observations of the use of a collection of sixty-six West Indian and Caribbean paintings that were converted into both digital images and mounted photographs and interviews with participants. The study found a relationship between categories of queries and the access points and computer manipulations used for the queries. (Hastings does not provide the queries used.) These categories were grouped into four different levels of complexity. Some of the queries in the least complex group, the Identification (who, where, when) class, could be answered without images. Some of the queries in the most complex group (including meaning, subject, and why queries) required secondary subject resources.

Relation to CLiMB: Hasting's recognition of a relationship between queries, access points, digital images, and computer manipulation serves as the context for the development of the CLiMB Toolkit. CLiMB addresses the meaning behind the subject queries that comprise Hastings' most complex level of query complexity.

Hastings, S. K. (1995). Query categories in a study of intellectual access to digitized art images. In T. Kinney (Ed.), ASIS '95: *Proceedings of the American Society for Information Science (ASIS) 58th Annual Meeting*, 3-8.

This article appears to be a longer version of the study discussed above. Here, Hastings described assertions relevant to each query category. Of particular interest: Level Two queries, such as queries about artists, were easier to answer with digital images than with the photographs. In order to answer these queries, some manipulation of images and text was used. Level Three queries such as "What is that object?" often involved comparison between images, or use of a magnification tool. Overall, the study found its categories of art historians' information needs support those previously

developed by Bakewell, Beeman and Ross (1988), Brilliant (1988), Markey (1986), Panofsky (1962), and Stam (1984).

Relation to CLiMB: The findings of this study could contribute to the development of a mental model of information seeking for digital images.

Holt, B. & Hartwick, L. (1994). "Quick, Who Painted Fish?" Searching a picture database with the QBIC Project at UC Davis. *Information Services & Use, 14,* 79-90.

The University of California at Davis's Department of Art and Art History developed a prototype database called QBIC (Query by Image Content) with IBM. QBIC is predominantly comprised of 1000 scanned images from the department's library. These images were selected by surveying faculty regarding their use of QBIC for teaching and a reference question log. Since users often search for images based on what things look like, rather than words used to describe things, QBIC was designed to compare images based on visual attributes. Preliminary findings showed that there were too many variations in the physical shape of the same subject, such as a horse drawn from the side or three quarters view, for QBIC to retrieve based on shape alone. Adding text to shape searching improved the retrieval results, though this search, as well as searches on faces and figures with similar skin colors, required five queries in order to yield a retrieval set in the mid-teens. Color and texture searches not related to shapes were the most accurate. While the query by example method is appealing, the effectiveness of this method is not explained in this study. Another weakness of the study is that the methods for developing the 9 possible search methods were not explained. Moreover, the results of QBIC's visual retrieval methods are not compared to results from purely textual retrieval methods.

Relation to CLiMB – CLiMB and QBIC both address the topic of searching for images based on their content, though QBIC proposes a non-linguistic technique for doing this. Art & Art History faculty, as well as students without art knowledge, are the intended audiences for both QBIC and CLiMB. QBIC and CLiMB also share the method of comparing similar entities for retrieving visual information.

Iyer, H., & Keefe, J. M. (2004, Spring). WordNet and keyword searching in art and architectural image databases. *VRA Bulletin, 30,* 22-27.

The use of WordNet to clarify concepts for searching, particularly, searching for architectural visual resources, is the focus of this exploratory study. Twenty participants were observed using the visual image database at Renssalaer Polytechnic Institute's Architecture Library to perform three tasks. For the first task, participants used WordNet to locate words similar to words they were given. 90% of participants found that WordNet helped them find terms for this task. For the second task, participants selected a group of images and used natural language or WordNet to create a caption that fit the group. The researchers observed that participants used a faceted approach, such as a time period, in order to create captions; and that general concepts were created from specific image attributes. Lastly, participants were shown images that they were asked to locate using natural language or WordNet terms. In this task, 30% of participants reported problems with conceptualizing the image, and 55% had terminology problems. Yet 70% of participants stated that WordNet clarified the terms or the images. Unfortunately, this study involved only 20 participants, and the results therefore lack generalizability.

Relation to CLiMB: Unlike other studies, this study tests the lexical resource used to find images or terms. A flaw of this study, which other studies may address, is how WordNet compares to similar resources, or to having no resources at all. This approach is similar to the four part study submitted to VRA members.

Keister, L. H. (1994). User types and queries: Impact on image access systems. In R. Fidel, T.B. Hahn, E.M. Rasmussen, P.J. Smith (Eds.), *Challenges in Indexing Electronic Text and Images* (pp. 7-22). Medford, NJ: Learned Information for the American Society for Information Science.

The Prints and Photographs Collection at the National Library of Medicine (NLM) analyzed user queries for the development of a videodisc access system. Retrospective analysis of reference requests showed that different user groups used the collection in different in ways. "Picture professionals (still picture researchers, TV, film, or media personnel)" thought visually and used art terms, like color; health professionals' requests consisted of queries such as "Do you have pictures of cholera?"; and museum professionals and academic researchers used precise citations to images (p. 9-10). The query analysis also identified different categories of requests: standard "subject topical term requests" for non-specific visual requirements ("Whatever you have on cholera in France…")

and topic-or subject independent "image construct queries" (p. 13). Image construct queries were distinguished by the use of words based on memory or need. They contained a visual construct behind them, for example, "the man sitting in the chair with a box on his head" or "surgeons standing". Keister also found that images were used in ways contrary to their original intent. Based on the findings from the query analysis, item-level cataloging of factual, visual elements ("the 'of' in 'of and about'") was incorporated into the videodisc access system. The image surrogate itself was also incorporated, because researchers' selection of images is based on seeing the actual image itself. While the use of an image varies from user to user, item-level cataloging of the "salient" visual elements provides a precise indicator of content. The use of even a few text based terms can serve to narrow a set retrieved for browsing.

Relation to CLiMB: Keister notes that pictures can't be retrieved if the words describing the picture are inaccessible (p. 13). This study indicates an interdependence between textual descriptions and image surrogates in the image retrieval process. This interdependent relationship suggests the need for images and text to be in close proximity to each other, a need which the CLiMB Toolkit facilitates. It should be noted that Keister opposes cataloging affective attributes of images, for the reason that requests such as "dramatic pictures" are too subjective to predict.

Matusiak, K. K. (2006). Information seeking behavior in digital image collections: A cognitive approach. *The Journal of Academic Librarianship, 32*, 479-488.

Pisciotta, H. A., Dooris, M. J., Frost, J., & Halm, M. (2005). Penn State's Visual Image User Study. *portal: Libraries and the Academy, 5(*1), 33-58.

This is an overview of some of the results of the Visual Image User Study (VIUS) conducted at Penn State University. (VIUS is described on its website as "an extensive assessment project to estimate needs for interdisciplinary image delivery" conducted over a 30 month period.) The research questions of the project were: "What are the design requirements for digital image delivery at Penn State? Will these requirements best be met by a single system or by multiple system?" The project resulted in 17 technical reports.

The authors discuss the VISOR and MESL projects, which influenced the VIUS project, and their methodology, briefly. 41 percent of faculty and 20.2 percent of students at Penn State replied to VIUS' largest surveys. Of these responses, 75.4 percent of faculty and 55 percent of students "use digital or analog pictures...for educational purposes". Yet "interestingly, a 'typical user' of digital images could not be rendered with the demographic data collected." The authors found that the use of image databases was moderate in comparison to use of University Libraries' other databases. They also found that, for both faculty and students, the most important factor for a digital image delivery system was content. Access to more pictures, a reduction of "labor in gathering and managing images," and "copyright and permissions sorted out" also ranked highly.

The project began with "a conceptual diagram of the functions that an image delivery system would need to meet in an academic setting": teaching, collection management, and independent learning. For faculty, facilitating teaching was considered more important than research for a digital image delivery system. Yet the authors state that "using large quantities of analog images is most associated with teaching", and attributes this to the difficulty in "technical support" of digital images. Faculty focused on use for teaching had less interest in a system with a breadth of searching techniques. The authors speculate that this may be because teachers are performing "'known item'" searching. A variety of search features and "well-described pictures" are needed for independent learning. Students placed more importance on search features than most faculty, except those conducting research.

One of the other findings is that 44.1 percent of the 639 faculty who use pictures maintain analog or digital collections for professional collections, and 44 percent of the 309 students who use pictures keep personal collections for education use as well. Many of the faculties' collections did not seem to be well described, one of the reasons why the VIUS group envisioned the ability for "associative metadata" as one of their "'ideal' database features."

The authors go on to describe their prototyping process, LionShare (a peer-to-peer system), and future research questions. These questions include "What are the more specific disciplinary characteristics of high-volume image users?" "Why and how do students collect images for their work? How do image-searching behaviors differ from search for bibliographic or textual information?" "What

can image systems do to provide associative data based upon the activities of users?" They conclude by suggesting that the project's inter-organizational team helped make it a success, and that users can contribute not only "media files" but different types of metadata as well to digital image delivery systems.

Relation to CLiMB: The VIUS reports themselves might be more helpful, but in providing an overview to the large scale VIUS project, this paper shows the full spectrum of user needs that VIUS addressed. The broad findings about academic user behavior that resulted from VIUS are the foundations for CLiMB's narrower focus on the linguistic aspects of images. CLiMB addresses the VIUS study's future research question of "What can image systems do to provide associative data based upon the activities of users?"

Specific

Armitage, L. H., & Enser, P. G. B. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Bakewell, E., Beeman, W. O., Reese, C. M. (1988). *Object, Image, Inquiry: The Art historian at work : Report on a collaborative study by The Getty Art History Information Program (AHIP) and The Institute For Research In Information and Scholarship (IRIS), Brown University* (Schmitt, M., General Ed.). Santa Monica, CA: Getty Art History Information Program.

Brilliant, R. (1988). How an art historian connects art objects and information. *Library Trends*, *37,* 120-129.

This article provides insight into the process of how an art historian uses information resources to establish an object as an *art* object. The process of studying an object begins with "It looks like…" and is then compared to similar objects. Generally these objects are accessed through information resources such as publications and reproductions, rather than through physical access to the objects themselves. Yet these critical resources may lack images of the objects themselves due to copyright restrictions. The use of online databases and high quality digitized image resources can facilitate access, but many online art resources, unlike those of the sciences, lack the older material required for art history scholarship. The narrow subspecialties in the field of art history result in the importance of subject-matter guides, and the field's changing nature of interpretation requires flexible databases.

Relation to CLiMB: This article shows how vital it is to make the images themselves accessible to art historians. The CLiMB Toolkit will make images more accessible by mining texts used to make historical associations between an art object and its historical context. Also, the enhanced subject access CLiMB will provide will assist those searching for an art object based on what "it looks like".

Chen, H.-L. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology, 52,* 260-273.

Through a two part user study, Chen examined the effectiveness of image classification categories proposed in earlier scholarly articles, namely Enser and McGregor (1992), Jörgensen (1995), and Fidel (1997). The first part of Chen's study involved 29 art history students formulating queries for a variety of art historical resources (it is important to note that the sources mentioned did not include image databases). For the second part of the study, these queries were reviewed by three information professionals and mapped to the most appropriate categories within each of the three image classification systems. To evaluate the effectiveness of each system in terms of user needs, Chen also had the information professionals rank how well the queries really fit into the chosen categories. Chen then compares the information professional's decisions to one another, and factors in their agreements and discrepancies. According to his findings, Enser and McGregor's categories of unique and non-unique were the most relevant categories for the information needs of art historians.

Relation to CLiMB – This study presents an interesting approach to evaluating image classification categories according to the needs of art history students - one of CLiMB's intended audiences. The significance Chen applied to the level of agreement between information professionals in categorizing queries stresses the importance of developing intuitive categories for consistent cataloging across databases. This will be an especially important consideration as we create the categories that CLiMB-mined terms will be mapped to.

Chen, H. L. (2007). A socio-technical perspective of museum practitioners' image-using behaviors. *Electronic Library, 25*(1), 18-35.

Garber, S. R., & Grunes, M. B. (1992). The art of search: A study of art directors. *CHI '92, Proceedings of the SIGCHI conference on human factors in computing systems,* 157-173; 703.

This study observed the process of art directors searching for images to create advertisements. Nine art directors, four art buyers, and two stock photo researchers were observed using a variety of techniques: interviews, analysis of steps in previous and current jobs, observations using think aloud protocol, and observation of art directors communicating image needs to art buyers. A work flow analysis based on this data showed that images are used at three stages of the search process: when the initial concept is created, during preparation of comps or demos used to obtain client approval, and at the beginning of the actual photo search. Throughout the process, images are set aside for later browsing. Image searching both altered and refined the initial query, as did viewing the images themselves. During the search process, images were used as "targets" or guides for finding similar images. The prototype developed from the study's findings allowed a searcher to use a known image to locate semantically similar or dissimilar items, and to weigh terms.

Relation to CLiMB: The use of the think-aloud protocol in this study provided step-by-step insight into the searching process.

Goodrum, A. A. (2003, February). Visual resource reference: Collaboration between digital museums and digital libraries. *D-Lib Magazine*, *9*(2). Retrieved April 3, 2007, from http://www.dlib.org/dlib/february03/goodrum/02goodrum.html

This is an overview of the goals for the Digital Museum and Library Project (DML), located at the Information Institute of Syracuse. The intent of the project was to extend the Virtual Reference Desk model to digital image reference in museums. Surveys and interviews of digital image reference providers in museums, analysis of image queries and responses, and the development of taxonomies of needs and iterative models would inform the creation of a digital image reference services for museums and libraries. Evaluation of the system would be conducted through analysis of transaction logs, surveys of experts and patrons, and "tracking question assignment patterns." The aim of DML's work was to advance digital reference capabilities and research on visual information seeking needs.

Relation of CLiMB: DML's proposed analysis of reference service providers and expert users is similar to CLiMB's analysis of both catalogers and users with subject domain knowledge. DML, like CLiMB, intended to develop a system based on the creation of image need taxonomies.

Goodrum, A. (2000). Image information retrieval: An overview of current research. *Informing Science*, *3*, 63-67.

Goodrum presents an overview of current research in image information retrieval and recommendations for future research. The article focuses on text-based retrieval, content-based retrieval, and "user interactions with image information retrieval". Text-based retrieval allows for precision in searching. The uniqueness of in-house collections resulting in in-house classification schemas; the need for description of both what an image is of and what it is about (Shatford, 1986); the lack of interindexing inconsistency; and the lack of accompanying textual information for images are some of the issues facing text-based image systems. Content Based Image Retrieval (CBIR) lacks research on which attributes are useful for an image-based approach. Image user interaction studies tend to focus on specific groups, and on non-digitized collections. Some studies have noted the use of narratives to describe images (O'Connor, 1999; Jörgensen, 1995); the tendency for either specific searching or browsing in image information retrieval is another observation of image user studies (Goodrum, 1997; Fidel, 1997). Research on interface support of image information retrieval systems is recommended in order to facilitate browsing, search modification, and to discover which attributes would be useful for systems design. Goodrum notes that a cross-disciplinary approach combining textual and content-based retrieval features may improve retrieval effectiveness. Goodrum concludes by recommending large scale image related projects similar to TREC.

Relation to CLiMB: The observations of catalogers and end-users conducted by CLiMB focuses on the browsing, search modification, and observation that Goodrum emphasizes. CLiMB seeks to test and examine the effectiveness of different image attributes for retrieval. CLiMB also seeks to gather data from catalogers and end users specifically for interface design purposes.

Layne, S. S. (1994). Artists, art historians, and visual art information. *The Reference Librarian*, *47,* 23-36.

This discussion of visual art information focuses on why it is needed, how it is used, what its characteristics are, and how to make it accessible (p. 24). Artists need images *of* things and images with particular evocative meanings. Images of artwork, as well as the literature describing the artwork, is needed for art historians. Both groups browse and compare images in order to select them. Technological developments such as videodiscs resolve physical access problems, yet there is also a need for intellectual access to the often problematic characteristics of images, such as artists' names, artwork titles, physical characteristics, and subjects, etc. A summary of the cataloging and indexing tools that help solve intellectual access problems is provided, as well as the existing methods for obtaining access to the characteristics of images. The discussion concludes by stating that the role of reference librarians is to understand artists and art historians' needs, the problems of visual information, and the available tools and methods for providing access to visual information (p. 34). Reference librarians can use their knowledge of the problems of visual information access to influence the creation of visual information research tools.

Relation to CLiMB: This overview highlights distinctive issues in artists' and art historians' access to visual information resources. The aspect of serendipitous discovery as discussed in artists' information seeking behavior could be observed in the CLiMB user studies.

Marty, P. F. (2007). The changing nature of information work in museums. *Journal of the American Society for Information Science and Technology, 58,* 97-107.

This study sought to explore the changing nature of museum information work by investigating the challenges faced by museum information professionals (MIPs), and the coping strategies employed for managing these challenges. From semi-structured interviews of 21 museum information professionals, a set of coping strategies and challenges common to the majority of MIPs studies was found. The findings include a need to first understand the impact of new capabilities provided by new technologies, through the "ability to implement new information systems and solve technological problems" (p.100). Another challenge, that of understanding museum users' changing needs, can be met by understanding both current and future users' needs, and by balancing the needs of all users of museum information resources needs, including those of visitors and scholars. Finally, in order to meet these diverse needs, MIPs need to adapt to users' changing expectations of museums. One means MIPs can employ for managing changing expectations is to encourage other museum professionals to share and communicate information with each other in order to break down barriers. Further research into the information behavior of museum information professionals will not only help MIPs understand the changing capabilities, needs, and expectations of museums, and means for coping with them, but also help guide future museum information research.

Relation to CLiMB: This article's concern for museum information professionals and museum informatics might be beyond the scope of the CLiMB project. However, future research topics described in the article's discussion and conclusion intersect with the work of CLiMB: "How can museums websites blur boundaries (e.g through podcasting or personal digital collections) between online and in-house museum resources?" and "Another possible line of inquiry concerns the growing dependence of MIPs on their abilities to work with online repositories of shared collections data or digital images."

Stam, D. C. (1984, Winter). How art historians look for information. *Art Documentation* 3(4), 117-119.

Relevance Assessments

Chang**,** C-C., & Wu, T-C. (1992). Retrieving the most similar symbolic pictures from pictorial databases. *Information Processing & Management*, *28*, 581-8.

Choi, Y. & Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing and Management 38*, 695 – 726.

Choi and Rasmussen report on their user study involving historians searching for images in LC's American Memory photo archives. The study examined both the aspects of the image, and any

associated text, that were considered when making relevance judgments. In addition, Choi and Rasmussen discuss observed shifts in the prioritization of relevance criteria from the problem solving (query formulation) stage to the information retrieval stage. On average, users' top three *pre*-search priorities were topicality, accuracy, and completeness. After viewing retrieval sets, the top three were topicality, time-frame, and accessibility. Overall, topicality remained the greatest determining factor but the exact mean changed throughout the process (p.708). They also tracked the impact of associated text on users' relevance judgments. According to their findings, date, notes, title, and subject descriptors were the most useful categories of text for evaluating an image's relevance. The authors do, however, acknowledge that these factors will likely vary for different user groups—art historians, for example, may find the name of the creator or photographer highly critical in their decision making.

Relation to CLiMB – The categories of relevance criteria (see pages 701 and 705) discussed in this report can provide a jumping off point for identifying types of terms we would like the CLiMB toolkit to identify. Additionally, the report includes the interview questionnaires employed for their study and these may help inform the development of our own written and oral surveys. Finally, Choi and Rasmussen suggest applications of their findings to the design of an image retrieval system, including browsing and comparison capabilities, user controlled displays, and interactive query refinement.

Layne, S. S. (2000). Modeling Relevance in art history. In T. H. Connell, and Maxwell, R. L. (Eds.), *The future of cataloging: Insights from the Lubetzky Symposium, April 18, 1998, University of California, Los Angeles* (pp. 33-41). Chicago: American Library Association.

Lu, Y., Hu, C., Zhu, X., Zhang, H., Yang, Q. (2000). A unified framework for semantics and feature based relevance feedback in image retrieval systems. *Proceedings of the eighth ACM international conference on multimedia*, p.31-38.

Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., & Field, M. Semi-automatic image annotation. Retrieved July 20, 2007, http://research.microsoft.com/asia/dload_files/group/mediasearching/fu-107-Semi-Automatic-Annotation-4th.pdf

Wenyin proposes a system that relies on both automatic and manual indexing, with end user input for relevance ranking. As new images are entered, the system automatically identifies content-based features, such as color and texture, and uses these features to query similar images already in the system. The system then retrieves a subset of similar images for user feedback. From those images which the user ranks as most relevant, the system extracts the text annotations and generates a list of keywords, ordered by frequency, and applies them to the new image. Keywords are weighted according to their source (i.e., manually entered, automatically entered, and/or end-user ranked). However, there does not seem to be much consideration of how the perception of relevancy varies from user to user, especially in instances where the same keywords are used. In addition, there is no defined user group for submitting user feedback, presenting obstacles to the creation of an authoritative source.

Relation to CLiMB: Despite these shortcomings, Wenyin's system offers an interesting model for measuring the relevancy of CLiMB-mined terms. Incorporating user feedback, especially in terms of keyword evaluation, will be an important part of our cataloger and end-user studies. Weighing relevance, as in Wenyin's system, will also be worth consideration.

HCI

Bates, M. J. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science, 49,* 1185-1205.

Human behavior, statistical properties of databases, and domain specific issues are factors which should be taken into account when designing indexing and access mechanisms for information retrieval. In the context of the increasing amount of Internet information, Bates presents information science findings which should be well known to information scientists, and which information systems designers should be made aware of. One finding relevant to CLiMB work is the mismatch between indexing behavior and end user searching behavior. Another finding is the

recognition of human language's universal tendency towards folk classification. Consequently, Bates suggests using folk classification as a means of access in interface design. In describing research on statistical indexing properties of databases, Bates explain that Bradford's Law and Zipfian distributions are human phenomena which should be taken into account when information retrieval systems are designed. Finally, the importance of domain is described. Research on domain vocabularies suggest that the different vocabularies unique to different domains result in different information needs – and therefore, the need for systems tailored for different domains. Bates cites the BIOSIS database as an example of domain specific design. Yet other examples of research into and design of human, statistical database, and domain factors are lacking – especially in the area of folk classification (not to be confused with folk taxonomies and social tagging).

Relation to CLiMB: Bates discusses the *AAT* as an example of a faceted approach useful to indexing, stating, "Using this and other indexing technique in combination with a sensitivity to database and domain factors can produce great retrieval power." CLiMB is using the AAT, but research into database and domain factors will need to be accomplished. Research conducted by BIOSIS designers, discussed in this paper, might provide an exemplar for CLiMB, and leads to the questions of: will catalogers really want all of the words matched to the Getty Vocabularies retrieved? And will users want the same amount and types of words retrieved? Observations of Getty vocabulary use might help answer these questions. Bates indicates that the ULAN provides three separate uses: it collocates artist's names, thus saving time for researchers, aiding in retrieval, and functioning as a "scholarly resource in its own right." The domain and database use of the Getty vocabularies online might provide findings helpful for the Toolkit's development.

Beard, D.V. (1991). Computer human interaction for image information systems. *Journal of the American Society for Information Science*, *42*(8), 600-608.

The development of viable image information systems involves the Computer Human Interaction (CHI) techniques of requirements analysis, image interface design, feedback and evaluation. Methods for conducting CHI are presented and applied to FilmPlane, a radiology workstation.

Relation to CLiMB: This proved to be a helpful reminder of the steps involved in computer-human interaction testing. Application of CHI methods to a domain outside of art-history serve as a reminder of the breadth of methods available for interface testing. Its concise presentation of CHI methods as applied to the FilmPlane project is so readable that it is recommended as a brief introductory reading to CHI for all CLiMB project members.

Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2003), 5*(1), 401-408.

Two image interfaces were compared in this study. A baseline interface was designed to incorporate the best features of six web image search engines. The other interface was designed using faceted categories, and was intended to facilitate both browsing and searching. Using WordNet, categories were extracted from descriptions of the images. To compare the interfaces, 32 art students used structured and unstructured queries to search a collection of approximately 35,000 images from the Fine Arts Museum of San Francisco.

The faceted category design performed well overall: 91% of the participants stated that that they preferred this design to the baseline interface. Some of the other findings include an 81% successful retrieval rate using the faceted interface to find all aquatints in France, compared with a 57% success rate using the Baseline interface. For finding all woodcuts made in the United States, a 77% success rate with the faceted category interface and a 21% success rate with the Baseline interface were reported. Also, participants constructed queries from multiple facets in unstructured tasks 19% of the time, and in structured tasks, 45% of the time. The authors conclude that their interface allows for flexible navigation of hierarchical, faceted metadata.

Relation to CLiMB: This research on a faceted interface will facilitate CLiMB toolkit interface design, especially thru its positive comparison to pre-existing Web image search engine interfaces. The comparison of unstructured and structured queries, or known item searching, is an area that CLiMB could expand upon in its interface testing.

System Design

Chang, Shi-Kuo. (1989). *Principles of pictorial information systems design.* Englewood Cliffs, NJ: Prentice Hall.

Chang, S.-F., Smith, J. R., Beigi, M., & Benitez A. (1997). Visual information retrieval from large distributed online repositories. *Communications of the ACM, 40*(12), 63-71.

WebSEEk and MetaSEEk were developed to address the need for large, Internet-based, visual information retrieval systems. The visual information retrieval system WebSEEk extracted key terms from URL addresses and html tags of images and videos, then mapped the terms to its semantic ontology (p. 67). This mapping was over 90% accurate. 1,600,000 queries were conducted on WebSEEk. 53.5% of these queries were subject-based. While 3.7% of queries were found to be content-based, the lack of content-based queries may have been due to search limitations of the content-based search functions. Targeting the VisualSEEk/WebSEEk, QBIC, and Virage search engines, the meta-search engine MetaSEEk allowed users to search "based on image examples or text" (p. 68). The authors conclude with a discussion of the problems of heterogeneity, complexity, and bandwith in scalable internet visual retrieval systems. Based on these problems, the authors recommend eight critical components to be incorporated in large online visual repositories.

Relation to CLiMB: The various features and recommendations described could be useful for CLiMB Toolkit design.

Constantopoulos, P., & Doerr, M. (1995). An approach to indexing annotated images. In D. Bearman (Ed.), *Multimedia Computing and Museums: Selected Papers from the Third International Conference on Hypermedia and Interactivity in Museums (ICHIM'95/MCN '95)* (pp. 278-298). Pittsburgh, PA: Archives & Museum Informatics.

A semantic data modeling approach to image indexing provides a flexible structure for evolving and heterogeneous user needs. The Semantic Index System (SIS) is described and proposed for use in CLIO, a proposed cultural documentation system. The rationale for the semantic data modeling approach is based on the wide variety of end user goals. Image end-users interact with images both as resources to study, and to re-use elsewhere. Unfortunately, no data to support this perspective is presented. An example of an image with semantic relationships expressed is provided.

Relation to CLiMB: The SIS functions as a purely semantic intermediary between images and end-users. The purpose of the intermediary is to reduce "the search space, without loss of recall." This is a perspective which CLiMB should consider, as it studies the efficacy of the Toolkit. Another point made Like Fidel (1997), this is task-based perspective to image information retrieval.

Fidel, R. (1997). The image retrieval task: Implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia, 3,* 181-199.

An analysis of 100 requests for stock photos indicate that images are used either as sources of data and information, as objects in-and-of themselves, or as a combination of these two categories. Building off of Jorgensen's research (1995), Fidel found that attributes of images used for these requests differed from the attributes used for describing, searching and sorting - though most resembled the searching task. The use of different attributes for different tasks leads to the conclusion that images may be used for simply information sources at one extreme, and as objects that convey ideas, such as in stock photo use for an advertisement, at the other. Searching behavior for these two poles, called the Data Pole and the Objects Pole, have different characteristics: images used for data would require less browsing, yet greater use of textual cues, whereas browsing is highly important when images are used as objects; ofness equals aboutness on the Data Pole, though ofness and aboutness differ on the Object Pole. The difference in attributes indicates the need for variable measures for relevance, performance measures, and precision in image database design.

Relation to CLiMB: Fidel's research is particularly applicable to designing the end-user searching function of the Toolkit. First, CLiMB must be both aware of the tasks that Toolkit users will desire and analyze the searching behavior of Toolkit end-users, including eliciting unintended uses from anticipated end users. CLiMB end-user research may also fill the gap of the lack of research in human browsing.

Jose, J. M., & Harper, D. J. (1995). An integrated approach to image retrieval. *The New Review of Document & Text Management*, *1*, 167-181.

This paper proposes a semi-automatic image retrieval method that combines text and picture retrieval techniques. The proposed method consists of an indexer identifying an object and its location by drawing a box around the object and naming it. When queried, the Éclair system would match the documents' textual and visual descriptors to those of the query. Documents with the largest score would be returned. Thru the use of both textual and visual retrieval techniques, the prototypical system could facilitate information discovery and recovery and ease of use for both indexers and users.

Relation to CLiMB: Jose and Harper note that the key problem of images in text-based retrieval system is their lack of associated text (p. 168). CLiMB's solution to this problem is to first find the associated text for images, then retrieve words matching the controlled vocabulary of the AAT, and finally, to have catalogers decide which AAT term should be presented to end-users. Interestingly, Jose and Harper similarly mention investigating the use of a thesaurus in the future (p. 180). This paper also leads to the question of how effective a CBIR based thesaurus in combination with the CLiMB toolkit's textual thesaurus matching might be for image retrieval systems.

Shen, H. T., Ooi, B. C., & Tan, K.-L. (2000). Giving meanings to www images. *Proceedings of the eighth ACM international conference on Multimedia,* 39-48.

The Weight ChainNet model proposes retrieving images on the World Wide Web by creating lexical (semantically related) chains out of the text surrounding the images. A preliminary study determined that the parts of the webpage to be used for retrieval would be: the image's file title, the image's ALT tag (alternate text), the image's caption, and the page title for the image. (The researchers used these sections, rather than the entire HTML document, because they expected the entire webpage to introduce too much unrelated information into the retrieval process.)

The resulting weighted lexical chains formed the Weight ChainNet model. To test different feedback methods of the model, the researchers used fifteen queries on 5232 images from over 2000 different URLs. The first feedback method tested, semantic accumulation (Accu), allows users to choose the single most relevant image from their search results for further refinement. The semantic integration and differentiation method (I & D) allows both relevant and irrelevant results to be selected. This experiment found that the Accu method had a higher recall rate than the I & D method, but that the I & D method outperformed the Accu method overall. The authors attribute their results to two reasons: first, Accu uses only one image, and unlike I & D, doesn't eliminate irrelevant images (or "noise"). Second, I & D uses "the most semantically related LC from each relevant image's ChainNet" to form a new query (p. 43), serving to eliminate irrelevant information. The authors conclude their paper by proposing to extend their Weight ChainNet Model for content based information retrieval, and to use WordNet to extend each lexical chain.

Some questions are left unanswered by this article. First, the preliminary study determining the most useful parts of HTML text for image retrieval should have been presented or cited. Providing this information would not only serve to substantiate the study, but would also serve to provide insight into where semantic meaning lies in websites. Second, the article did not seem to validate its claim that the Weight ChainNet model "methods outperform existing techniques" (p. 39). These "existing techniques" did not appear to be presented, nor were they compared against Weight ChainNet results.

Relation to CLiMB: Like CLiMB, this article is concerned with retrieving images based on their associated texts. The authors' perspective that the Weight ChainNet model captures "the semantic representation of the content of an image" makes explicit a principle motivating CLiMB work as well, and again, provides a reason why providing results of testing on various parts of the webpage would've been helpful in showing how these parts "represent" images differently. [It is difficult for me personally to assess the mathematical formulas in this article, and to show how this article is related to current computational linguistic work. Is CLiMB also weighting and re-ordering lexical chains?]

Srihari, R. K., Zhang, Z., & Rao, A. (2000). Intelligent indexing and semantic retrieval of multimodal documents. *Information retrieval, 2*, 245-275

Srihari, R.K., Rohini K., & Zhang Z. (1999). Exploiting multimodal context in image retrievalT. *Library Trends*, *48,* 496-520.

Stoica, E., Hearst, M., & Richardson, M. (in press). Automating creation of hierarchical faceted metadata structures. *Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Projects

Hastings, S. K. (Project Director). User Evaluation in the Retrieval of Digital Art Images: The Bryant Art Images. School of Library and Information Sciences, University of North Texas, 1999. Project report retrieved July 20, 2007, http://courses.unt.edu/shastings/HastingsWWW/digital_images/

4.4

4.5 Hearst, M. (Project leader). The Flamenco Search Interface Project: Search Interfaces that Flow. Retrieved July 20, 2007, from http://flamenco.berkeley.edu/

4.6 This NSF funded project taking place as of July, 2007,

Rose, T., Elworthy, D., Kotcheff, A., & Clare, A. ANVIL: A system for the retrieval of captioned images using NLP techniques. *Challenge of Image Retrieva*l, Brighton, 2000. Available through Google Scholar, last accessed July 31, 2006.

This article discusses the functionalities and general design of ANVIL, a system for applying NLP techniques to Japanese and English captions. Functionalities include query expansion, enabled through ID's shared by hyponyms, hypernyms, and synonyms. Two interfaces were designed and tested for usability. The first displays a traditional list of recalled images. The second has a 2-d spatial layout that clusters recalled images based on relevance. Precision tests found the ANVIL system to achieve between 85 and 86% precision for keyword searches, and between 92 and 95% precision for phrase matching.

Relation to CLiMB: Though focused on captions rather than more extensive structured and unstructured texts, ANVIL appears to be very much a parallel project.

Appendix 7: Internal Advisory Board Member List

1. Joseph Danks, Research Professor
   Center for Advanced Study of Lanuage (CASL), University of Maryland

2. Chris Higgins, Manager of Learning Technologies
   University of Maryland

3. Matt Kirschenbaum, Associate Professor of English; Associate Director of the Maryland Institute for Technology in the Humanities (MITH)
   University of Maryland

4. Kari Kraus, Assistant Professor at the iSchool
   University of Maryland

5. Charles Lowry, Dean of Libraries
   University of Maryland

6. Doug Oard, Associate Dean for Research for the College of Information Studies; Associate Professor in the iSchool and the Institute for Advanced Computer Studies (UMIACS)
   University of Maryland

7. Jennifer Preece, Professor and Dean of the iSchool
   University of Maryland

8. Philip Resnik, Associate Professor in Linguistics Department and UMIACS
   University of Maryland

9. Lauree Sails, Director of the Visual Resources Center
   University of Maryland

10. Susan Schreibman, Assistant Dean of Libraries, Head of Digital Collections and Research
    University of Maryland

11. Martha Nell Smith, Professor of English and Founding Director of MITH
    University of Maryland

12. Joan Stahl, Branch Manager, Art & Architecture Libraries
    University of Maryland

Appendix 8: External Advisory Board Member List

1. Marcia Bates, Profesor
   Graduate School of Education and Information Studies, University of California, Los Angeles

2. David Fenske, Dean of the iSchool
   Drexel University

3. Clifford Lynch, Director
   Coalition for Networked Information

4. Carol Mandel, Dean of the Libraries
   New York University

5. Tom Moritz, Associate Director of Administration and Chief of Knowledge Management
   Getty

6. Brian Schottlaender, University Librarian
   University of California, San Diego

7. James Shulman, Executive Director
   ARTstor

8. John Unsworth, Dean and Professor
   Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign