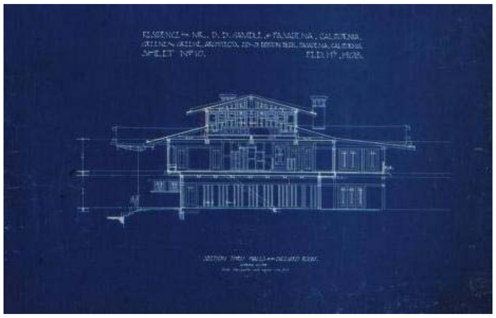
CLiMB

Computational Linguistics for Metadata Building



Center for Research on Information Access
Columbia University

First Year Report
March 2002 - February 2003

CLiMB Progress Report
March 1, 2002 to February 28, 2003
(Report covers developments through June 1, 2003)

Executive Summary
Goals of the CLiMB Project iii
Highlights of Progress - The First Year of CLiMB iii

Preface: Elaborating the CLiMB Problem 1

1. CLiMB Collections 2
1.1 Establishing Criteria for Selecting Texts and Images 2
1.2 Generating the Initial Dataset 4
1.3 Understanding the Collections – the TOI Concept 5
1.4 Working with the Collections 6
1.41 Greene & Greene Architectural Images 6
1.42 Chinese Paper Gods Collection 8
1.43 South Asian Temples Images and Metadata 9

2. CLiMB Tools 10
2.1 Designing System Architecture 10
2.2 Choosing and Testing Pre-existing Tools 14
2.3 Understanding the Datasets – the TOI Concept 15
2.4 Creating Original Algorithms and Refining the Toolset 16
2.41 Identifying Target Objects Automatically 17
2.42 Using Subject-oriented Vocabulary 20
2.43 Employing Segmentation Techniques 21

3. CLiMB Access Platform 24
3.1 Presuppositions about Users 24
3.2 Building a CLIMB Testing Platform 25
3.3 Collecting Data on How People Use CLiMB Information 25

4. CLiMB People 27
4.1 CLiMB Organizational Chart 27
4.2 CLiMB Project Teams 27
4.3 Developing the External Advisory Board 28

5. Taking Next Steps 30

6. Selected References 31

Appendices: CLiMB Supporting Material 32
A. Additional Information on CLiMB 32
B. Collections and Related Material 32
C. Sample Catalog Records 33
D. CLiMB Presentations & Talks: May 2002 – May 2003 38
E. CLiMB Glossary of Terms 39

Acknowledgements

This report was a collaborative effort from the entire CLiMB project team. However, the bulk of the compilation and editing was done by Mark Weber, the CLiMB project assistant.

For reference purposes, authors are:

Judith Klavans, PI
Roberta Blitz
Peter Davis
Stephen Davis
David Elson
Angela Giral
Amy Heinrich
David Magier
Patricia Renfro
Bob Scott
Mark Weber
Bob Wolven

Executive Summary

Goals of the CLiMB Project

The CLiMB project at Columbia University aims to discover to what extent and under which circumstances automatic techniques can be used to extract descriptive metadata from texts associated with image collections. Ordinarily, descriptive metadata (in the form of catalog records and indexes) are compiled manually, a process that is slow, expensive, and often tailored to the purpose of a given collection. Our goal as a research project is to explore the potential for employing computational linguistic techniques to alleviate some of the cataloging bottleneck by enhancing descriptive metadata with automatic procedures. If successful, the CLiMB project will enable the identification of highly-ranked terms by extracting them from written material associated with images in digital collections. Additionally, among our objectives is the creation of a set of tools that can ultimately and easily be used by other projects, in order not only to enable sophisticated automatic indexing procedures, but also to enhance access for end-users by labeling descriptive metadata for review by experts. CLiMB includes an extensive evaluation component, which will measure the effectiveness of the tools at extracting desired information. This will lead to the ability to assess the usefulness of this information once included in image search platforms. To our knowledge, CLiMB is engaged in a unique approach to issues of automatic metadata extraction from selected authoritative texts.

Highlights of Progress - The First Year of CLiMB

CLiMB's project teams have made substantial progress in several areas, beginning with the creation of a strong conceptual and organizational foundation, which was then applied to the development of the CLiMB collections and the implementation of the CLiMB toolset.

Technical

- Designed an overall project architecture to facilitate the realization of the project's goals and to integrate smoothly the inherently interdisciplinary aspects of the project
- Mapped a strategy for using computational linguistic tools and techniques in conjunction with external resources such as book indexes, authority files, and other controlled vocabularies
- Established the concept of the Target Object Identifier (TOI) as a way of structuring curatorial data and focusing technical developments

Collections

- Elaborated a set of working guidelines for choosing collections for the project
- Selected three diverse collections of images, and texts associated with them, as datasets:
    - Images of Greene & Greene architectural drawings at Avery Library
    - Images of Chinese Paper Gods from the Starr East Asian Library collection
    - Images of South Asian Temples from the American Institute of Indian Studies Photographic Archive

Highlights of Progress – continued

Collections – continued

- Began the process of formatting the images and associated texts as testbed datasets for testing
  - Scanned text of several key monographs
  - Developed catalog record templates for previously uncataloged image collections
- Established detailed guidelines for defining relevance in terms of metadata extraction
- Explored initial evaluation methodologies for
  - Determining the accuracy of subject related terms derived from texts
  - Verifying the usefulness of this metadata for users
- Secured access to the Getty Trust's *Art & Architecture Thesaurus*

Testing

- Performed initial tests of natural language processing software over selected documents for identification of
  - Common noun phrases
  - Proper noun phrases
  - Segment boundaries in texts
  - References to projects (Target Object Identifiers or TOIs)
  - Domain (or "subject") specific vocabularies
- Improved existing natural language processing tools by refining algorithms and developing innovative uses for these tools
- Began implementation of the CLiMB suite of tools, of which the TOI Finder was the first example; subsequently enlarged the suite to include other term matching tools
- Built a user interface for the TOI Finder, to display results for user evaluation and manipulation; available on-line at http://www.columbia.edu/cu/cria/climb/ tools.html
- Moved towards the development of a more comprehensive testing platform for collections and toolsets, of which the TOI Finder's interface is an early example

Administrative

- Determined the overall organizational structure of CLiMB project groups
- Fulfilled all internal staffing requirements
- Set a timeline for two year project completion
- Invited a select group of experts to serve on the External Advisory Board, which will hold its inaugural meeting in June

Preface: Elaborating the CLiMB Problem

The innovative contribution of the CLiMB project is its use of computational linguistic tools in order to extract metadata from text associated with images in large digital collections – hence **C**omputational **Li**nguistics for **M**etadata **B**uilding. Put in basic terms, the project applies techniques developed in computational linguistics to the problem of cataloging and describing the images in a given collection; it does this by putting to use information that already exists in the form of written material related to those images. Although manual cataloging is an established field, what is novel about the CLiMB approach is the notion that some cataloging might be accomplished automatically. The achievement of CLiMB's goals will not only address the cataloging bottleneck that arises as the volume of available data increases with new technology; it will also benefit end-users by providing a set of tools that will potentially aid in the access of information across collections and vocabularies. As a research project, CLiMB seeks to create a platform in which to determine whether such an approach is indeed useful, and to what extent results can be incorporated into existing metadata schema.

When a collection of images is cataloged, art librarians may provide intellectual access to the material by selecting relevant name and subject terms from existing authority lists and controlled vocabularies. If, for example, a collection contains images of artworks, an art cataloger can employ his or her knowledge of the material to select the applicable terms from among such lists and vocabularies, which can then be entered into a catalog or index. These terms answer a fundamental set of questions that might be put to the collection by potential users when they are searching for images: *What does this image depict? What is the object made out of? What other physical characteristics does it exhibit? Who created it? What is its historical and cultural significance?* The benefit of using experts to perform this task is that they possess an extensive and invaluable understanding of the material. However, given the incredible size of the digital collections now being created, it can be both too expensive and too time-consuming to accomplish this manually with the full level of detail to provide maximum user access.

In many cases, researchers have already described aspects of these images in contexts such as scholarly monographs and subject specific encyclopedias. CLiMB employs text sources that closely mirror the content of a given digital image collection to automatically extract descriptive metadata from those texts – in effect, making the writings of specialist scholars useful to enrich the catalog. The challenge is to identify the *meaningful* facts (or metadata) in the written material and distinguish them from among the thousands of other words that make up a text in its original form. This is the task that CLiMB is currently undertaking. The end result will be the CLiMB suite of tools: software that can be embedded in existing search platforms to extract metadata, which can then be used for cataloging images from collections that might otherwise be unmanageably large.

This first year report describes the three primary areas of progress. First, collections with varying characteristics were selected in order to ensure that our research would permit us to discover how different techniques might produce a range of results for distinct collections. Second, at the same time, we built a platform for running and testing tools to run over text data in real time. In the upcoming year of the project, we will initiate evaluation with users; thus, the third area is addressed as part of our future work, although we have already established a forward-looking infrastructure. The fourth and fifth sections of this report explain the organization of the CLiMB project teams, and the projected next steps of the project, respectively.

1. CLiMB Collections

1.1 Establishing Criteria for Selecting Texts and Images

CLiMB operates under the assumption, which was borne out by preparatory investigation, that there already exist image collections and search platforms into which the results of our research can be embedded. A fundamental step in developing the CLiMB toolset, however, is the establishment of criteria for what types of collections can most benefit from the automatic extraction of rich metadata. Because workable datasets are absolutely necessary for testing software tools and user interfaces alike, creating these sorts of guidelines is an imperative starting point for the project. Prior to any testing of computational linguistic tools, image collections must be chosen or developed such that automatically extracted metadata can be applied to the particular problems of cataloging or searching an actual collection. Likewise, CLiMB extracts information from texts such as scholarly monographs, which provide implicit descriptions of the items in image collections; thus, it is just as important to determine which texts are most strongly "associated" with a given collection—a task that itself demands an examination of the relationship between texts and the images they describe. Currently, the CLiMB project is working with three image and text collections:

- Greene & Greene Architectural Records and Papers Collection, housed in the Avery Architectural and Fine Arts Library at Columbia University
- Chinese Paper Gods Collection, housed in the C.V. Starr East Asian Library, also at Columbia University
- South Asian Temples images and metadata, from the American Institute of Indian Studies (AIIS) photographic archive in the Digital South Asia Library

These collections are described in detail below in section 1.4. This section describes the considerations that led to their selection for use on the CLiMB project.

Early during the first year of CLiMB, as well as during the exploratory research period that preceded the actual start date of the project, the CLiMB Curatorial Group established a detailed set of specifications for what sorts of image collections and associated texts would be applicable to our goals. The group recognized that before preparing workable electronic datasets, it was necessary to categorize the types of relationships that exist between images and texts for the purposes of this project. For example, in certain books and collections, written material serves to explicate—and thus in a sense is generated by—particular images. In other instances this is not the case; the relationship can be reversed, as when images illustrate written text, or when there is little or no clear relationship between text and image at all. In order for the project to move forward smoothly it was necessary to choose an initial dataset that did not present avoidable complications for preliminary tests of natural language processing software tools. They therefore decided that it was essential to begin by selecting data in which the written material had a clear explanatory relationship to the images.

The group also determined that it was important to prevent too large a variety of text types from diffusing the initial results, which could potentially lead to a lack of depth in the early stages of the study. This problem can be addressed in two ways, the first of which is careful consideration of the ratio of the number of texts to the number of images in any given dataset, an issue that can be managed during the planning and formatting stages of data preparation. At present, CLiMB employs a small number of carefully selected texts to a comparatively large number of images. The second solution is to narrow down the subject areas of the image collections and related texts under consideration. Thus, the group decided to focus on collections that were art and architecture related. Image collections in these fields provide a good balance of images and written material; in addition, for these types of collections the relationship between text and image is usually such that the former (text) more clearly serves the purpose of describing the latter (image).

Once this set of preliminary considerations had been outlined, members of the Curatorial Group set about defining more specific criteria that could be applied to the process of compiling and preparing image collections and texts for use on the CLiMB project. Four significant items, which were applied throughout the selection and development process, emerged from these discussions.

- The group decided that one important aspect of selecting images should be image content type, which can be broadly categorized into two groups: primary items (for example, a building) and secondary items (for example, blueprints of a building). Each collection now being used by the CLiMB project exhibits a different concentration of images across these two groups. The Greene & Greene architectural images are largely secondary; the Chinese Paper Gods Collection is comprised entirely of primary items; the South Asian Temples images are largely, though not exclusively primary. The images in the last collection differ from the Chinese Paper Gods images in that they are not directly scanned artworks, but photographs of temple sites.
- Another criterion for image selection takes account of the properties of the text associated with a particular group of images. Questions that should be asked about associated text include that of whether a given work is generated specifically around (or for) the images in a given collection, and, just as importantly, whether the accompanying text is in some sense authoritative. The choice of initial texts for each of the three collections reflects this spectrum of possibilities. For the Greene & Greene collection, the text consists of a recent and well-respected scholarly monograph that is not directly related to the image collection, but discusses at length the major projects that also appear in the images; for the Chinese Paper Gods a classic study was selected, which was written about the collection itself by the donor of the images; for the South Asian Temples, work will begin with a comprehensive encyclopedia that contains specific images from the collection, as well as descriptive metadata about images discussed in the text.
- A third consideration is whether pre-existing authority lists are related to a given text or collection. For example, the selection process had to account for indexes or catalog records related to the written or visual material in a given collection; a similar matter is whether the images in a particular collection have captions. Also important is the question of whether there is already existing database information for any potentially applicable collection, as well as whether that data could be made available to CLiMB for use on the project. The Greene & Greene collection at Columbia University has already been digitized and cataloged at the project level; a selection of the Chinese Paper Gods images are currently being scanned by library staff and given preliminary catalog records in conjunction with the CLiMB project; the South Asian Temples images already exist in a digital, on-line format, but the associated scanned text will have to be processed before it can be put to use.
- A final important issue in image selection was that of whether a "gold standard" could be easily built for a given collection. In other words, the group realized that it was valuable to consider whether there was already enough catalog information for a collection such that

> CLiMB could measure its impact in terms of enhancing or replicating the information that already exists. This issue to a large extent depends upon the results of point three above, but all of the collections CLiMB currently utilizes have presented many different options for measuring the effectiveness of CLiMB metadata extraction capabilities. These range from employing existing catalog records or back-of-book indexes as benchmarks, to having users evaluate the toolset in a search platform or other interface. Thus, exploring the value of these different "gold standards" themselves—deciding which are the most relevant for measuring effectiveness—has become another important part of the project.

1.2 Generating the Initial Dataset

The consensus among the members of the CLiMB Curatorial Group was that multiple testbed collections were desirable as datasets for the long-term goals of the project; for this reason, three collections were decided upon, and their development has proceeded during the first year on a staggered timeline. As the section above suggests, the selection process was guided by a desire both to limit the dataset to workable proportions and at the same time to achieve the breadth necessary to make the results of software testing valuable.

However, the group also agreed, in conjunction with the Technical Group, that it was important to begin with a single collection of images and a single associated text before eventually adding others. The decision was thus made at the beginning of the project to use the Greene & Greene Collection as the initial set of images and catalog records, and to use Edward R. Bosley's book *Greene & Greene* as the related monograph over which to start testing the metadata extraction capabilities of computational linguistic software. Although Bosley's book is not the only available text for this image collection, it does represent an authoritative text on the topic; Chapter 5 of that book was chosen because it presented an interesting and clearly workable dataset that discussed major architectural projects. The Greene & Greene image collection at Columbia University's Avery Library is comprised mainly of architectural drawings and papers, and it represents a good starting point on several accounts. Foremost among these is that the relationship between text and image is not one of direct correlation: the challenge of using *Greene & Greene* with these images is that the text only "implicitly" speaks about the items in the collection. In order to prepare it for testing, the text was converted to TEI Lite XML in three working versions: plain XML, XML with a minimal stylesheet, and plain text.

Despite the clear benefits of selecting Bosley's text, the Curatorial Group was able to anticipate a characteristic of the scanned material that could prove inherently problematic, which is that the written texts associated with images in the collection do not describe those images, but rather what those images depict. In other words, specifically in the context of Bosley's book *Greene & Greene*, the images from Avery Library are most frequently *drawings of buildings* (plans, elevations, sections, and so forth), whereas Bosley's written text describes *the buildings themselves*, but does not often comment on the architectural drawings as such. Thus on occasion the data requires a kind of mental leap, whereby the written material refers to the building project generally, rather than to any particular image or plan of it. At the same time, this divergence between the image content and the content of the associated text that CLiMB proposes to mine for metadata will inevitably exist to some degree for most image collections. For instance, a photograph of a South Asian Temple or a scan of a Chinese Paper God is not the temple or paper god itself, but an image of it. More frequently than not, associated text will describe the object, rather than the image.

1.3 Understanding the Collections – the Target Object Identifier (TOI) Concept

Because associated text often describes specific art objects, as opposed to particular images or related items in a collection, the Curatorial and Technical Groups had to develop a method for identifying what an image or a section of text is about. Catalog records for images are structured around the names of the objects those images depict. These authoritative names generally serve as the key to a database entry for the image or collection of images. In order to mine text for robust metadata, which can then be automatically loaded into catalog records or employed by end-users, we must be able to confidently identify what object is being discussed in particular sections of the text. Because the CLiMB project employs collections of art or architecture related images, we have designated these object names and their variants as TOIs, or Target Object Identifiers. The development of this concept has been essential for much of the recent work of CLiMB's Technical and Curatorial Groups. For the former group, as this report will discuss below in sections 2.3 and 2.41, it has influenced many of the advances that have been made in developing the CLiMB toolset, and provides the foundation for CLiMB's first user interface, the TOI Finder. For the latter group, the TOI has been instrumental to the ongoing development of workable datasets, as well as for the creation and formatting of electronic catalog records.

In any image collection, TOIs can serve as the key to records associated with a particular image. If we know that a given segment of text is about a particular art object, we can then use the extracted metadata that is about that target object to enhance that object's catalog records, or make the extracted information available to end-users. Before we can do this, however, we have to determine what the TOIs for a given collection are. In some cases, designating certain terms as TOIs is straightforward from a conceptual standpoint, because lists of these objects may already exist, whether in existing catalog records, back-of-book indexes, encyclopedias, or other similar texts.

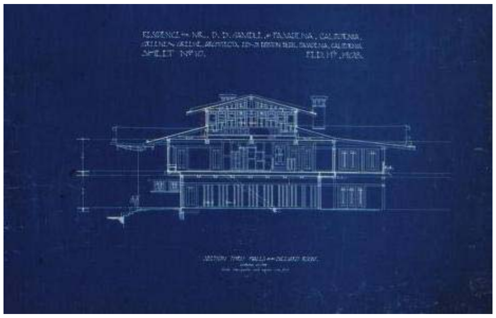
However, the process is complicated by the fact that, across collections, TOIs can differ greatly, because the objects depicted in those collections can be of different types. This is the case for each of the three sets of images the CLiMB project is working with at present. In conjunction with the Technical Group, the Curatorial Group defined the TOIs for these groups of images as follows: for the Greene & Greene collection, the TOIs are *architectural projects*; for the Chinese Paper Gods collection, the TOIs are the *names of the god*s depicted; finally, for the South Asian Temples images, the TOIs are *temple sites*, identified first by geographic location and secondarily by the names of temples themselves, which can be the same in multiple locations. Having established these designations, the Curatorial Group has been able to structure catalog records or create templates for them that include the TOI as a digital library data field. In addition, "authority lists" of TOIs created by the Curatorial Group for a given collection can provide the Technical Group with an important starting point for the process of extracting metadata that will be capable of enhancing a collection's catalog records, or the search capabilities of an image database.

1.4 Working with the Collections

After completing initial preparation for the Greene & Greene dataset, the Curatorial Group began work on two additional collections. The first of these was the Chinese Paper Gods Collection, housed at the C.V. Starr East Asian Library at Columbia University; the second was the South Asian Temples images and metadata, which are part of the Digital South Asia Library, a collaborative project in which Columbia University and the University of Chicago are the lead institutions. While each collection purposefully represents a different kind of material than that found in the images and texts associated with Greene & Greene, they both fulfill the requirements for applicability to the CLiMB project that the Curatorial Group established in its early meetings. Additionally, the primary objects in the former collection (the Chinese Paper Gods) are housed at Columbia, which permits close collaboration with Library staff. By contrast, the latter group of images (South Asian Temples) is already digitized, and thus represents the kind of material to which the CLiMB toolset might eventually be applied when embedded in existing image access platforms.

Because the CLiMB project endeavors to develop tools that will be applicable to a broad range of material, the Curatorial and Technical Groups decided that for each of the CLiMB image collections, at least three different associated texts should be prepared in electronic format for testing with the CLiMB toolset. As was the case with the Greene & Greene Collection, technical work for each of the other two datasets will begin with a single associated authoritative text for initial testing purposes. However, the final group of texts that will be selected for each dataset is designed to be representative of a broader spectrum of possibilities than one text alone could hope to provide. For the Greene & Greene and Chinese Paper Gods collections, a full complement of texts has already been selected; for the South Asian Temples collection, which represents the third to have entered development on the Curatorial Group's staggered timeline, two texts have been selected, one of which will serve as the primary critical work. Complete lists of the relevant texts associated with each collection can be found below in sections 1.41-1.43 and in Appendix B: Collections and Related Material. Also, because one of the major goals of the CLiMB project is to use descriptive metadata to enhance catalog records, a sample image and record from each of the CLiMB collections is provided below in Appendix C: Sample Catalog Records; these examples were taken from the CLiMB website, and can be consulted there at http://www.columbia.edu/cu/cria/climb/collections.html.

1.41 Greene & Greene Architectural Images



Residence for Mr. D.B. Gamble of Pasadena, California
Section through halls and billiard room looking south.
Avery Architectural and Fine Arts Library,
Columbia University Libraries.

From the Sample Images and Metadata available on the CLiMB web pages at:
http://www.columbia.edu/cu/cria/climb/collections-gg.html.

More detail about this image can be found in Appendix C: Sample Catalog Records.

This collection documents the work of the American architects Charles Sumner Greene (1868-1957) and his brother, Henry Mather Greene (1870-1954). It includes over 4,800 architectural drawings, all of which have been digitized. As this report has indicated above, the Greene & Greene

Architectural Records and Papers Collection, housed in the Avery Architectural and Fine Arts Library at Columbia University, was the first of CLiMB's three collections to be prepared and formatted for software testing purposes.

The Greene & Greene architectural drawings are cataloged at the project level in 253 MARC records. (MARC, or MAchine Readable Cataloging, is the bibliographic data structure standard developed by the Library of Congress.) Using project level cataloging means that a single catalog record was created for every architectural project in the collection. The project level records contain constituent unit entries or item level information for all drawings associated with a particular project. For instance, all items in the collection relating to the "William R. Thorsen house" will be cataloged under that particular project. In this case, the project names are the TOIs for the collection.

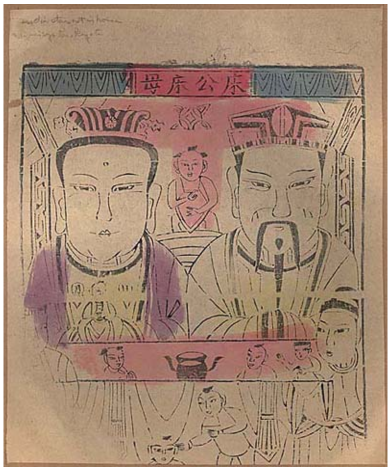
Using these MARC records as a starting point, the Curatorial Group developed specifications for the creation of CLiMB descriptive metadata records for the collection. The group identified which of the MARC fields were relevant to the CLiMB project, and plans to extract information for as many of the 4800 drawings as possible. At the project level, the following fields will be harvested from the existing metadata: personal/corporate name, title, date, physical description, notes, provenance, geographic name, collection title, and repository. Fields harvested at the item level include: accession number (a unique identifier for each drawing), personal/corporate name, title, date, physical description, and drawing number. Additional fields will be included at the project and item levels for CLiMB generated subject terms and keywords. These records could be displayed in Columbia's online catalog (CLIO), Columbia's Master Metadata File (MMF), and in the Luna Insight image platform.

The Curatorial Group has already been able to employ this catalog data to generate an authority list for Greene & Greene projects. As this report mentioned above in section 1.42, the authority list served as the basis for the project names that have been employed for testing the CLiMB toolset. The list developed by the Curatorial Group serves as the master list for Target Object Identifiers (TOIs) for work with the Greene & Greene Collection and texts.

The first text to be scanned and prepared for use in testing metadata extraction capabilities was Chapter 5 of Edward R. Bosley's *Greene & Greene*. Since that time, the remaining texts that will be employed for this dataset have also been digitized. The complete list of texts for this collection is as follows:

1. Bosley, Edward R. *Greene & Greene*. London: Phaidon, 2000.
2. Current, William R. *Greene & Greene: Architects in the Residential Style*. Fort Worth, TX: Amon Carter Museum of Western Art, 1974.
3. Makinson, Randell. *Greene & Greene: Architecture as a Fine Art*. Salt Lake City: Pergrine Smith, 1977.
4. Makinson, Randell. *Greene & Greene: The Passion and the Legacy*. Salt Lake City: Gibbs Smith, 1998.
5. Smith, Bruce. *Greene & Greene Masterworks*. San Francisco: Chronicle Books, 1998.
6. Strand, Janann. *A Greene & Greene Guide*. Pasadena, Calif.: G. Dahlstrom, 1974.

1.42 Chinese Paper Gods Collection



Chuang gong chuang mu
Wood-engraving, color.
C.V. Starr East Asian Library,
Columbia University Libraries.

From the Sample Images and Metadata available on the CLiMB web pages at:
http://www.columbia.edu/cu/cria/climb/collections-cpg.html.
More detail about this image can be found in Appendix C: Sample Catalog Records.

The Anne S. Goodrich Chinese Paper Gods Collection, housed in the Starr East Asian Library at Columbia University, was the second to enter development following the staggered timeline set for the three CLiMB collections. The collection consists of woodblock prints on paper that represent a great variety of Chinese deities, deriving from an incredibly diverse set of cultural sources. Prints of this kind have been produced and used for many centuries in many parts of China; today they are of particular interest both as art and as sources of information in a variety of fields, such as sociology, anthropology, literature, and religion.

The Columbia University Libraries Conservation Laboratory is in the process of conserving approximately 110 prints from the Goodrich collection from among a total of 239 paper gods, using techniques specifically developed for these prints by a paper conservator. The images are then being digitized to provide broader access to the materials and reduce physical handling. In conjunction with this conservation work, the CLiMB Curatorial Group has produced a template for minimal level MARC catalog records for the prints in the collection. Unlike the drawings in the Greene & Greene collection, no previous cataloging had been done for the prints. To date, thirty-two records have been created; these are currently available in the RLG Union Catalog. The MARC fields selected for these records provide basic identifying information for each print, such as title (usually supplied by the cataloger), date, physical description, collection name, and repository. For this collection, TOIs will be based upon the name of the depicted deity, which is established using *pinyin* and Wade-Giles transliteration.

As in the work on the Greene & Greene collection, which thus far has focused on Edward R. Bosley's book, technical work on metadata extraction for the Chinese Paper Gods collection will begin with a single associated text. In this case the selection was a monograph by the collection's donor, Anne S. Goodrich, entitled *Peking Paper Gods*. Goodrich began her collection in 1931, and sixty years later published her book, which is now the standard reference on the subject. This title was selected specifically because it is based upon the collection of images held at the Starr East Asian Library. The text of Goodrich's book has been scanned for technical use, as have two other key monographs that are related to the materials in the collection. The complete list of associated texts for this collection is, at present:

1. Day, Clarence Burton. *Chinese Peasant Cults: Being a Study of Chinese Paper Gods.* Taipei: Ch'eng Wen Publishing Co., 1974.
2. Goodrich, Anne Swann. *Peking Paper Gods: A Look at Home Worship.* Nettetal: Steyler Verlag, 1991.
3. Laing, Ellen Johnston. *Art and Aesthetics in Chinese Popular Prints: Selections from the Muban Foundation Collection.* Ann Arbor, MI: Center for Chinese Studies, University of Michigan, 2002.

1.43 South Asian Temples Images and Metadata

Sun temple - General view (Natamandir in foreground)
Location Konarak, Puri, Orissa, India.
From the photo archives of the American Institute of Indian Studies.

From the Sample Images and Metadata available on the CLiMB web pages at: http://www.columbia.edu/cu/cria/climb/collections-sat.html.
More detail about this image can be found in Appendix C: Sample Catalog Records.

The South Asian Temples collection will be the third to be developed by the Curatorial Group. The images for this dataset are from the American Institute of Indian Studies (AIIS) photographic archive, which is available on-line as part of the Digital South Asia Library project at http://dsal.uchicago.edu/images/aiis/. Unlike the other two collections, in this case both images and minimal metadata already exist in electronic format. CLiMB recently acquired access to the AIIS data in its raw form, which is being sent to CLiMB as of the date of this report. Also, CLiMB has been offered back-end read-only access to the archive itself, should it prove useful to the project. The Curatorial Group may also obtain access to the University of Pennsylvania's Penn/AIIS South Asia Art Archive, a related collection that is available at http://imagesvr.library.upenn.edu/a/aiis/. This database represents a subset of the holdings in the Penn Library's South Asia Collection.

Thus far the Curatorial Group has selected two of the associated texts that will be employed for work on this collection. As with the previous two collections, the initial work on the South Asian Temples will begin with a single text, which in this case is the *Encyclopaedia of Indian Temple Architecture*. The *Encyclopaedia* is a multi-volume reference work that, as for Goodrich's *Peking Paper Gods*, relates intimately to the images in the collection. The *Encyclopaedia* employs the images that appear in the AIIS photographic archive. Though the Curatorial Group had been hopeful that it would be possible to receive the electronic files from which the *Encyclopaedia* was printed, these appear to have been lost abroad. As a result, the group decided to scan relevant portions of text relating to an important group of temple sites. Because the photographic archive contains over

125,000 images and the *Encyclopaedia* consists of two volumes in fourteen parts of text and plates, it had already been deemed necessary to reduce the size of the initial working dataset. The group has made arrangements to consult with several scholars in this field to determine which temple sites are the most important from the point of view of academic research and cultural significance. On this basis of this information, it will be possible both to select additional monographs for the collection, as well as proceed with developing the image and text dataset. The information on the texts that have been selected already is as follows:

1. *Encyclopaedia of Indian Temple Architecture*. New Delhi: American Institute of Indian Studies; Philadelphia: University of Pennsylvania Press, 1983- .
2. *Architectural Survey of Temples*. New Delhi: Archaeological Survey of India, 1964- .

2. CLiMB Tools

2.1 Designing System Architecture

During the first year of the CLiMB project, the Technical Group developed an architecture for the CLiMB suite of tools. Figure 1A presents a general overview of project design, which can be

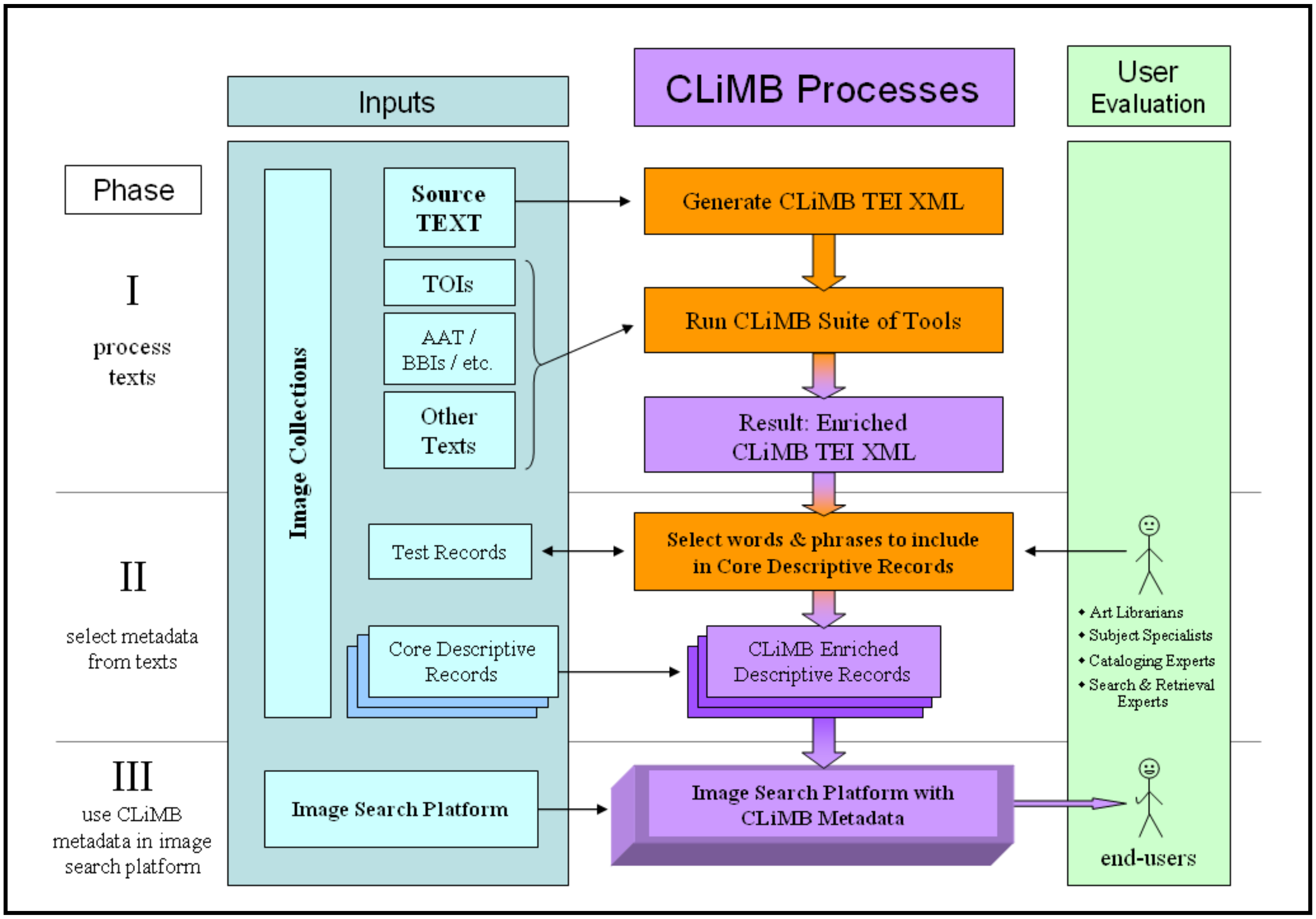


Figure 1A: General Overview of CLiMB Project Design

conceptualized in three phases. Each part is then elaborated in the subsequent detailed diagrams, which represent the current thinking for each particular phase (Figures 1B, 1C, and 1D). As the project evolves, these flowcharts may also evolve to reflect research results. The center section of each chart shows CLiMB processes and results in purple and orange boxes; the left side of each diagram, in blue, shows the "Inputs" to the system, while the right side, in green, shows user evaluation and involvement in the overall process.

Phase I, "process texts," describes the sequence by which the Technical Group converts texts associated with image collections into a manipulable format – "CLiMB TEI XML" – and then runs the CLiMB toolset over those texts. TEI XML is a standard markup language that has a document type description (or DTD) that is compliant with the guidelines set out by the Text Encoding Initiative Consortium (http://www.tei-c.org). CLiMB TEI XML has overlap with the current TEI DTD, and in the future we plan to make all CLiMB data entirely TEI compliant insofar as it is possible to do so. The first step in the CLiMB project architecture involves taking text that is already in TEI XML format, and, as Figure 1B shows, marking it for noun phrases using LTChunk, a Noun Phrase chunker (Finch and Mikheev, 1997; http://www.ltg.ed.ac.uk/software/chunk/index.html). The decision to use LT Chunk for this step was the result of an extensive set of tests made with several NP finders and similar tools, all of which are discussed in detail in section 2.2. LTChunk identifies a large number of noun phrases that can potentially be employed as metadata.

Once the text has been given preliminary markup as CLiMB XML, it is run through the CLiMB suite of tools. This suite employs a set of independent modules, which allow each particular

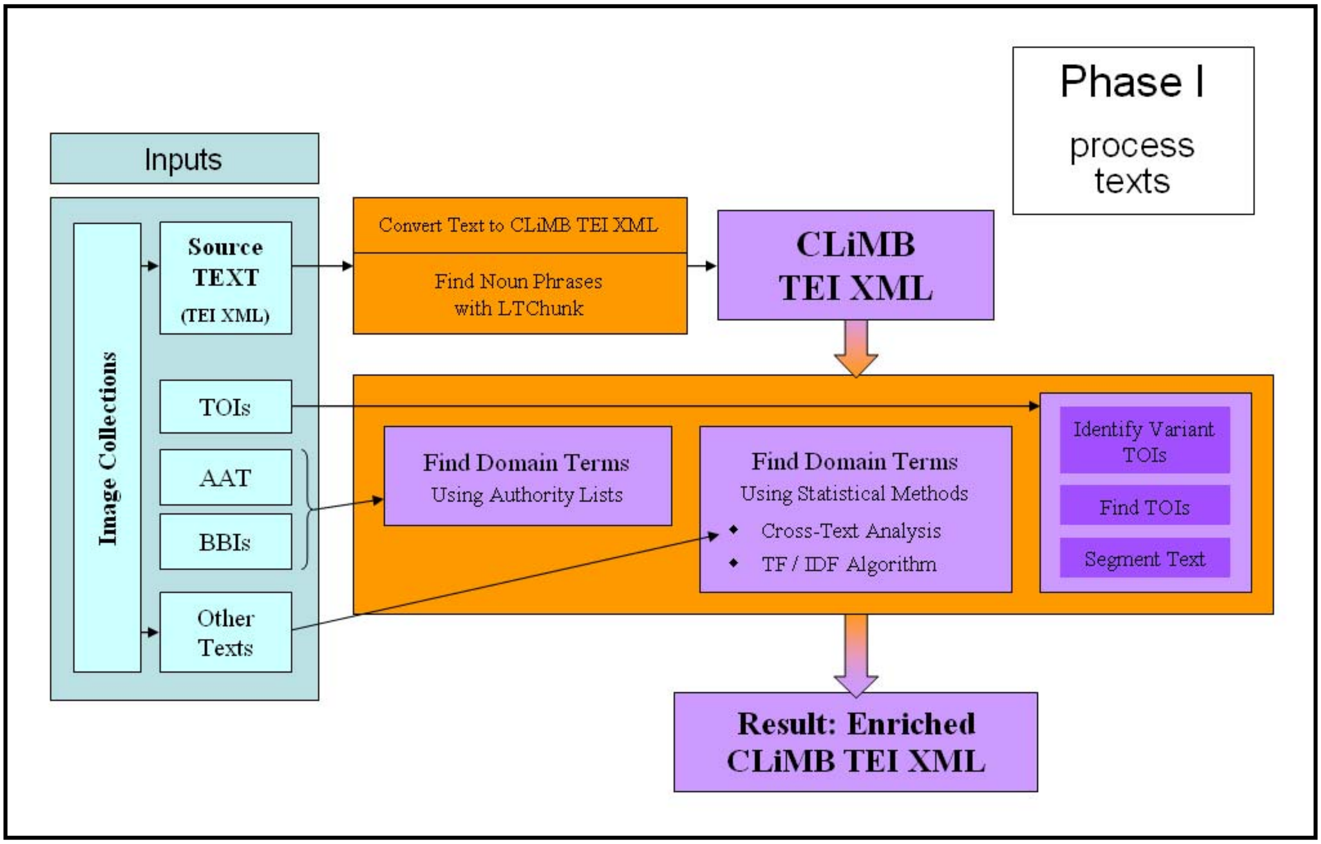


Figure 1B: Processing Texts (Phase I)

component of the toolset to be designed specifically for its purpose. These modules run parallel to one another, as opposed to sequentially. Three modules are depicted in the large orange box in the vertical center of the diagram. Each takes some form of input from the left side of the diagram and employs it in further markup of the CLiMB XML. One module, for instance, will "Find Domain Terms Using Authority Lists" derived from structured vocabularies such as the Getty *Art and Architecture Thesaurus* (AAT); in the future this will be extended to include back-of-book indexes (BBIs). Another will "Find Domain Terms Using Statistical Methods" with techniques such the Term Frequency / Inverse Document Frequency algorithm (TF/IDF) and variants (Salton, 1971), or by performing cross-text analysis by comparing several texts associated with a collection of images. Finally, another will take Target Object Identifiers (TOIs), automatically derive their variants, locate them in the text, and then segment the text into topical sections relating to each Identifier. The flexible design of the toolset allows new features to be added at any time as subsequent modules; it also makes possible the adjustment of each particular component without affecting any others. The result of this phase is "Enriched CLiMB TEI XML," which is text that has been marked up by each part of the toolset.

Phase II, "select metadata from texts," is shown in Figure 1C. This is the stage in which the data collected in the first phase is applied to catalog records for image collections. The orange box in the center of the diagram shows the process of selecting which text to use for this purpose. A specialized group of technical experts – art librarians, subject specialists, cataloging experts, and

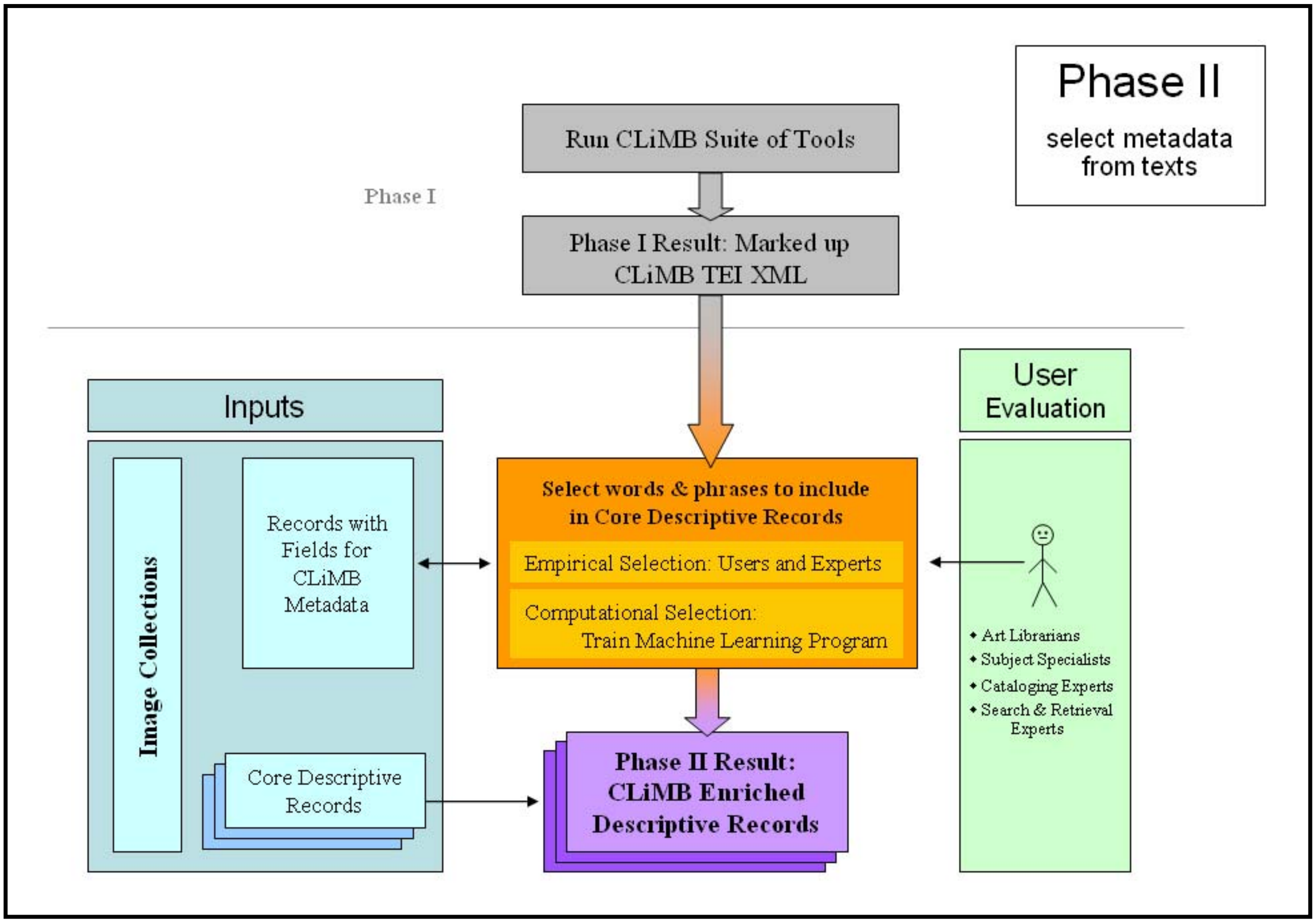

Phase II
select metadata from texts
Run CLiMB Suite of Tools
Phase I
Phase I Result: Marked up CLiMB TEI XML
Inputs
Image Collections
Records with Fields for CLiMB Metadata
Core Descriptive Records
Select words & phrases to include in Core Descriptive Records
Empirical Selection: Users and Experts
Computational Selection: Train Machine Learning Program
Phase II Result: CLiMB Enriched Descriptive Records
User Evaluation
• Art Librarians
• Subject Specialists
• Cataloging Experts
• Search & Retrieval Experts

Figure 1C: Build Metadata Records with CLiMB-derived Terms (Phase II)

access experts – provide one way of indicating what parts of the marked up information are valuable as metadata; in conjunction with this "Empirical Selection," "Computational Selection" can take place using a machine learning program, such as Ripper (Cohen 1996). Programs like this one can be trained to identify what qualifies to be extracted as metadata. During this part of the process, the extracted text will be applied to catalog records that include fields designed for CLiMB metadata, such as those described in section 1.41 above. When the selection process has been refined to the point at which it can be employed for records with an adequate level of confidence, harvested metadata can be inserted into "Core Descriptive Records," which are sets of multiple catalog records that have fields for CLiMB Metadata. The result will be "CLiMB Enriched Descriptive Records," which can be inserted into existing image search platforms.

Phase III will "use CLiMB metadata in an image search platform" in the form of the descriptive records resulting from the previous phase. The first two phases are also shown in Figure 1D in order to emphasize the potential for continual adjustment of the CLiMB toolset based upon users' needs. The evaluation loop allows the suite of tools to be modified at the crucial point in each of the two previous phases: running the toolset for markup and selecting metadata for insertion into records. This process is implicit in the system architecture. If, for instance, we learn from users that Phase I has not identified all of the potentially useful words and phrases as candidates for metadata,

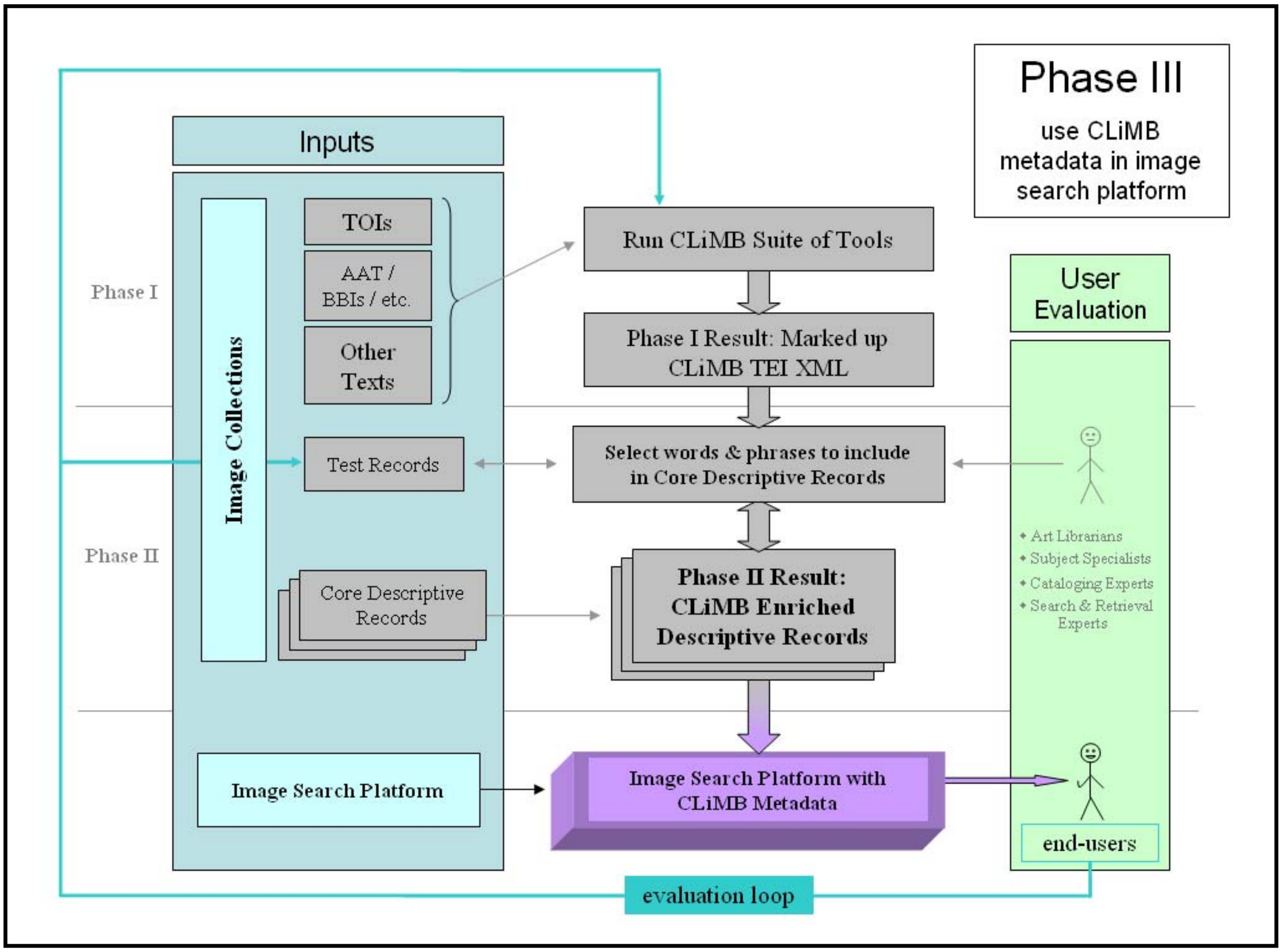

Phase III
use CLiMB metadata in image search platform
Inputs
Image Collections
TOIs
AAT / BBIs / etc.
Other Texts
Run CLiMB Suite of Tools
Phase I Result: Marked up CLiMB TEI XML
User Evaluation
Phase I
Test Records
Select words & phrases to include in Core Descriptive Records
Phase II
Core Descriptive Records
Phase II Result: CLiMB Enriched Descriptive Records
• Art Librarians
• Subject Specialists
• Cataloging Experts
• Search & Retrieval Experts
Image Search Platform
Image Search Platform with CLiMB Metadata
end-users
evaluation loop

Figure 1D: Testing CLiMB Metadata for Image Access (Phase III)

modules can be adjusted or added to the suite accordingly; likewise, if we learn that we have identified the proper candidates for metadata, but have not selected them properly in Phase II, the machine learning program can be retrained.

The thinking behind the first phase of the CLiMB project, as well as the steps that were taken to arrive at a substantial picture of this phase, are discussed in detail throughout the rest of section 2 of this report. The subsequent phases are addressed in sections 3 and 5.

2.2 Choosing and Testing Pre-existing Tools

The process of extracting metadata from texts by employing software tools designed for computational linguistics begins with the assumption that many of the terms we hope to find in a given text will be either simple nouns or more complex noun phrases. Because noun detection is not a new endeavor for computational linguistics, the CLiMB Technical Group began work by testing software tools that already exist: POS (Part of Speech) taggers and NP (Noun Phrase) chunkers. As their generic names indicate, POS taggers are software programs that determine the part of speech of each word in a written text or transcript, whereas NP chunkers combine nouns into complex noun phrases. Locating particular kinds of words, rather than searching for specific predetermined terms (which resembles the search function of a word processing application), was the first step in developing techniques for automatically extracting relevant metadata from text.

The CLiMB Technical Group thus began by testing three software tools: Alembic Workbench version 2.8, a POS tagger made by Mitre (Day et al, 1997; http://www.mitre.org/); LinkIT, an NP chunker created by Columbia University's NLP Group (Evans, Klavans, & Wacholder 2000; Wacholder 1998); and LTChunk, an NP chunker created by the University of Edinburgh Language Technology Group (Finch and Mikheev, 1997). The software was applied to Chapter 5 of Edward R. Bosley's book *Greene & Greene*, which documents the work of Pasadena architects Henry Mather Greene (1870-1954) and Charles Sumner Greene (1868-1957). (The process that led to the selection of Greene & Greene materials for the initial dataset is described in detail above in the first part of this report.) The text of Chapter 5 spans 47 paragraphs and contains close to 11,500 words; its content is primarily devoted to detailed descriptions of houses designed during the "most demanding phase" of the architects' careers.

For testing the software on this initial dataset, the group measured the performance of each of the three tools against provisional "gold standards," which were lists of terms directly related to the Greenes' projects and to architecture-specific terminology. The initial tests revealed that Alembic Workbench (AWB) performed better than either LinkIT or LTChunk at identifying relevant *proper* nouns (which are referred to in the computational linguistics literature as "named entities") with a high level of precision. On the basis of these preliminary results, the Group decided that AWB would provide the best initial basis for CLiMB's metadata extraction.

This judgment was based upon the early assumption that the most relevant terms for extraction were proper noun phrases, which AWB located well. Proper nouns are often extremely valuable as metadata because they can describe something important about a given image. In the initial Greene & Greene dataset, for example, the city Pasadena, the name Greene, or the name of a person for whom an architectural project was undertaken would likely be valuable. On the other hand, not all proper nouns make for constructive metadata, and in a given text not all metadata are limited to terms that are proper nouns.

Because terms such as these are not readily identified by AWB, the Technical Group expanded its preliminary assumptions and returned to using LTChunk for testing purposes, eventually choosing this tool over AWB as the first component of CLiMB's suite of tools. AWB has a higher level of *precision*, because the terms it identifies are likely to be proper nouns, at the cost of a certain amount of comprehensiveness in identifying all possible terms; LTChunk has a higher level of *recall* across a wider range of noun phrases, which means that it is more likely to identify more of the noun phrases, at the cost of a certain amount of exactness. Thus, although LTChunk identifies a higher percentage of less relevant terms, it is better at identifying a wider range of them overall.

The "wide net" cast by this kind of processing is at the foundation of the CLiMB software suite, which now has a user interface developed by the CLiMB Technical Group at the end of the first year of the project. Beginning with LTChunk's broad capabilities, the software behind the interface identifies matches to noun phrases, terms from the Getty *Art and Architecture Thesaurus*, and TOIs. The CLiMB-designed software automatically generates a list of variants for each TOI, and then disambiguates between choices for less reliable matches. This part of the suite of tools is referred to as the TOI Finder; the processes behind it are discussed below in sections 2.3, 2.4, and 2.41.

The different capabilities of AWB and LTChunk point to an important characteristic of the datasets themselves, which is that the terms they contain can be either *project specific* (for example, the name of a house designed by Greene & Greene), or *domain specific* (for example, the name of a particular aspect of an architectural project, such as an architectural style or the term for an element of a building, such as "sleeping porches"). Recognizing the distinction between domain specific terms and project specific terms represents a valuable insight that can eventually be addressed by the user evaluation component of the CLiMB project. At present, the Technical Group has mapped strategies that will allow the suite of tools to locate *either* sort of term, depending upon what users decide is valuable. As section 2.1 explained, the system architecture has been designed to allow for refining the software at many points. The thinking of the Technical Group is that all possible candidates for metadata should be identified in the first part of the process, or Phase I. The foundation for locating these terms in text is provided by LTChunk, and then elaborated and made more specific by subsequent modules in the suite. At that point, the relative significance of particular terms—i.e., those that would make for valuable metadata—can be determined by users and machine learning software to weigh additional selection factors, in Phase II.

2.3 Understanding the Datasets – the Target Object Identifier (TOI) Concept

In order to develop a set of software tools that will be able to extract relevant metadata from text associated with images, it is crucial to be able to identify which segments or parts of a given text can be confidently classified as being *about* the object shown in a given image. As this report mentioned above, catalog records for the images in CLiMB's collections can be structured around the names of the objects those images depict. We have termed these names Target Object Identifiers, or TOIs; these identifiers can serve as the key to catalog records associated with particular images. Thus, before text can be mined for metadata that can be put to use in catalog records or employed by end-users, CLiMB must necessarily develop a way to confidently identify which sections of a text are about which particular images – or, in other words, we must be able to identify which sections of text contain references to, or are about, particular TOIs.

The challenge of automatically locating TOIs in text arises from the fact that, for any single target object, references to TOIs can vary greatly within a given text. For instance, in Chapter 5 of Edward R. Bosley's *Greene & Greene*, which was discussed in section 1.2 above, the William R. Thorsen house, one of the Greenes' projects, might be referred to as "the Thorsen house," "the William Thorsen house," or by a less direct referential term such as "the house."

The catalogers for the collection of Greene & Greene architectural drawings at the Avery Architectural and Fine Arts Library supply the comprehensive and clear-cut "William R. Thorsen house (*Berkeley, Calif.*)" as the term for identifying this particular object. This subject term, however, does not appear in the text of Bosley's chapter; searching for it would not return any matches. For this reason, it is necessary to develop a method for identifying not only the obvious references to a given target object in a text, but also the less obvious ones, which tend to fluctuate significantly. To address this issue, the Technical Group created a method for locating these types of variants in Chapter 5 of Bosley's *Greene & Greene*. Starting with an "authority list" of TOIs for Greene & Greene architectural projects, which was developed by the Curatorial Group, the Technical Group crated a software tool, the TOI Finder, that first "decays" the TOI into a list of possible variants, and then searches for them, assigning different levels of confidence to individual matches. This process represents one module of the CLiMB toolset; it is elaborated below in section 2.41. Since its first implementation, an additional module, which identifies terms appearing in the Getty *Art and Architecture Thesaurus*, has been added to the user interface for the toolset, as section 2.42 indicates. Finally, segmentation techniques, which are discussed below in section 2.43, will likely employ TOIs in another module to identify topical boundaries in associated text.

A particularly interesting situation arises for each of CLiMB's other two datasets, the Chinese Paper Gods Collection and the South Asian Temples images and metadata. In both cases, the name of a given target object can exist in several forms as a result of translation, different conventions for transliteration, or variations in dialect. This is particularly true of the Chinese Paper Gods, which can have as many as twenty-five different names for any one god. Future technical applications of the TOI concept will include an inquiry into searching for TOIs across languages and different orthographic and transliteration renderings.

2.4 Creating Original Algorithms and Refining the Toolset

The most significant recent accomplishment for the Technical Group was the creation of the interface that allows users to locate occurrences of TOIs (Target Object Identifiers) in text. This tool is called the TOI Finder, and is currently available on-line at CLiMB's "Tools and Prototypes" page, http://www.columbia.edu/cu/cria/climb/tools.html. Clicking the link for the "Current TOI Finder Prototype" will open the interface. It allows users to provide their own target object names, and a text over which to search for them; the results of the search are then returned to the user as a visual display highlighting TOIs, including variants of the object names provided by the user, that the software located in the text. Also during the first year of the CLiMB project, the Technical Group began to develop techniques for expanding upon the capabilities of LTChunk, which is the foundation of the CLiMB suite of tools. These developments have thus far taken place in two main areas, which have run parallel to one another, as the overall plan for the suite allows. The first area the Group focused on was the use of subject-oriented (or "domain") vocabularies derived from existing authority lists, which identify terms specific to a given discipline or domain to improve metadata extraction results. The second technique the Group considered was segmentation, which is a computational method for dividing texts into discrete topical segments pertaining to specific themes.

2.41 Identifying Target Objects Automatically

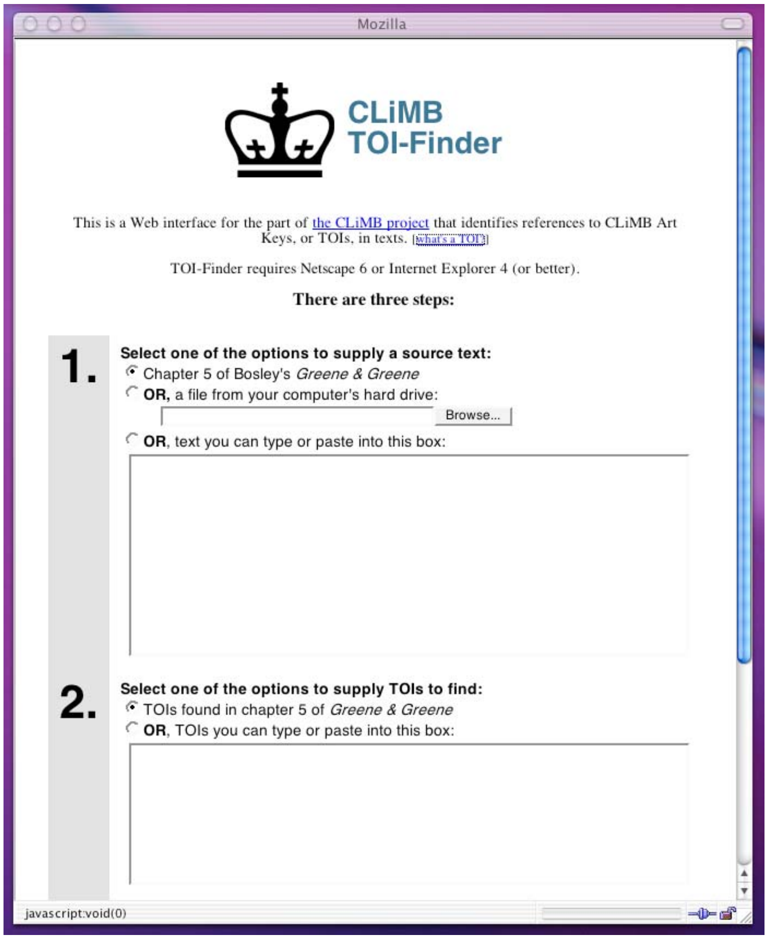
As this report has explained in sections 1.3 and 2.3, a crucial step towards the goal of extracting robust metadata about particular target objects from associated text is being able to determine which parts of a given text refer to a particular object. In many situations, the names for target objects can be provided by users, catalogers, and, at times, existing authority lists. The interesting challenge in applying these names to associated text, however, is that the references to the target object itself are made with any number of possible variants, which are sometimes shortened versions of a proper name (for instance, the title of an artwork or architectural project), or even more typically, less specific referential terms (common nouns or pronouns referring to a title, a project, or the like). In order to address this problem, the CLiMB Technical group developed the TOI Finder, the major purpose of which is to locate variants for the names of an art or architectural object. It is able to take target object names and associated text supplied by a user, and then locate TOIs (Target Object Identifiers), which include the target object name and its variants, in the given text. The TOI Finder will also run Chapter 5 of Bosley's *Greene & Greene* as a default text, supplying the user with TOIs referring to Greene & Greene projects.

A paper describing this tool in detail was recently presented at this year's Joint Conference on Digital Libraries (JCDL); it was authored by Peter Davis, David Elson, and Judith Klavans, and entitled "Methods for precise named entity matching in digital collections." The interface for this tool, which is shown below in Figures 2A-2D, was first publicly presented in April of 2003 as a demo for a meeting at the Coalition for Networked Information (CNI). Both the demo and the paper itself are available via a link on the CLiMB "Publications and Presentations" page at http://www.columbia.edu/cu/cria/climb/presentations.html.

The TOI Finder locates TOIs by taking the target object name or names supplied by a user and repeatedly "decaying" them into more general variants. Returning to the example discussed in section 2.3, the subject heading used for one of the Greene & Greene architectural projects is "William R. Thorsen house (*Berkeley, Calif.*)." The process of decay sequentially removes modifiers, yielding a progressively more general set of terms, so that the subject heading above provides variants such as "The William R. Thorsen house," "William Thorsen house", "Thorsen house", and "the house" (a definite article is automatically added to the last of these to ensure that occurrences of the variant are mentions of a specific project). The TOI Finder will then run LTChunk (Finch and Mikheev, 1997) over the text, locating all noun phrases, as well as this automatically generated list of variants.

Most of the texts that can be associated with images in collections, however, discuss more than a single target object at once. Chapter 5 of Bosley's *Greene & Greene*, for example, considers several of the projects undertaken by the Greenes during a particularly active period of their careers. As the variants created by the TOI Finder become more general, they also become more ambiguous, which means that they are less likely to refer to a particular project – in technical terms, the *recall* for such variants is high, but the *precision* with which matches refer to a particular project is low. Thus, occurrences of "Thorsen house" tend to refer to "William R. Thorsen house (*Berkeley, Calif.*)" with a high level of *precision*, but *recall* is low because those occurrences do not represent a large percentage of the total number of references the chapter makes to this project; on the other hand, matches for "the house" demonstrate less *precision* in that they are less likely to refer to the particular project – Bosley speaks about more than one "house" in the chapter – but *recall* is increased because the likelihood of having located a higher percentage of all possible references to the project is greater.

To achieve a balance between precision and recall, the algorithm behind the TOI Finder was designed to use matches with high precision but low recall as "seeds" for disambiguating matches with low precision and high recall. This approach uses more specific matches to determine what the less specific ones refer to. It is reasonable to assume, for example, that when an occurrence of "the house" follows closely upon an occurrence of "Thorsen house," the more ambiguous term is likely to be related to the more precise one. Refining the algorithm such that it could employ the more precise matches as "seeds" in this way improved the TOI Finder's overall performance substantially by increasing a combined score for precision and recall.



This is a Web interface for the part of the CLiMB project that identifies references to CLiMB Art Keys, or TOIs, in texts. [what's a TOI?]

TOI-Finder requires Netscape 6 or Internet Explorer 4 (or better).

There are three steps:

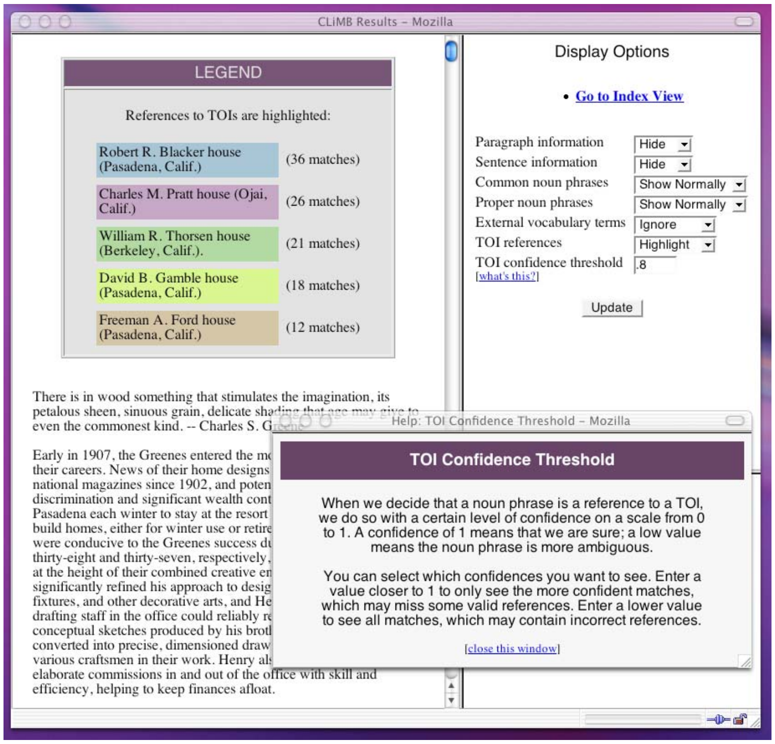
1. **Select one of the options to supply a source text:**
   ● Chapter 5 of Bosley's *Greene & Greene*
   ○ **OR,** a file from your computer's hard drive:
   ○ **OR**, text you can type or paste into this box:

2. **Select one of the options to supply TOIs to find:**
   ● TOIs found in chapter 5 of *Greene & Greene*
   ○ **OR**, TOIs you can type or paste into this box:

Figure 2A: TOI-Finder Input Page

The screen shot in Figure 2A shows the starting point of the TOI Finder. The user has chosen to run the finder over text from Chapter 5 of *Greene & Greene*, the option that has been selected at step 1; the target objects, or TOIs, for five of the Greenes' projects have been selected at step 2. Though Chapter 5 of *Greene & Greene* and the related TOIs may be used as a default, as in this example, the TOI Finder can run on over any text supplied by the user, whether it is entered directly into the text field at step 1 or taken from a file on the user's hard drive. The user may then enter TOI names of his or her choosing into the text field for step 2. After clicking the "submit" button for step 3, which is off the bottom in the screen in this example, the results are returned on a separate page.

The top of the results page for this particular search, shown in Figure 2B, provides the user with a "legend" to identify the TOIs that the Finder has highlighted in the text below. This image also shows the "what's this?" pop-up window for the "Confidence Threshold," which is a score between 0 and 1 that refers to the level of confidence the TOI Finder has that a particular match in the text refers to a particular TOI. A user can adjust the Confidence Threshold in the last of the Display Options in the top-right frame. The default threshold is 0.8. Selecting a threshold of 1.0 will return only the most confident matches, whereas entering a lower number will return more matches, some of which may be incorrect. In the technical terms introduced in the sections above, a high confidence threshold will return matches with high *precision* but low *recall*; a lower confidence threshold increases *recall* while sacrificing a certain amount of *precision*.



LEGEND

References to TOIs are highlighted:

Robert R. Blacker house (Pasadena, Calif.)	(36 matches)
Charles M. Pratt house (Ojai, Calif.)	(26 matches)
William R. Thorsen house (Berkeley, Calif.).	(21 matches)
David B. Gamble house (Pasadena, Calif.)	(18 matches)
Freeman A. Ford house (Pasadena, Calif.)	(12 matches)

There is in wood something that stimulates the imagination, its
petalous sheen, sinuous grain, delicate sha
even the commonest kind. -- Charles S. G

Early in 1907, the Greenes entered the m
their careers. News of their home designs
national magazines since 1902, and poten
discrimination and significant wealth cont
Pasadena each winter to stay at the resort
build homes, either for winter use or retire
were conducive to the Greenes success d
thirty-eight and thirty-seven, respectively,
at the height of their combined creative en
significantly refined his approach to desig
fixtures, and other decorative arts, and He
drafting staff in the office could reliably r
conceptual sketches produced by his brot
converted into precise, dimensioned draw
various craftsmen in their work. Henry al
elaborate commissions in and out of the office with skill and
efficiency, helping to keep finances afloat.

Display Options

- Go to Index View

Paragraph information	Hide
Sentence information	Hide
Common noun phrases	Show Normally
Proper noun phrases	Show Normally
External vocabulary terms	Ignore
TOI references	Highlight
TOI confidence threshold [what's this?]	.8

Update

TOI Confidence Threshold

When we decide that a noun phrase is a reference to a TOI, we do so with a certain level of confidence on a scale from 0 to 1. A confidence of 1 means that we are sure; a low value means the noun phrase is more ambiguous.

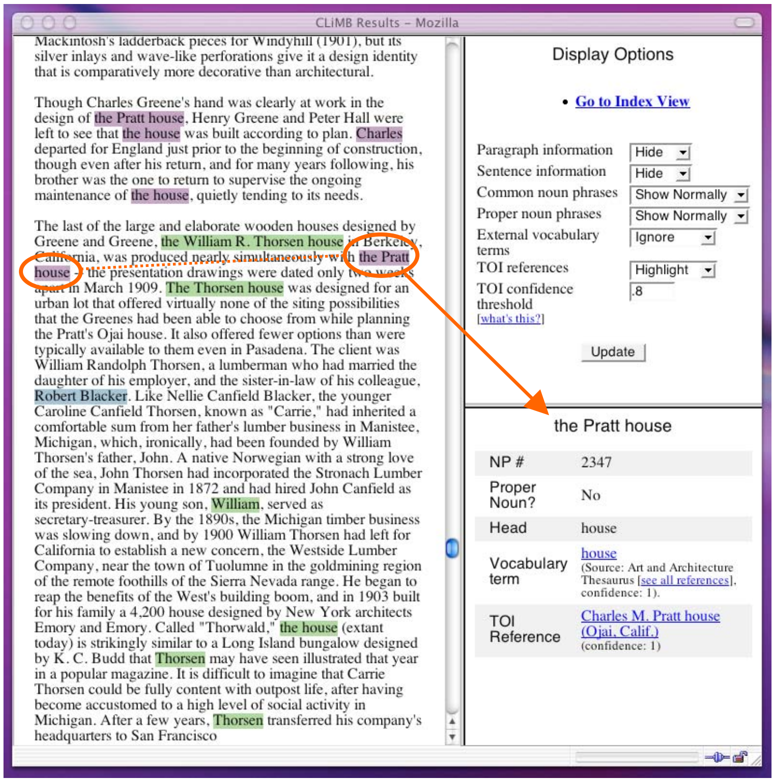
You can select which confidences you want to see. Enter a value closer to 1 to only see the more confident matches, which may miss some valid references. Enter a lower value to see all matches, which may contain incorrect references.

[close this window]

Figure 2B: Top of TOI-Finder Results Page

Legend and Confidence Threshold

The display assigns a color to each TOI, which is then used to highlight each match in the text. This is shown in Figure 2C. By clicking on any given match, information about that item will appear in the "Detail Panel," which is the lower frame on the right; in this example, the match selected is the second highlighted occurrence of "the Pratt house," here circled in orange. The TOI Finder has tagged the phrase as a reference to the TOI "Charles M. Pratt house (Ojai, Calif.)" with a confidence score of 1, which is the maximum.



Mackintosh's ladderback pieces for Windyhill (1901), but its silver inlays and wave-like perforations give it a design identity that is comparatively more decorative than architectural.

Though Charles Greene's hand was clearly at work in the design of the Pratt house, Henry Greene and Peter Hall were left to see that the house was built according to plan. Charles departed for England just prior to the beginning of construction, though even after his return, and for many years following, his brother was the one to return to supervise the ongoing maintenance of the house, quietly tending to its needs.

The last of the large and elaborate wooden houses designed by Greene and Greene, the William R. Thorsen house in Berkeley, California, was produced nearly simultaneously with the Pratt house -- the presentation drawings were dated only two weeks apart in March 1909. The Thorsen house was designed for an urban lot that offered virtually none of the siting possibilities that the Greenes had been able to choose from while planning the Pratt's Ojai house. It also offered fewer options than were typically available to them even in Pasadena. The client was William Randolph Thorsen, a lumberman who had married the daughter of his employer, and the sister-in-law of his colleague, Robert Blacker. Like Nellie Canfield Blacker, the younger Caroline Canfield Thorsen, known as "Carrie," had inherited a comfortable sum from her father's lumber business in Manistee, Michigan, which, ironically, had been founded by William Thorsen's father, John. A native Norwegian with a strong love of the sea, John Thorsen had incorporated the Stronach Lumber Company in Manistee in 1872 and had hired John Canfield as its president. His young son, William, served as secretary-treasurer. By the 1890s, the Michigan timber business was slowing down, and by 1900 William Thorsen had left for California to establish a new concern, the Westside Lumber Company, near the town of Tuolumne in the goldmining region of the remote foothills of the Sierra Nevada range. He began to reap the benefits of the West's building boom, and in 1903 built for his family a 4,200 house designed by New York architects Emory and Emory. Called "Thorwald," the house (extant today) is strikingly similar to a Long Island bungalow designed by K. C. Budd that Thorsen may have seen illustrated that year in a popular magazine. It is difficult to imagine that Carrie Thorsen could be fully content with outpost life, after having become accustomed to a high level of social activity in Michigan. After a few years, Thorsen transferred his company's headquarters to San Francisco

Display Options

- Go to Index View

Paragraph information	Hide
Sentence information	Hide
Common noun phrases	Show Normally
Proper noun phrases	Show Normally
External vocabulary terms	Ignore
TOI references	Highlight
TOI confidence threshold [what's this?]	.8

Update

the Pratt house

NP #	2347
Proper Noun?	No
Head	house
Vocabulary term	house (Source: Art and Architecture Thesaurus [see all references], confidence: 1).
TOI Reference	Charles M. Pratt house (Ojai, Calif.) (confidence: 1)

Figure 2C: Body of TOI-Finder Results Page

Text and Detail Panel

In addition to decaying and identifying TOIs, the suite of tools also locates all proper nouns in the text using LTChunk, and all terms that appear in the Getty *Art and Architecture Thesaurus* ("External vocabulary terms"). The top frame on the right allows a user to show or hide the various marked elements of the text, with different formatting options. These aspects of the software are discussed in detail in the next section.

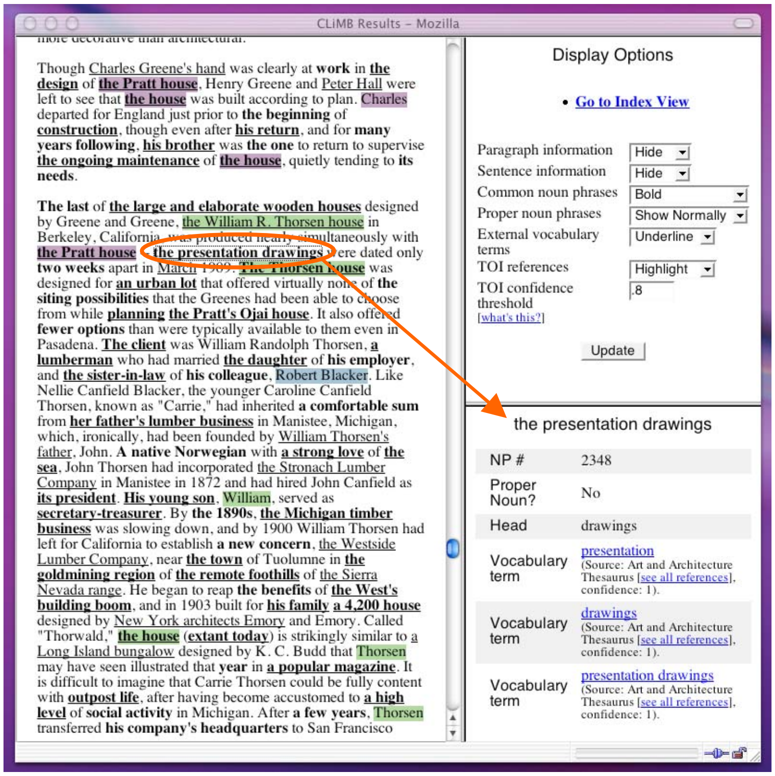
Though the Technical Group has been very encouraged by the results obtained using the TOI Finder, their goal is to increase the tool's precision and recall abilities for locating TOIs to well above the benchmark usually set for computational linguistic research. In addition, the Group has begun to explore methods for modifying the TOI Finder such that it will be able to identify multilingual and transliterated variants. This is particularly important for collections like the two CLiMB is currently developing as part of ongoing research, the Chinese Paper Gods collection and the South Asian Temples images and metadata. For the latter, many of the proper names referring to the objects depicted in the images are terms from Sanskrit, for which there are two standardized transliteration methods. The problem is similar for the former collection, because there are also two methods for transliterating Chinese characters; however, the Chinese Paper Gods collection is arguably even more complex, because the gods depicted in the images are often known by more than one name, each of which might itself exist in different transliterated and translated forms. On at least one occasion, this has yielded over twenty-five different names by which to identify a single god, or target object.

2.42 Using Subject-oriented Vocabulary

Decaying TOIs into variants and locating them in text is an important means of augmenting the accuracy of high-recall, low-precision matches, like those that would be found by a tool such as LTChunk; another method is to employ specific lists of terms that the software can then be made to search for. The kind of searching performed by the TOI Finder is based upon project names, which often already exist, for instance in back-of-book indexes or catalog records. The default search on the TOI Finder employs the five project TOIs that represent the main focus Chapter 5 of Bosley's *Greene & Greene*. These terms, which can be seen at the top of Figure 2B, were taken from the authority list for the collection of Greene & Greene architectural images that was developed by CLiMB's Curatorial Group using information from catalog records. It is possible that similar lists of projects could be established for many image collections. On the other hand, a cataloger or an end-user might wish to locate information and images dealing with a particular architectural component, rather than a specific project. This issue can be addressed by deriving these kinds of terms from pre-existing lists or vocabularies; it may also be possible to generate such vocabularies automatically.

Existing lists for subject-specific (or domain specific) terms can be found in some of the same places that one is likely to find project specific terms: back-of-book indexes potentially refer to both types of information. Thus, one way to develop authority lists would be to identify subject-specific terms in such sources by employing automatic or manual techniques (or a combination of both). A more promising option for establishing such vocabularies exists in subject-specific dictionaries and thesauri. Just prior to the writing of this report, the Curatorial Group secured a site license to the electronic version of the *Art and Architecture Thesaurus* (AAT), a resource published in print and electronic formats by the Getty Trust (http://www.getty.edu/research/tools/vocabulary/aat/). The AAT is a structured vocabulary of terms specific to the fields of art and architecture. The Technical Group recently added the structured vocabulary of the AAT as a feature of the CLiMB suite of tools, and adjusted the user interface in its demo version to include marking AAT terms as a formatting option. On the right side of the results page, a user may choose to display AAT (or

"External vocabulary") terms with bold, underlined, or highlighted text. Figure 2D displays the combined matches for the default Greene & Greene TOIs, which are highlighted in various colors; all common noun phrases located by LTChunk, which are in bold text; and all terms with matches in the AAT, which are underlined. The "Detail Panel" in this figure provides information about the match for "the presentation drawings," which has been tagged as both a noun phrase and as an external vocabulary term. Clicking on the "Vocabulary term" links in the panel will open a separate window showing the term's meaning and location in the AAT hierarchy.



more decorative than architectural.

Though Charles Greene's hand was clearly at work in the design of the Pratt house, Henry Greene and Peter Hall were left to see that the house was built according to plan. Charles departed for England just prior to the beginning of construction, though even after his return, and for many years following, his brother was the one to return to supervise the ongoing maintenance of the house, quietly tending to its needs.

The last of the large and elaborate wooden houses designed by Greene and Greene, the William R. Thorsen house in Berkeley, California, was produced nearly simultaneously with the Pratt house ; the presentation drawings were dated only two weeks apart in March 1909. The Thorsen house was designed for an urban lot that offered virtually none of the siting possibilities that the Greenes had been able to choose from while planning the Pratt's Ojai house. It also offered fewer options than were typically available to them even in Pasadena. The client was William Randolph Thorsen, a lumberman who had married the daughter of his employer, and the sister-in-law of his colleague, Robert Blacker. Like Nellie Canfield Blacker, the younger Caroline Canfield Thorsen, known as "Carrie," had inherited a comfortable sum from her father's lumber business in Manistee, Michigan, which, ironically, had been founded by William Thorsen's father, John. A native Norwegian with a strong love of the sea, John Thorsen had incorporated the Stronach Lumber Company in Manistee in 1872 and had hired John Canfield as its president. His young son, William, served as secretary-treasurer. By the 1890s, the Michigan timber business was slowing down, and by 1900 William Thorsen had left for California to establish a new concern, the Westside Lumber Company, near the town of Tuolumne in the goldmining region of the remote foothills of the Sierra Nevada range. He began to reap the benefits of the West's building boom, and in 1903 built for his family a 4,200 house designed by New York architects Emory and Emory. Called "Thorwald," the house (extant today) is strikingly similar to a Long Island bungalow designed by K. C. Budd that Thorsen may have seen illustrated that year in a popular magazine. It is difficult to imagine that Carrie Thorsen could be fully content with outpost life, after having become accustomed to a high level of social activity in Michigan. After a few years, Thorsen transferred his company's headquarters to San Francisco

Display Options

- Go to Index View

Paragraph information	Hide
Sentence information	Hide
Common noun phrases	Bold
Proper noun phrases	Show Normally
External vocabulary terms	Underline
TOI references	Highlight
TOI confidence threshold [what's this?]	.8

the presentation drawings

NP #	2348
Proper Noun?	No
Head	drawings
Vocabulary term	presentation (Source: Art and Architecture Thesaurus [see all references], confidence: 1).
Vocabulary term	drawings (Source: Art and Architecture Thesaurus [see all references], confidence: 1).
Vocabulary term	presentation drawings (Source: Art and Architecture Thesaurus [see all references], confidence: 1).

Figure 2D: Body of TOI-Finder Results Page

Format Showing TOIs, Common Noun Phrases, and External Vocabulary Terms

Resources such as the AAT, however, do not exist for every possible subject area, and it would be impractical to manually create such a set of terms for every subject-specific vocabulary. For this reason, the Technical Group also explored ways during the early stages of the project to create such lists automatically using subject specific texts. This method, if successful, could conceivably be applied to any subject, and added to the CLiMB suite of tools as an additional module. The group began with the hypothesis that words occurring more frequently in subject-specific texts than in common English usage would represent formal terminology, belonging to that particular field. To gauge this frequency, the group composed and implemented a variant of the Term Frequency / Inverse Document Frequency (TF/IDF) equation (Salton 1971), which is a method for measuring the relative frequency of words appearing in different sets of data. This algorithm was then run over both Chapter 5 of *Greene & Greene* and the "Brown corpus" (Francis & Kucera 1982), which aims to represent a balanced model of common usage in English. The corpus contains nearly one million words taken from news articles, literature, technical documents, and other sources. Those terms that appear significantly more often in Bosley than in the Brown corpus were determined to be part of the "domain" vocabulary, or architecture-specific (subject-oriented) terms. Though the algorithm was able to identify more subject-oriented terms than other natural language processing tools the group tested, it sacrificed a certain amount of specificity to Greene & Greene projects.

2.43 Employing Segmentation Techniques

Parallel to its work with POS taggers and NP chunkers, the Technical Group tested segmentation software. Segmentation is a technique for automatically breaking up a text into smaller parts, which tend to reflect coherence around particular topics. When a cataloger or an end-user deals with a collection of digital images, it is often the case that only certain parts of a text associated with those images will be applicable to their needs. For example, a discussion of the exterior of a given architectural project might take place within a few paragraphs, whereas the next paragraphs might discuss the front entrance. Those segments are not necessarily clear from headers in the text, but rather might be sequences of paragraphs or sentences within a text. Taking Chapter 5 of Bosley's *Greene & Greene* as a concrete example, a user searching for information on driveways for one of the Greenes' projects will not find all of the 47 paragraphs equally pertinent. Different paragraphs or sentences dealing with "driveways" might be related to the "Blacker house" or the "Thorsen house." As opposed to a standard key-word or phrasal search method that would simply identify "driveway," "Blacker," or "Thorsen" in the text of the chapter, segmentation identifies entire sections of written material that might pertain to particular terms in context.

The Technical Group has compared four linear segmentation techniques in order to determine which approach will best meet expert and end-user needs in the CLiMB suite of tools. Linear segmentation is a common method in computational linguistics that involves identifying non-overlapping and non-hierarchical segments in which there are no sub-units by topic. For instance, a news article might have a large segment on analysis that includes hierarchical sub-topics of criticism and commendation; linear segmentation marks these as two distinct segments, but does not identify the hierarchical relationship. A demonstration of the segmenters the group is comparing, each of which has been run over Chapter 5 of *Greene & Greene*, can be found on the CLiMB "Tools and Prototypes" page at http://www.columbia.edu/cu/cria/climb/tools.html. Results are displayed using a simple common single delimiter, represented by short lines of equal signs ('=====‘). For one tool, Segmenter, the results are also available in a more complex format, which marks paragraph and segment boundaries using labeled sequences.

The first tool on the demonstration page is TextTiling, a segmenter discussed in Hearst 1994. This tool identifies segments using cosine similarity, which is a standard mathematical technique for identifying comparable sections of text. Computation takes place over a window that slides across the text in a way similar to the window on a slide rule, making calculations for every possible position of the window. The standard window size is 100 words for news articles, though experiments on varying window size for different genres and purposes have been run. TextTiling identifies segments by looking for areas of low cohesion in the content of text; after smoothing dips and peaks in cohesion to eliminate any local artifacts, the remaining peaks and valleys are used to determine segment boundaries.

The second tool is Segmenter, which was developed within Columbia's NLP Group (Klavans, Kan & McKeown 1998). It employs a method called lexical chaining, which connects words by topic using independent lexical hierarchies such as WordNet (Miller 1994, Fellbaum 1998). Lexical chains link different elements of the text, in this case common nouns, proper nouns, and, when possible, co-referents (different words or phrases that refer to the same topic). Segmenter makes a guess about the strength of cohesion between paragraphs by weighing different features in the text. The higher the measured cohesion between elements of the text, the more likely it is that these elements are part of a topical segment.

The third segmenter, C99, was proposed by Choi 2000. This segmenter uses the sentence as a basic unit, and marks segments using the rank of cosine similarity over them, rather than using random arbitrary text blocks as in TextTiling or paragraphs as in Segmenter. C99 creates a sentence-by-sentence matrix over which a cosine-based density function is computed. This density function is maximized, and segments are selected based on high density. In contrast, neither TextTiling nor Segmenter employs maximization. The performance of Choi exceeds that of the first two tools for certain kinds of text where segment boundaries correlate with radical vocabulary changes.

Finally, the fourth tool, TextSeg (Utiyama and Isahara 2001), treats the text as a graph in which each word is a node. TextSeg computes a cost-function, which rewards word similarity as determined by stem overlap, and finds the lowest cost path using a Bayesian probabilistic minimization algorithm, i.e. guessing which word is likely to come after another one. Informally, the goal is to maximize the probability that a set of segments is related given the words as independent units. Unlike Hearst 1994, who uses word contiguity, or Klavans, Kan, and McKeown 1998, who use the paragraph, or Choi 2000, who uses the sentence, Utiyama and Isahara 2001 assume each word is independent of all other words in the document. This assumption fares well in shorter documents, although the computational complexity might be unwieldy for longer texts.

The Technical Group also experimented with an original segmenter based upon the Term Frequency / Inverse Document Frequency (TF/IDF) equation. The TF/IDF equation, as described above in the previous section, measures the relative concentration of a particular word or phrase across given texts; the group employed it here in a unique way to segment a document based upon where TF/IDF scores rise and fall significantly, which could indicate changes in topic.

In recent months the Group has built upon these initial tests using the TOI concept to focus work on segmentation processes. A segmentation module for the CLiMB suite of tools will employ the ability to locate TOIs to identify also which segments of the text discuss particular TOIs. The first part of this procedure involves segmenting text based upon co-reference, which identifies segments based upon the number of words or phrases that refer to a TOI in a given paragraph. Thus, returning to Bosley's *Greene & Greene* as an example, if a paragraph has more references to "the Blacker house" than it does to "the Thorsen house," the segmenter would identify that paragraph as being about the former because references to that house appear more often.

In some cases, however, the number of references to more than one TOI may be equal. The second part of the CLiMB segmentation procedure will "disambiguate" these paragraphs based upon co-occurrence, using a version of the modified TF/IDF equation discussed above. This involves using other elements of the text that appear in less ambiguous paragraphs to disambiguate others. If, for example, references to "the Blacker house" and "the Thorsen house" are equal in a given paragraph, the segmenter will look at paragraphs that are more readily identifiable as being about each of these two TOIs. In those less ambiguous paragraphs, other words or phrases may occur frequently near the TOI. In one paragraph that deals with the Blacker house early in Chapter 5 of *Greene & Greene*, there are several occurrences of the word "pergola," which the AAT identifies as "garden structures with open wood-framed roofs, often latticed, supported by regularly spaced posts or columns; often covered by climbing plants such as vines or roses, shading a walk or passageway." If the ambiguous paragraph also includes a number of occurrences of the word "pergola," along with other words or phrases that occur in the less ambiguous paragraph, the segmenter will determine that the ambiguous paragraph is likely about the Blacker house, rather than the Thorsen house.

3. CLiMB Access Platform

3.1 Presuppositions about Users

The CLiMB system architecture, which is described in detail in section 2.1 of this report, presupposes that there are two different types of users that must be considered in developing the CLiMB suite of tools. On the right-hand column of Figure 1A, "General Overview of CLiMB Project Design," users are seen entering the process at two points. The first of these is in phase II of the project, in which a set of experts evaluate the results of CLiMB metadata extraction; the second is in phase III, in which end-users can employ image search platforms that feature CLiMB metadata.

The role of the first group of users is elaborated in Figure 1C, "Build Metadata Records with CLiMB-derived Terms." After the CLiMB suite of tools has been employed to extract metadata from text, experts in the domain (or subject area), catalogers, reference librarians, and specialists in search and retrieval, will help us evaluate the results and develop standards for inserting CLiMB metadata into descriptive records. As the direction of the arrow in the Figure indicates, in this phase of the project CLiMB will be employing user input to modify extraction processes and refine the content of what we will insert into records. The idea behind employing a select group of this kind is that experts will know what metadata are valuable additions to existing catalog records: in other words, they will be able to indicate what items, in their judgment, are the right ones for the toolset to find. Furthermore, in later stages of the project, we can use their suggestions to build additional tools to guess when terms will be valuable. Our plan is to explore the use of machine learning techniques, which can take a set of hand-labeled examples, examine them automatically, and develop a "best-guess" approach to handling new sets of terms. In this way, the CLiMB toolset can be iteratively reworked to identify more accurately the terms and phrases that this group of users deems important.

The second group, which we represent as "end-users," will include general users of image search platforms as well as experts such as those in phase II. In phase III, however, the relationship of the user to the CLiMB project architecture is somewhat different than it is the previous stage. Here, the CLiMB suite of tools will not be "visible" as such; rather, it will be embedded behind the user interface of an existing image search platform in order to enhance search and retrieval processes as well as provide enriched descriptive metadata. As Figure 1D, "Testing CLiMB Metadata for Image Access" indicates, this phase also includes an important evaluation component. The "evaluation loop" that encircles the Figure transforms the one-way arrow at the bottom of the diagram into a cyclical process, whereby user evaluation can be continually employed to further refine the CLiMB toolset, in terms of its capacity both to identify and to select metadata. This group of users is purposefully conceived in broader terms than the first group, and includes anyone who might use an image search platform, rather than experts alone. A more detailed explanation of these aspects of the project can be found below in section 3.2.

As section 5 of this report indicates, user evaluation and testing is one of the two focused goals for CLiMB's second year. This will begin with the expert input in phase II. CLiMB will also be holding a special plenary meeting in Fall 2003, at which a select group of experts, some of whom are members of CLiMB's External Advisory Board, will advise the project teams about how best to approach the process of evaluation.

3.2 Building a CLIMB Testing Platform

The ultimate result envisioned for the CLiMB project is testing our results using existing image search and retrieval systems. This means that words and phrases identified by CLIMB processes must be able to be stored in standard library or museum catalog record fields, such as those for subject, name, or keyword. These records must then be capable of being loaded into existing image access systems. The two widely available commercial systems proposed for CLiMB testing are Endeavor's Voyager (http://www.endinfosys.com/prods/voyager.htm) and Luna's Insight (http://www.luna-imaging. com/insight.html), both of which will be available at Columbia during the second year of the CLiMB project. These systems vary sufficiently in their respective capabilities to allow us test a broad range of standard configuration and searching options. The Insight system in particular is optimized for image searching and display, and will be an important demonstration platform for CLiMB.

Since one of the anticipated issues in testing will be helping users to understand the nature of the results they get from CLiMB-based searching – results that may be at a different level of detail, density, or relevance than they have encountered previously – we also plan to create a prototype access system for testing. In order to accomplish this, CLiMB extracted metadata must be placed in catalog records that can be employed with the Columbia Libraries' production Structured Query Language (SQL) metadata infrastructure and database publishing system. SQL is an efficient and highly usable database query language that allows users to access information in a robust platform. This will allow us to customize user displays and messages and provide more direct ways for users to understand searching options, interpret results, and provide specific input to the project.

Once these testing platforms are in place we will be able to gather feedback from end-users. As described in section 3.1 above, this user input can be employed to improve the CLiMB metadata extraction and selection processes; it will also be able to help us evaluate our technology strategy overall.

3.3 Collecting Data on How People Use CLiMB Information

The first year of the CLiMB project has been focused on selection of appropriate collections, and on developing a basic set of tools. Once we have enough data to test with users, we will then be able to move into the phase of user testing. This phase is thus in very early stages, since at this point, without adequate data to test, we can design experiments and approaches but cannot yet engage in user evaluation. Understanding how users describe images in words can guide us in the development of the CLiMB toolset.

The CLiMB Curatorial Group has defined four aspects to the user testing process. The first is to collect information from colleagues on what kinds of queries tend to be posted that return images as the answer. This could be achieved by interacting with librarians involved in projects like the Art Museum Image Consortium (AMICO), who have vast user logs that could possibly be mined for information on user query behavior. This would have to be achieved for research purposes, maintaining privacy. Of course, without knowing the outcome of the query, such queries might not be as useful as if the full query session complete with patron feedback, could be obtained. In order to mimic this situation, we could pursue a second approach. This would involve gathering a workable list of questions that users are likely to ask, and then analyzing those questions to see what components might be useful for CLiMB. In this case, we would ask the information intermediary to construct sample queries, based on their experience. Thirdly, we could set up an experimental situation in which we give users a task, where success on the task is finding a given image. In this case, we could observe users' behaviors in using terms to satisfy the task. This would provide us with ample data on how users might use CLiMB terms, as opposed to existing authoritative terms, for finding images. Finally, we could provide a number of librarians with a succinct description of the project, and then ask them to list recent questions that they have encountered from users about image collections.

As CLiMB uses digitized images, a web survey might also be a potential vehicle for a study collecting user input on the CLiMB metadata terms they encounter. The suggested methodology for this approach might be as follows:

- Members of the CLiMB project teams and External Advisory Board participants can recommend groups and subject specialists from a variety of fields for participation.
- Targeted users will receive an email request for assistance. Included will be an explanation of the survey and a link to the CLiMB homepage.
- The survey will present dissimilar images and minimal descriptive metadata (no subjects) for each of the testbed collections.
- The users' task is to enter into query boxes words or phrases they would use to retrieve images similar to the ones they are viewing.

Survey participants will have the option of selecting one, two, or all three collections.

The Curatorial Group also recognized that it would be a valuable part of the information gathering process to consult with those who have created (as opposed to those whose primary responsibility is to maintain) existing image collection databases. It is a safe assumption that people in these positions have query logs that could provide a summary of how their collections have been approached by users. A difficulty that arises in obtaining information on users' queries of image collections is that many such searches are limited to the controlled vocabularies that structure the data in those collections. One of the major advantages of the CLiMB project is that the metadata extraction capabilities of the CLiMB toolset may enable users to employ words and word groups that are not normally thought of (or capable of being employed) as search terms, precisely because they don't appear in controlled vocabularies. The impact of this threatens to "google-ize" carefully selected and accurate metadata. The goal of CLiMB is to explore a middle ground where index terms are filtered and refined. Thus, users can search more broadly, but take advantage of CLiMB curatorial pre-processing and filtering.

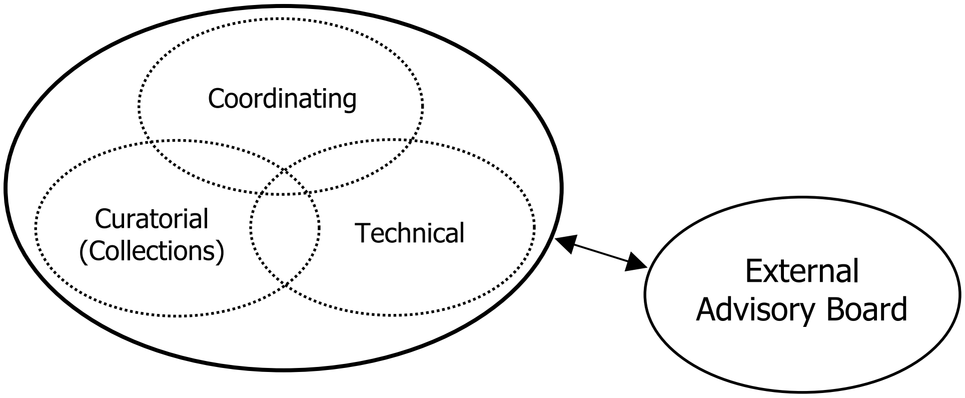
One of the goals of the External Advisory Board meeting is to gather input on user testing, and to formulate precise experiments for Year Two of CLiMB.

4. CLiMB People

4.1 CLiMB Organizational Chart

The chart below provides a visual representation of CLiMB's organizational structure. Each of the three CLiMB project groups (Coordinating, Curatorial, and Technical) works independently to some degree in accomplishing particular tasks. Communication between them is essential, however, and this is accomplished in part as a result of the fact that the groups overlap, as the diagram is meant to indicate. In the first instance, this overlap represents the distribution of CLiMB personnel. Two members of the CLiMB staff participate in all three groups (Judith Klavans and Stephen Davis); three people participate in two groups (Bob Wolven, Angela Giral, and Roberta Blitz). Just as importantly, the overlap is also functional. The work that each group undertakes separately both influences and is influenced by the others. For example, the selection and preparation of collections for use on the CLiMB project is the major objective for the Curatorial Group; at the same time it is necessary for these objectives to be accomplished in close collaboration with the Technical Group, which develops tools around these datasets as they are prepared.

CLiMB also works with an External Advisory Board, which is comprised of a carefully selected group of experts in fields related to the overall goals of the project. All of the members of the Board are taken from outside of the project itself, as the name should imply. The Board is the only part of the project that serves in a specifically consultative role.



4.2 CLiMB Project Teams

As the Organizational Chart shows, CLiMB is divided into three overlapping project teams, or Committees. The Coordinating Committee is charged with overseeing the CLiMB project as a whole, and thus serves both a conceptual and an administrative role. The Technical Group takes the development of the CLiMB toolset as its primary goal, and thus is responsible for creating computational linguistic tools capable of extracting robust metadata from texts associated with image collections. The Curatorial Group is responsible for producing and maintaining the project's collections, which entails a combination of several duties, among which are digitizing materials, securing access to electronic resources, and designing catalog formats for use on the CLiMB project.

After an extended round of regularly scheduled meetings at the beginning of the first year, each project group now meets primarily on an "as needed" basis. For the most part, this involves meetings every one to four weeks, with smaller groups meeting whenever required. For instance, the Curatorial Group has found that smaller scale meetings are most productive when dealing with particular collections, just as members of the Technical Group meet informally on a regular basis to discuss particular aspects of the project. Because members of the Coordinating Committee are actively involved in the workings of the other two groups, much of its supervisory role is integrated into the project itself. Below are complete lists of the members of each group.

Project Leader / Principal Investigator: Judith Klavans

CLiMB Coordinating Committee

- Judith Klavans, Director, Center for Research on Information Access (klavans@cs.columbia.edu)
- Patricia Renfro, Deputy University Librarian (pr339@columbia.edu)
- Stephen Davis, Director, Libraries Digital Program (daviss@columbia.edu)
- Bob Wolven, Director of Bibliographic Control & Library Systems (wolven@columbia.edu)
- Angela Giral, Director, Avery Architectural and Fine Arts Library (giral@columbia.edu)

CLiMB Curatorial Group

- Judith Klavans
- Angela Giral
- Stephen Davis
- Bob Wolven
- Roberta Blitz, Digital Collections / Art Research Librarian (rlb179@columbia.edu)
- Bob Scott, Head, Electronic Text Service (scottr@columbia.edu)
- Amy Heinrich, Director, Starr East Asian Library (heinrich@columbia.edu)
- David Magier, Director, Area Studies (magier@columbia.edu)

CLiMB Technical Group

- Judith Klavans
- Stephen Davis
- Roberta Blitz
- Peter Davis, Graduate Research Assistant (ptd7@cs.columbia.edu)
- David Elson, Programmer (delson@cs.columbia.edu)

4.3 Developing the External Advisory Board

The CLiMB External Advisory Board is comprised of fifteen people from outside of the project who possess backgrounds and interests that will enable them to contribute useful input and guidance as the project moves forward. The Board will meet annually to hear the reported results of the project; in addition, this will enable the CLiMB group to hear about related projects with which board members are involved. In this way these meetings are envisioned both as a means of enriching the CLiMB project through feedback and interaction with a select group of experts, and as an opportunity for an interdisciplinary group of participants to exchange ideas on a common set of concerns. The CLiMB Coordinating Committee extended official invitations during the Winter, and finalized Board Membership during the Spring in preparation for the Inaugural meeting, which is scheduled for June 13 of this year. Below is a list of Board members; this is also available on the CLiMB website at http://www.columbia.edu/cu/cria/climb/people.html.

CLiMB External Advisory Board	
Alfred Aho	**Columbia University** Chair, Department of Computer Science
Caroline Arms	**Library of Congress** Coordinator, National Digital Library Program
Murtha Baca	**The Getty Research Institute** Head, Standards and Digital Resource Management Program
Hilary Ballon	**Columbia University** Chair, Department of Art History and Archaeology
Michael Buckland	**University of California, Berkeley** Professor, School of Information Management & Systems
Jeff Cohen	**Bryn Mawr College** Director of Digital Media / Visual Resources
Greg Crane	**Tufts University** Editor-in-Chief, Perseus Project
Marilyn Deegan	**Oxford University** Digital Resources Manager, Refugee Studies Centre
David Fenske	**Drexel University** Dean, College of Information Science and Technology
Carl Lagoze	**Cornell University** Senior Research Associate, Information Science
Clifford Lynch	**Coalition for Networked Information** Executive Director
Merrilee Proffitt	**The Research Libraries Group** Program Officer
John Unsworth	**University of Virginia** Director, Institute for Advanced Technology in the Humanities
Nina Wacholder	**Rutgers University** Assistant Professor, School of Communication, Information & Library Studies
Clara Yu	**Middlebury College** Director, National Institute for Technology

5. Taking Next Steps

As we reach the end of Year One of the CLiMB project, we are optimistic about the experimental results we have achieved to date, but also conservative about possible outcomes. As a research project, our goal is to find out if and how computational linguistics can be used to help address the metadata issue. The aim is to extract high-quality metadata that is filtered and collection-specific.

We have two very focused goals for Year Two:

- Refine CLiMB tools to provide filtered output for loading into image access platforms
- Test our results with users

These two goals expand into many facets, including the continued development and testing of software for its ability to define and extract valuable descriptive metadata from text related to the three collections of digital images with which CLiMB is now working. The Curatorial Group will be focusing on two main areas: the establishment of a more authoritative set of "gold standards" for software testing, and the continued development of datasets. The former task will greatly enhance CLiMB's evaluation procedures in terms of measuring the relevance of metadata extracted from text; by using already-existing resources such as indexes and subject related vocabularies, the software being developed can be measured against the work of experts in scholarly fields associated with particular image collections. The second goal for the Curatorial Group will involve the continued creation of new datasets and the enhancement of existing ones. Whereas initial software tests were performed primarily on Greene & Greene materials, technical work now has a strong enough foundation to be able to branch out into more diverse sets of data. These will be developed from the Chinese Paper Gods collection and the South Asian Temples images. Finally, advice from the External Advisory Board meeting will be invaluable for taking the extensive technical and curatorial research already underway one step closer to integrating CLiMB suite of tools into a standard access platform, testing it with users, and packaging it for external use. A special meeting focusing on user testing and evaluation is planned for September of 2003. We plan to invite several of the members of our External Advisory Board, as well as selected experts in this area.

6. Selected References

Choi, Freddy Y. Y. (2000). "Advances in domain independent linear text segmentation." *Proceedings of the North American Chapter of the Association of Computational Linguistics* (NAACL). Seattle, Washington, 2000.

Cohen, William W. (1996). "Learning with set-valued features." *Proceedings of the Thirteenth National Conference on Artificial Intelligence*. Portland, Oregon, 1996.

Day, David, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson and Marc Vilain. (1997). "Mixed-Initiative Development of Language Processing Systems." *Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, Washington D.C., 1997.

Evans, David K., Judith L. Klavans, and Nina Wacholder (2000). "Document Processing with LinkIT." *Proceedings of the RIAO Conference*. Centre de Hautes Etudes Interationales d'Informatique Documentaire (CID) and the Center for the Advanced Study of Information Systems (CASIS), 2000.

Fellbaum, Christiane (1998). *WordNet, An Electronic Lexical Database*. MIT Press, Cambridge, MA.

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

Hearst, Marti A. (1994). "Multi-Paragraph Segmentation of Expository Text." *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics* (ACL). Las Cruces, New Mexico, 1994.

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown (1998). "Linear segmentation and segment relevance." *Sixth Annual Workshop on Very Large Corpora* (WVLC6). Montreal, Quebec, Canada, 1998.

Klavans, Judith L., Kathleen McKeown, Min-Yen Kan, and Susan Lee (1998). "Resources for the evaluation of summarization techniques." *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Grenada, Spain, 1998.

Francis, W. Nelson and Henry Kucera (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, MA.

Miller, George (1995). "WordNet: a lexical database for English." *Communications of the ACM* (Association for Computing Machinery), vol. 11, 1995.

Salton, Gerard, editor (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall: Englewood Cliffs, NJ.

Utiyama, Masao, and Hitoshi Isahara (2001). "A statistical model for domain-independent text segmentation." *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics* (ACL/EACL). Toulouse, France, 2001.

Wacholder, Nina (1998). "Simplex NPs sorted by head: a method for identifying significant topics within a document." Workshop on the Computational Treatment of Nominals. *Proceedings of the Joint 17th International Conference on Computational Linguistics – 36th Annual Meeting of the Association for Computational Linguistics* (COLING-ACL'98). Montreal, Quebec, Canada, 1998.

Appendices: CLiMB Supporting Material

A. Additional Information on CLiMB

- **CLiMB home page:** http://www.columbia.edu/cu/cria/climb/

  (The CLiMB home page includes links to "People," "Tools & Prototypes," "Presentations & Publications," and "Collections," as well as to the original grant proposal and press release.)
- **CRIA home page:** http://www.columbia.edu/cu/cria/

B. Collections and Related Material

Greene & Greene – Architectural drawings from the Avery Art and Architecture Library's Greene & Greene Collection

1. Bosley, Edward R. *Greene & Greene*. London: Phaidon, 2000.
2. Current, William R. *Greene & Greene: Architects in the Residential Style*. Fort Worth, TX: Amon Carter Museum of Western Art, 1974.
3. Makinson, Randell. *Greene & Greene: Architecture as a Fine Art*. Salt Lake City, UT: Pergrine Smith, 1977.
4. Makinson, Randell. *Greene & Greene: The Passion and the Legacy*. Salt Lake City, UT: Gibbs Smith, 1998.
5. Smith, Bruce. *Greene & Greene Masterworks*. San Francisco: Chronicle Books, 1998.
6. Strand, Janann. *A Greene & Greene Guide*. Pasadena, CA: G. Dahlstrom, 1974.

Finding Aid to the Greene & Greene Architectural Papers and Records Collection at Avery Library: http://www.columbia.edu/cu/lweb/eresources/archives/avery/greene/.

Greene & Greene Virtual Archives: http://www.usc.edu/dept/architecture/greeneandgreene/. This is a collaborative visual resource for Greene & Greene materials that includes the collection at Avery Library.

Chinese Paper Gods – Woodblock prints from the C.V. Starr East Asian Library's Anne S. Goodrich Collection

1. Day, Clarence Burton. *Chinese Peasant Cults: Being a Study of Chinese Paper Gods*. Taipei: Ch'eng Wen Publishing Co., 1974.
2. Goodrich, Anne Swann. *Peking Paper Gods: A Look at Home Worship*. Nettetal: Steyler Verlag, 1991.
3. Laing, Ellen Johnston. *Art and Aesthetics in Chinese Popular Prints: Selections from the Muban Foundation Collection*. Ann Arbor, MI: Center for Chinese Studies, University of Michigan, 2002.

South Asian Temples – Images with descriptive metadata from the AIIS Center for Art and Archaeology, Digital South Asia Library

1. *Encyclopaedia of Indian Temple Architecture*. New Delhi: American Institute of Indian Studies; Philadelphia: University of Pennsylvania Press. 1983-.
2. *Architectural Survey of Temples*. New Delhi: Archaeological Survey of India, 1964-.

Digital South Asia Library: http://dsal.uchicago.edu.

Authority Sources / Structured Vocabularies

The Getty *Art & Architecture Thesaurus On Line*: http://www.getty.edu/research/tools/vocabulary/aat/.

CLiMB has also employed the following two books as references on this topic:

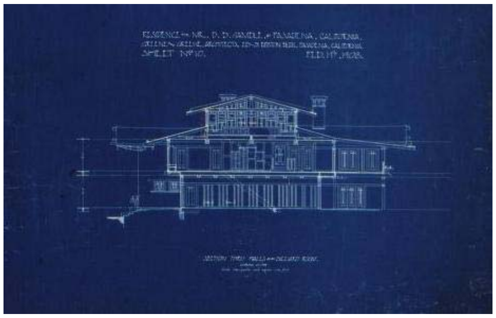
Baca, Murtha (editor). *Introduction Metadata: Pathways to Digital Information*. Los Angeles: Getty Information Institute, 2002.

Baca, Murtha (editor). *Introduction to Art Image Access: Issues, Tools, Standards, Strategies*. Los Angeles: J. Paul Getty Museum Publications, 2002.

C. Sample Catalog Records

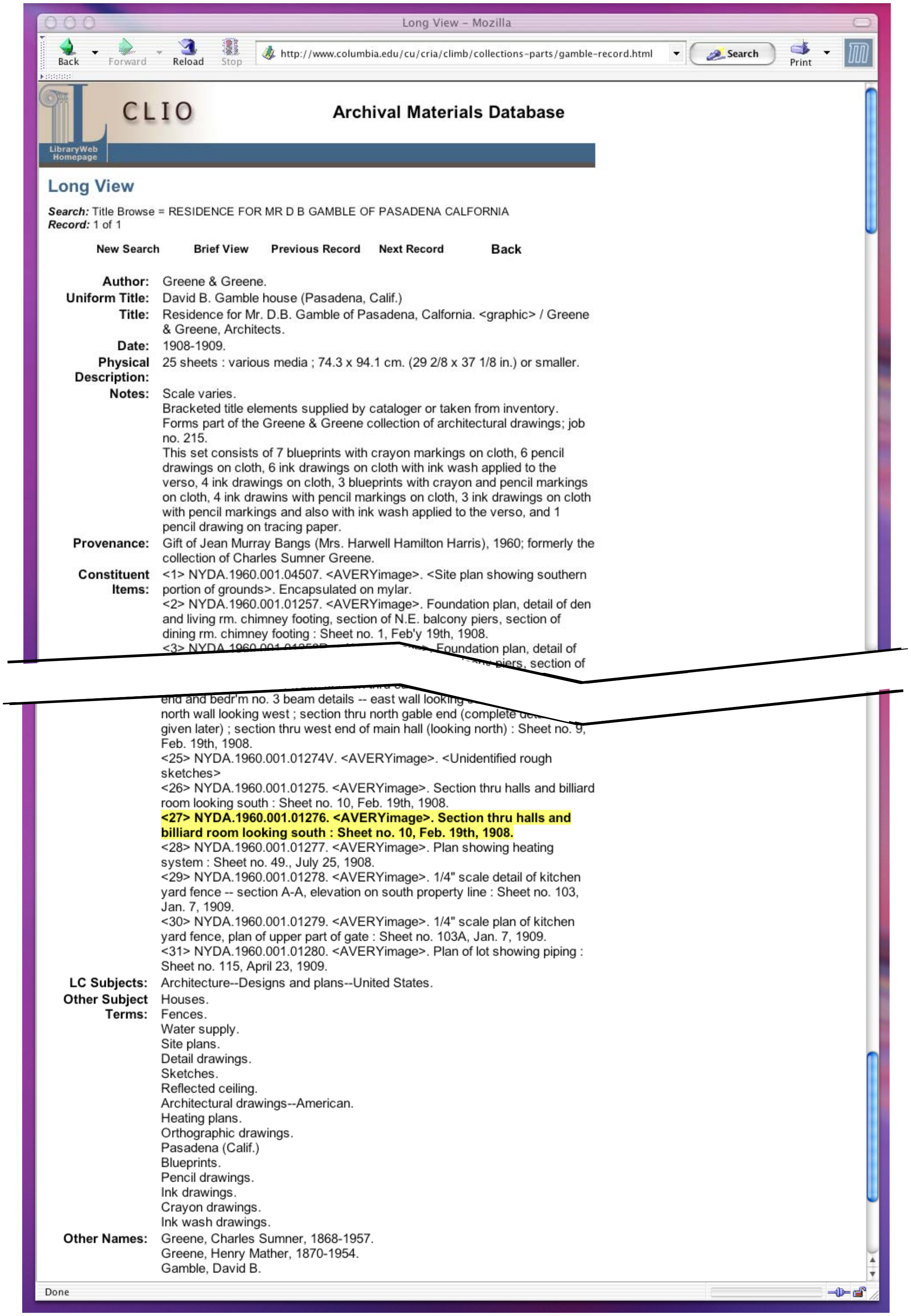
Greene & Greene Architectural Images

Greene & Greene Architectural Records and Papers Collection
Avery Architectural and Fine Arts Library, Columbia University



Residence for Mr. D.B. Gamble of Pasadena, California
Section through halls and billiard room looking south
Avery Architectural and Fine Arts Library,
Columbia University Libraries

Columbia University Catalog record for the Gamble house drawing:



Archival Materials Database

Long View

Search: Title Browse = RESIDENCE FOR MR D B GAMBLE OF PASADENA CALFORNIA
Record: 1 of 1

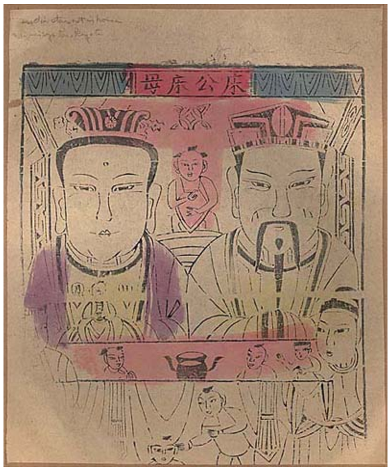
New Search   Brief View   Previous Record   Next Record   Back

Author: Greene & Greene.
Uniform Title: David B. Gamble house (Pasadena, Calif.)
Title: Residence for Mr. D.B. Gamble of Pasadena, Calfornia. <graphic> / Greene & Greene, Architects.
Date: 1908-1909.
Physical Description: 25 sheets : various media ; 74.3 x 94.1 cm. (29 2/8 x 37 1/8 in.) or smaller.
Notes: Scale varies.
Bracketed title elements supplied by cataloger or taken from inventory.
Forms part of the Greene & Greene collection of architectural drawings; job no. 215.
This set consists of 7 blueprints with crayon markings on cloth, 6 pencil drawings on cloth, 6 ink drawings on cloth with ink wash applied to the verso, 4 ink drawings on cloth, 3 blueprints with crayon and pencil markings on cloth, 4 ink drawins with pencil markings on cloth, 3 ink drawings on cloth with pencil markings and also with ink wash applied to the verso, and 1 pencil drawing on tracing paper.
Provenance: Gift of Jean Murray Bangs (Mrs. Harwell Hamilton Harris), 1960; formerly the collection of Charles Sumner Greene.
Constituent Items: <1> NYDA.1960.001.04507. <AVERYimage>. <Site plan showing southern portion of grounds>. Encapsulated on mylar.
<2> NYDA.1960.001.01257. <AVERYimage>. Foundation plan, detail of den and living rm. chimney footing, section of N.E. balcony piers, section of dining rm. chimney footing : Sheet no. 1, Feb'y 19th, 1908.
<3> NYDA.1960.001.0125 . Foundation plan, detail of
piers, section of

end and bedr'm no. 3 beam details -- east wall looking
north wall looking west ; section thru north gable end (complete
given later) ; section thru west end of main hall (looking north) : Sheet no. 9, Feb. 19th, 1908.
<25> NYDA.1960.001.01274V. <AVERYimage>. <Unidentified rough sketches>
<26> NYDA.1960.001.01275. <AVERYimage>. Section thru halls and billiard room looking south : Sheet no. 10, Feb. 19th, 1908.
<27> NYDA.1960.001.01276. <AVERYimage>. Section thru halls and billiard room looking south : Sheet no. 10, Feb. 19th, 1908.
<28> NYDA.1960.001.01277. <AVERYimage>. Plan showing heating system : Sheet no. 49., July 25, 1908.
<29> NYDA.1960.001.01278. <AVERYimage>. 1/4" scale detail of kitchen yard fence -- section A-A, elevation on south property line : Sheet no. 103, Jan. 7, 1909.
<30> NYDA.1960.001.01279. <AVERYimage>. 1/4" scale plan of kitchen yard fence, plan of upper part of gate : Sheet no. 103A, Jan. 7, 1909.
<31> NYDA.1960.001.01280. <AVERYimage>. Plan of lot showing piping : Sheet no. 115, April 23, 1909.
LC Subjects: Architecture--Designs and plans--United States.
Other Subject Terms: Houses.
Fences.
Water supply.
Site plans.
Detail drawings.
Sketches.
Reflected ceiling.
Architectural drawings--American.
Heating plans.
Orthographic drawings.
Pasadena (Calif.)
Blueprints.
Pencil drawings.
Ink drawings.
Crayon drawings.
Ink wash drawings.
Other Names: Greene, Charles Sumner, 1868-1957.
Greene, Henry Mather, 1870-1954.
Gamble, David B.

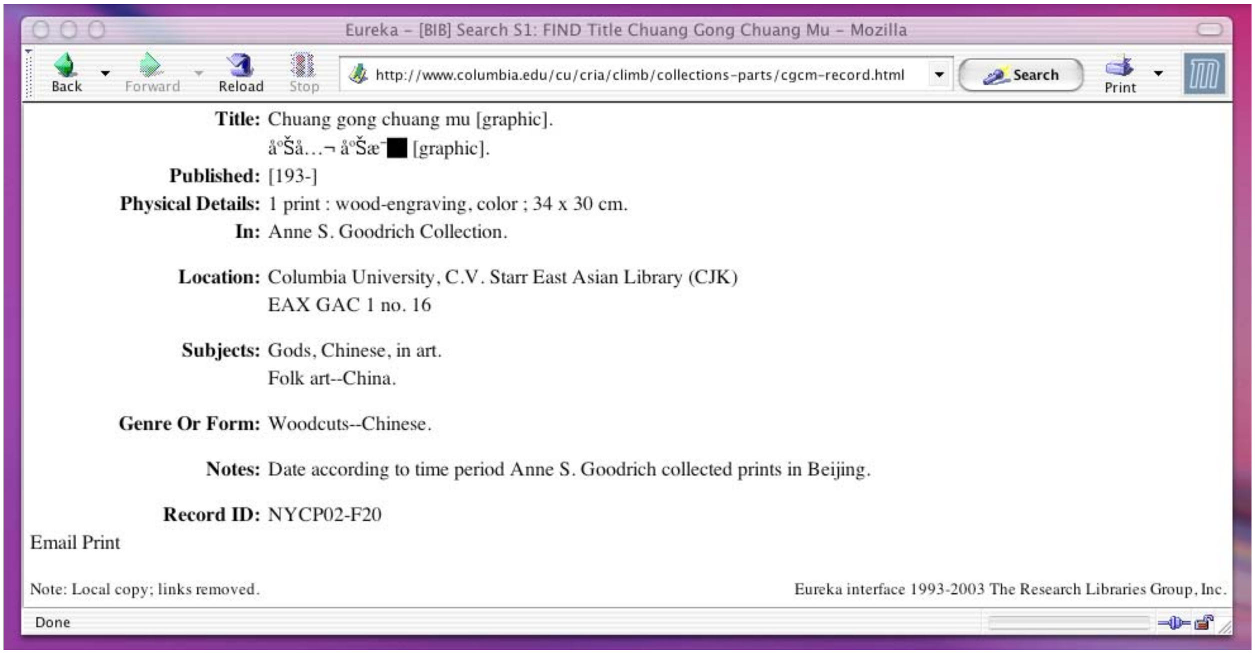
Chinese Paper Gods Collection

The Anne S. Goodrich Chinese Paper Gods Collection
C.V. Starr East Asian Library, Columbia University



Chuang gong chuang mu
Wood-engraving, color
C.V. Starr East Asian Library,
Columbia University Libraries

Catalog Record for *Chuang gong chuang mu*:


Eureka – [BIB] Search S1: FIND Title Chuang Gong Chuang Mu – Mozilla
Back
Forward
Reload
Stop
http://www.columbia.edu/cu/cria/climb/collections-parts/cgcm-record.html
Search
Print

Title: Chuang gong chuang mu [graphic].
å°Šå…¬ å°Šæ¯█ [graphic].

Published: [193-]

Physical Details: 1 print : wood-engraving, color ; 34 x 30 cm.

In: Anne S. Goodrich Collection.

Location: Columbia University, C.V. Starr East Asian Library (CJK)
EAX GAC 1 no. 16

Subjects: Gods, Chinese, in art.
Folk art--China.

Genre Or Form: Woodcuts--Chinese.

Notes: Date according to time period Anne S. Goodrich collected prints in Beijing.

Record ID: NYCP02-F20

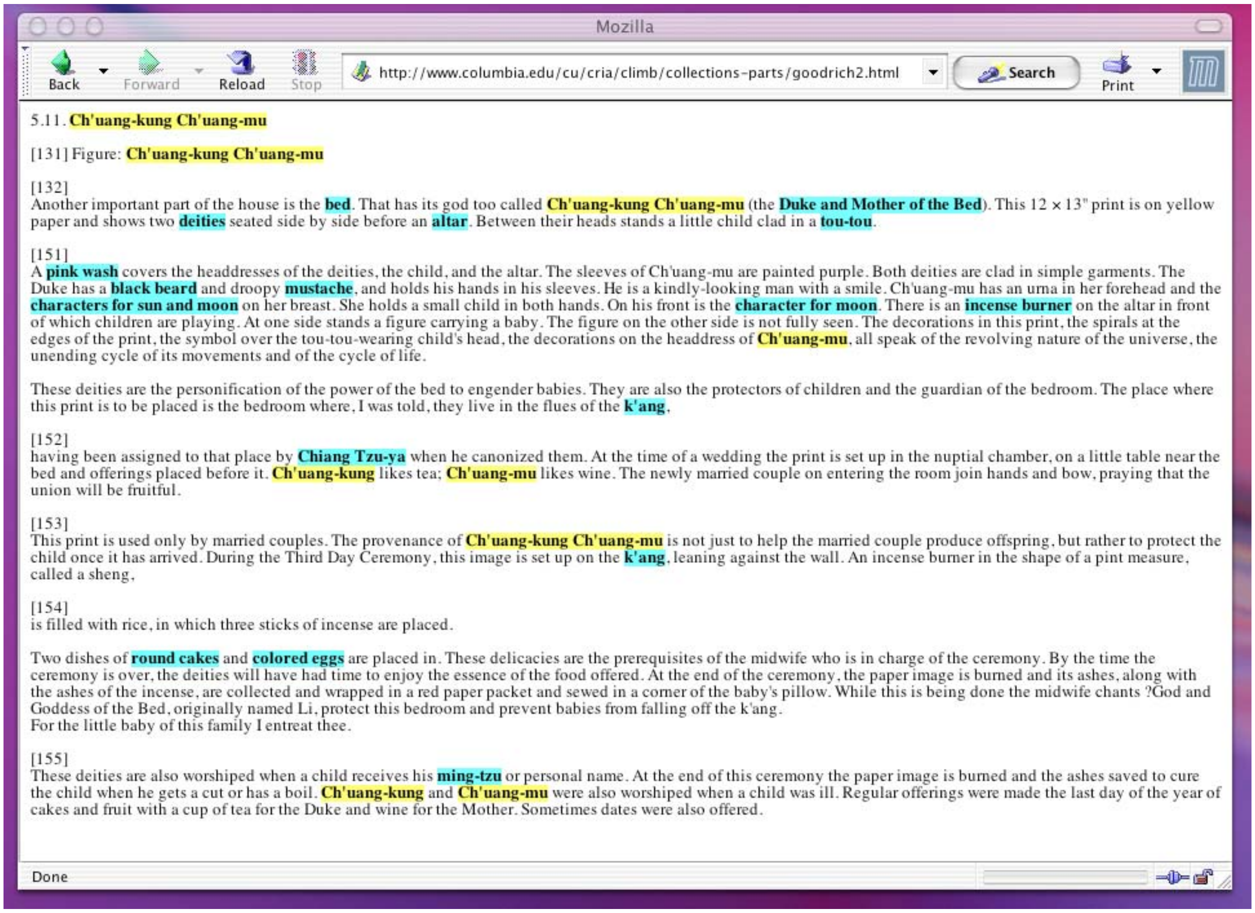
Email Print

Note: Local copy; links removed.

Eureka interface 1993-2003 The Research Libraries Group, Inc.

Selected text from Anne S. Goodrich, *Peking Paper Gods*.

TOIs and descriptive terms are highlighted.



Mozilla
Back
Forward
Reload
Stop
http://www.columbia.edu/cu/cria/climb/collections-parts/goodrich2.html
Search
Print
Done

5.11. **Ch'uang-kung Ch'uang-mu**

[131] Figure: **Ch'uang-kung Ch'uang-mu**

[132]
Another important part of the house is the **bed**. That has its god too called **Ch'uang-kung Ch'uang-mu** (the **Duke and Mother of the Bed**). This 12 × 13" print is on yellow paper and shows two **deities** seated side by side before an **altar**. Between their heads stands a little child clad in a **tou-tou**.

[151]
A **pink wash** covers the headdresses of the deities, the child, and the altar. The sleeves of Ch'uang-mu are painted purple. Both deities are clad in simple garments. The Duke has a **black beard** and droopy **mustache**, and holds his hands in his sleeves. He is a kindly-looking man with a smile. Ch'uang-mu has an urna in her forehead and the **characters for sun and moon** on her breast. She holds a small child in both hands. On his front is the **character for moon**. There is an **incense burner** on the altar in front of which children are playing. At one side stands a figure carrying a baby. The figure on the other side is not fully seen. The decorations in this print, the spirals at the edges of the print, the symbol over the tou-tou-wearing child's head, the decorations on the headdress of **Ch'uang-mu**, all speak of the revolving nature of the universe, the unending cycle of its movements and of the cycle of life.

These deities are the personification of the power of the bed to engender babies. They are also the protectors of children and the guardian of the bedroom. The place where this print is to be placed is the bedroom where, I was told, they live in the flues of the **k'ang**,

[152]
having been assigned to that place by **Chiang Tzu-ya** when he canonized them. At the time of a wedding the print is set up in the nuptial chamber, on a little table near the bed and offerings placed before it. **Ch'uang-kung** likes tea; **Ch'uang-mu** likes wine. The newly married couple on entering the room join hands and bow, praying that the union will be fruitful.

[153]
This print is used only by married couples. The provenance of **Ch'uang-kung Ch'uang-mu** is not just to help the married couple produce offspring, but rather to protect the child once it has arrived. During the Third Day Ceremony, this image is set up on the **k'ang**, leaning against the wall. An incense burner in the shape of a pint measure, called a sheng,

[154]
is filled with rice, in which three sticks of incense are placed.

Two dishes of **round cakes** and **colored eggs** are placed in. These delicacies are the prerequisites of the midwife who is in charge of the ceremony. By the time the ceremony is over, the deities will have had time to enjoy the essence of the food offered. At the end of the ceremony, the paper image is burned and its ashes, along with the ashes of the incense, are collected and wrapped in a red paper packet and sewed in a corner of the baby's pillow. While this is being done the midwife chants ?God and Goddess of the Bed, originally named Li, protect this bedroom and prevent babies from falling off the k'ang.
For the little baby of this family I entreat thee.

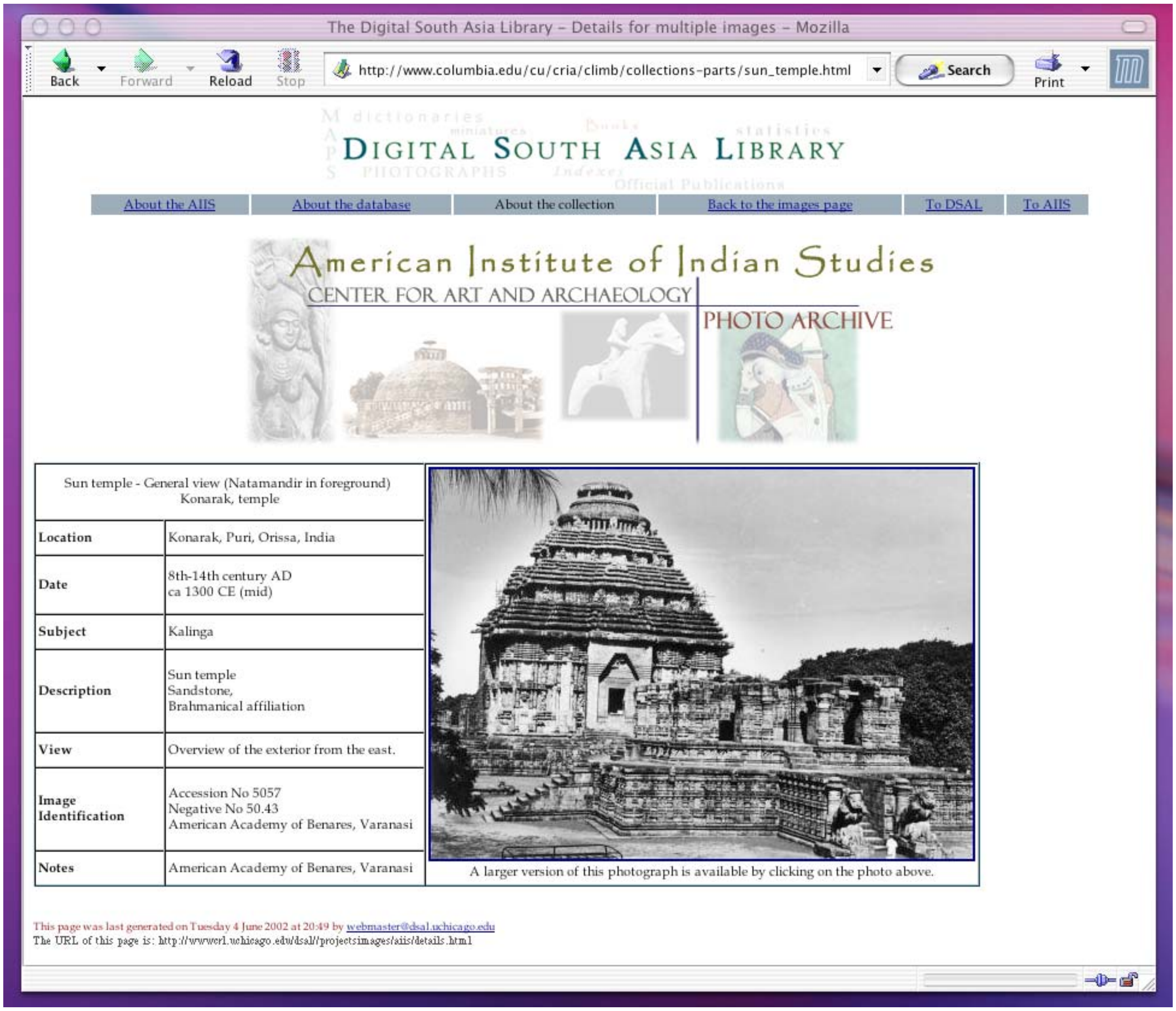
[155]
These deities are also worshiped when a child receives his **ming-tzu** or personal name. At the end of this ceremony the paper image is burned and the ashes saved to cure the child when he gets a cut or has a boil. **Ch'uang-kung** and **Ch'uang-mu** were also worshiped when a child was ill. Regular offerings were made the last day of the year of cakes and fruit with a cup of tea for the Duke and wine for the Mother. Sometimes dates were also offered.

South Asian Temples Images and Metadata

The Digital South Asia Library

Sun temple - General view (Natamandir in foreground)
Location Konarak, Puri, Orissa, India
From the photo archives of the American Institute of Indian Studies

Individual Image with Metadata from the Digital South Asia Library:



The Digital South Asia Library – Details for multiple images – Mozilla

http://www.columbia.edu/cu/cria/climb/collections-parts/sun_temple.html

About the AIIS | About the database | About the collection | Back to the images page | To DSAL | To AIIS

Sun temple - General view (Natamandir in foreground) Konarak, temple	
Location	Konarak, Puri, Orissa, India
Date	8th-14th century AD ca 1300 CE (mid)
Subject	Kalinga
Description	Sun temple Sandstone, Brahmanical affiliation
View	Overview of the exterior from the east.
Image Identification	Accession No 5057 Negative No 50.43 American Academy of Benares, Varanasi
Notes	American Academy of Benares, Varanasi

A larger version of this photograph is available by clicking on the photo above.

This page was last generated on Tuesday 4 June 2002 at 20:49 by webmaster@dsal.uchicago.edu
The URL of this page is: http://wwwcrl.uchicago.edu/dsal//projectsimages/aiis/details.html

D. CLiMB Presentations & Talks: May 2002 – May 2003

- External Advisory Board Meeting, June 12th and 13th, 2003.

- Third ACM/IEEE Joint Conference on Digital Libraries (JCDL), May 28th, 2003.
  Presentation of paper on "Methods for Precise Named Entity Matching in Digital Collections." The paper was authored by Peter Davis, David Elson, and Judith Klavans. Available on line at http://www.columbia.edu/cu/cria/climb/presentations.html.

- Presentation on CLiMB for the Professional Staff at Avery Fine Art and Architectural Library, Columbia University. Given by Roberta Blitz and Angela Giral, May 15th, 2003.

- Coalition for Networked Information, April 28th, 2003.
  Presentation on the CLiMB project at CNI. Live demo and presentation slides available at http://www.columbia.edu/cu/cria/climb/presentations.html.

- CLiMB talks on "Computational Linguistics and Digital Libraries" delivered by Judith Klavans:

  - Hong Kong University Library, Hong Kong, February 14th, 2003.
  - Tsinghua University, Department of Computer Science, Beijing, China, February 28th, 2003.
  - National Institute of Informatics, Information Science, Tokyo, March 5th, 2003.
  - Toyo Bunko Oriental Library, Tokyo, March 5th, 2003.
  - Waseda University Library, Waseda, Japan, March 5th, 2003.
  - Kyoto University, Department of Computer Science, Kyoto, Japan, March 6th, 2003.

  Abstract: As digital libraries have grown, so has the need for developing more effective ways to access collections. This talk will present an overview of the CLiMB project (Computational Linguistics for Metadata Building), funded by the Mellon Foundation and currently underway at Columbia University. The goal of the project is to use computational linguistic techniques to extract metadata relevant to image collections, and thus to improve cataloging access. This research addresses the access bottleneck by applying the latest natural language processing techniques to the problem of identifying descriptive metadata. Our goal is to load our results into a database for image search, although we have not yet reached this phase of the project. This talk will report on research in CLiMB's first phase. In addition, the talk will provide an overview of selected digital library projects at Columbia, in terms of collections, access and technology.

- Mellon Technical Meeting, November 21st, 2003.
  Presentation given to Mellon Foundation participants on the technical and methodological aspects of the CLiMB project. Slides for this "Report on Technical Progress to the Mellon Foundation" can be found at http://www.columbia.edu/cu/cria/climb/presentations.html.

E. CLiMB Glossary of Terms

Because CLiMB is an inherently interdisciplinary project, many of the terms we employ are either specialized for use in certain fields, or they have been developed by the CLiMB project internally in order to address particular conceptual needs. With the hope of making the CLiMB project accessible to people in fields as diverse as those represented in our project teams, this short glossary provides definitions and explanations for some of the terms that this report and the project generally tend to rely upon.

associated text. A term used by the CLiMB project for texts that are potentially related to the items in a given image collection. For instance, in the first year of testing, much of the work focused on Edward Bosley's book *Greene & Greene*, which is a text associated with the Greene & Greene Collection in Columbia's Avery Library.

associational context. The extent of the text surrounding an occurrence of a Target Object Identifier or its variants (or any other term located by the CLiMB toolset) that is deemed likely to yield relevant associated words and concepts. Associational context may, for instance, include a certain number of words before and after the occurrence of a TOI. The exact limit would be rule-based but flexible for different cases.

authority list. A set of relevant terms, such as names of places and subject vocabularies, that is provided a priori to CLiMB software by outside experts. The authority list for Greene & Greene architectural projects, for instance, comes from the master list of project names decided upon by the catalogers at Columbia's Avery Library.

Brown Corpus. A 1,000,000-word corpus of edited English prose compiled by W. Nelson Francis and Henry Kucera at Brown University in 1964. Designed to represent common English usage, it draws from sources such as newspapers, humanities texts, and fiction.

catalog record, item level. A catalog record that describes a specific item in a collection. In the catalog record for the Greene & Greene Collection drawing shown above in Appendix C, the highlighted text near the bottom of the record represents item level cataloging: "<27> NYDA. 1960.001.01276. <AVERYimage>. Section thru halls and billiard room looking south : Sheet no. 10, Feb. 19th, 1908." These item level records are subordinated to records at the project level.

catalog record, project level. A catalog record that describes part of a collection based upon the "project" it refers to. CLiMB uses this term mostly with reference to the Greene & Greene collection. The catalog record in Appendix C provides project level information near the top of the screen shot; for instance, the "Uniform Title" is "David B. Gamble House (Pasadena, Calif.)." Thus, the record encompasses those constituent items that relate to this project.

cataloging. Describing works of art, architectural projects, visual representations of works of art or architecture, or bibliographic materials. Cataloging most often involves the creation of a systematic record that provides a basic classification and description for the object being cataloged. Frequently the terms used to catalog an object come from standardized vocabularies.

co-occurrence. A relationship among words or phrases in a text based upon near proximity. Two words or noun phrases are co-occurrent when they appear close to each other in a text on at least one occasion.

co-reference. A relationship among words or phrases in a text determined by whether they describe the same object. Two words or noun phrases are co-referents when they both refer to the same thing: for example, "our nation's first President" and "George Washington," or "the Thorsen house" and "the house designed for William Thorsen."

DTD. Document Type Definition. A DTD is the grammar that describes the structure of an XML file, and tells the computer how that particular file should be read. The CLiMB DTD, for example, indicates that for a given text each chapter is made up of sections, that each section is made up of paragraphs, that each paragraph is made up of sentences, and so forth.

external vocabulary (also *domain dictionary* or *domain vocabulary*). A set of nouns and noun phrases that describe objects relevant to the subject at hand as a whole. For the Greene & Greene collection, where the domain is architecture, external vocabulary terms may include "porte cochere" and "inlaid brick," but not terms specific to Greene projects such as "the Blacker entry." At present the CLiMB suite of tools will locate external vocabulary terms in text from a Getty Structured Vocabulary, the *Art and Architecture Thesaurus*.

gold standard. A benchmark against which to test a software tool. For instance, in running the CLiMB toolset over a text, we might use experts to develop a list of terms and phrases to represent a standard against which to measure the results of automatic metadata extraction; another option would be to employ existing catalog records as standards by which to measure the effectiveness of the toolset in replicating or enhancing existing record information.

MARC. MAchine-Readable Catalog format. MARC represents a standard for describing bibliographic items (or objects in a collection) in catalog records. It is used by the Library of Congress, and aids in the exchange of data among information systems.

metadata. Data describing other data. For example, the size of a Web page, and the word that conveys the gist of its content, are metadata. In the CLiMB project, metadata usually refers to words and phrases in text that describe the items in a given image collection.

metadata schema. A set of rules for structuring metadata information such that it can incorporate specific elements of that information in encoded form.

MMF. Master Metadata File. According to the Columbia University Libraries' web pages, "Columbia's Master Metadata File (MMF) is a locally-developed metadata repository built around a MARC-based relational database schema. As of Sept. 2002, it holds over 75,000 metadata records for digital items & collections held locally or accessed remotely. The schema was designed to be able to represent multiple versions, collections, aggregations such as pages in a book, and hierarchies of digital objects. Information may be imported and exported in several formats. The database also may be used as an intermediate architectural component and may be queried interactively."

NP chunker. Noun Phrase chunker. A software tool that automatically locates noun phrases in a text.

POS tagger. Part of Speech tagger. A software tool that automatically locates (or "tags") the parts of speech in a text.

pinyin (also see *Wade-Giles*)**.** A transliteration system for the romanization of Chinese written characters that has been in use since the 1950s. This method of transferring written Chinese into a Latin alphabet has already been the standard for the United States Government for more than two decades; it is also the standard used by the United Nations and most of the world's media. Pinyin is largely coming to replace Wade-Giles, an older system that is now thought to provide a much less accurate representation of Chinese phonemes. A common instance of this replacement is the substitution of "Beijing" (pinyin) for "Peking" (Wade-Giles).

precision (also see *recall*)**.** In computational linguistics, a characteristic of the results obtained by running a software tool over a text in order to identify certain terms or types of terms. Precision is a measure of the accuracy with which the tool identifies the correct terms. For example, if a tool is designed to identify noun phrases, and does so with a precision of 90%, this means that 90% of the terms identified by the tool are in fact noun phrases. Precision does not account for whether the tool has located all of the desired terms (this is expressed by 'recall').

precision versus recall. Often, results in computational linguistics are expressed by a combined score for precision and recall. Generally, as one rises, the other tends to fall. There are several methods for obtaining satisfactory overall scores. At present, the CLiMB suite of tools begins by searching for high-recall but low-precision matches in a given text, and then seeks to increase the precision of those matches with further computation.

recall (also see *precision*)**.** In computational linguistics, a characteristic of the results obtained by running a software tool over a text in order to identify certain terms or types of terms. Recall is a measure of the tool's ability to locate all of the desired terms. For example, if a tool is designed to identify noun phrases, and does so with a recall of 90%, this means that the tool has located 90% of all possible noun phrases. Recall does not account for whether the tool has identified the terms accurately (this is expressed by 'precision').

segmenter. A software tool that automatically breaks a text up into smaller parts, or segments, based upon content or topic. "Segmentation" thus refers to a technique for dividing texts into discrete topical segments pertaining to specific themes.

SQL. Structured Query Language – pronounced "sequel". A standard language used for the expression of database queries. SQL allows users to search for and retrieve information from databases.

TEI XML / CLiMB TEI XML. Text Encoding Initiative Extensible Markup Language. TEI XML is XML that adheres to the standard for encoding texts in electronic form that is issued by the TEI Consortium (http://www.tei-c.org/). CLiMB currently uses XML that has overlap with the TEI standard. Our eventual goal is to achieve full compliance with TEI guidelines; in the report we refer to this as "CLiMB TEI XML."

term, domain specific (or *subject specific*). A term or phrase that is specific to a given subject area or discipline. For instance, the "domain" for Greene & Greene projects is architecture, and would include words that are specific to the field, such as "porte cochere" or "pergola."

term, project specific. A term or phrase that is specific to a particular project or object; often a TOI, but also words associated with a project that are not domain specific. For instance, the city

"Pasadena" is a project specific term for the "Freeman A. Ford house," because it is the project's location.

TF/IDF. Term Frequency / Inverse Document Frequency. An equation for measuring the relative frequency of words appearing in different sets of data. CLiMB has applied this equation to Chapter 5 of Edward R. Bosley's *Greene & Greene* in order to compare frequency of words in that chapter to the Brown Corpus, which is a model representing standard English usage.

TOI / Target Object Identifier. A noun phrase in a text that directly refers to a discrete target object. For the target object "the David B. Gamble house" (Greene & Greene collection), TOIs may include "the Gamble house," "the house," and "his residence in Pasadena," but not "the Gamble entry" or "the site." However, TOIs vary from collection to collection. Each of CLiMB's three collections uses a different kind of term for TOIs: for Greene & Greene, TOIs are terms referring to architectural projects; for the Chinese Paper Gods, TOIs are the names of the gods depicted; for the South Asian Temples Images, TOIs are place names for temple sites.

Wade-Giles (also see *pinyin*)**.** An older transliteration system for the romanization of Chinese written characters. This method of transferring written Chinese into a Latin alphabet is mostly being replaced by pinyin, a more modern system that is thought to provide a superior representation of Chinese phonemes. A common instance of this replacement is the substitution of "Beijing" (pinyin) for "Peking" (Wade-Giles).

XML (also see *TEI XML*)**.** Extensible Markup Language. A version of SGML (Standard Generalized Markup Language), which is a standard for defining the structure of different types of electronic document. XML allows organizations to customize their own markup languages for structuring data.

CLiMB – Computational Linguistics for Metadata Building

Center for Research on Information Access
Columbia University

508 Butler Library – MC 1103
535 W 114th Street
New York, NY 10027
212.854.7443

http://www.columbia.edu/cu/cria/climb