The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0737-8831.htm

OTHER ARTICLE

Integrating thesaurus relationships into search and browse in an online photograph collection

Michelle Dalmau, Randall Floyd, Dazhi Jiao and Jenn Riley
Indiana University Digital Library Program, Bloomington, Indiana, USA

Received 31 August 2004
Revised 10 November 2004
Accepted 10 November 2004

Abstract

Purpose – Seeks to share with digital library practitioners the development process of an online image collection that integrates the syndetic structure of a controlled vocabulary to improve end-user search and browse functionality.

Design/methodology/approach – Surveys controlled vocabulary structures and their utility for catalogers and end-users. Reviews research literature and usability findings that informed the specifications for integration of the controlled vocabulary structure into search and browse functionality. Discusses database functions facilitating query expansion using a controlled vocabulary structure, and web application handling of user queries and results display. Concludes with a discussion of open-source alternatives and reuse of database and application components in other environments.

Findings – Affirms that structured forms of browse and search can be successfully integrated into digital collections to significantly improve the user's discovery experience. Establishes ways in which the technologies used in implementing enhanced search and browse functionality can be abstracted to work in other digital collection environments.

Originality/value – Significant amounts of research on integrating thesauri structures into search and browse functionalities exist, but examples of online resources that have implemented this approach are few in comparison. The online image collection surveyed in this paper can serve as a model to other designers of digital library resources for integrating controlled vocabularies and metadata structures into more dynamic search and browse functionality for end-users.

Keywords Controlled languages, Photography, Digital storage, Collections management

Paper type Technical

Background

The Charles W. Cushman Photograph Collection (www.dlib.indiana.edu/collections/cushman/) is an online resource providing access to nearly 15,000 color photographs

The four authors of the paper all contributed to this project in significant and unique ways, as did approximately 15 other members of the project team from the Indiana University Digital Library Program and Indiana University Archives. Development of the Cushman collection web site was funded through an Institute of Museum and Library Services National Leadership Grant. The authors would like to thank everyone involved with the project for their hard work, patience, and inspiration. Special thanks to Kristine Brancolini, Jon W. Dunn and John A. Walsh for their helpful review of this paper.

Library Hi Tech
Vol. 23 No. 3, 2005
pp. 425-452
© Emerald Group Publishing Limited
0737-8831
DOI 10.1108/07378830510621829

shot by Charles W. Cushman, an amateur photographer, from 1938 to 1969. The vast majority of the photographs in the collection were taken on Kodachrome color slide film, which was originally introduced in 1936. The slides were taken at hundreds of locations all over the world, and there are particularly large quantities of images taken in Chicago from 1941 to 1951 and in San Francisco from 1954 to 1969. The collection shows a similar breadth of subject matter. Cushman was apparently fond of photographing plants, but also seemed to favor shooting architecture and people, often showing these in various states of decay or misfortune.

Cushman, an Indiana University alumnus, bequeathed his collection of slides, along with notebooks documenting each slide and some additional related materials, to Indiana University upon his death in 1972, where they were deposited in the Indiana University Archives (http://www.indiana.edu/ ~ libarch/). The collection was rediscovered in the Archives in late 1999 and recognized as remarkable for its breadth, level of documentation, and representation of color photography in a time we today generally envision in black and white. The Indiana University Digital Library Program (www.dlib.indiana.edu/) and Indiana University Archives collaborated on the project to digitize and build a delivery system for the images, which was funded by an Institute of Museum and Library Services (IMLS) National Leadership Grant (www.imls.gov/). The large amount of description that came with the slides allowed us to focus our development efforts on creating robust metadata for the collection and using it in novel ways for searching and browsing.

Metadata issues

Metadata for image collections

Creators of image collections commonly follow the traditional metadata model of providing a combination of free-text descriptions and access terms from controlled vocabularies. Free-text descriptions serve many functions in metadata records. For historical or archival collections, these descriptions can preserve the terminology used to describe an item by its creator or an important collector. They provide human-readable, in-depth details about an item, such as "pen and brown (iron-gall) ink and wash, graphite, watercolor, gouache and opaque white, with gum arabic and scraping out, on gray wove paper" for an art drawing (Visual Resources Association, 2004, p. 97). Descriptions of this sort can supply context and expert interpretation for end-users.

Appropriate terms, called "authorized terms," from controlled vocabularies are commonly added to metadata records for personal and corporate names, geographical locations, form and genre, and the topical nature of the item being described, whether or not the concept represented is already present in the descriptive fields. This is done for a number of reasons. First, controlled vocabularies increase the number of access points available for an item. By specifying ahead of time a fixed set of information fields – such as geographical location at country, state, county and city levels; names; genre terms; and topical terms – records become more consistent and thus more useful for searching. Second, the use of controlled vocabularies ensures that the same term is used to describe the same concept, same person, or same place among all records, a practice which results in the collocation of all relevant records under a single form of a term. Similarly, controlled vocabularies provide disambiguation between different meanings of the same term or different people and places with the same name by ensuring every authorized term is unique through the use of qualifiers and other devices. Finally, the use of controlled vocabularies promotes interoperability between collections. The same mechanisms that increase access points and provide collocation and disambiguation functions within a discrete collection can do the same among disparate collections when the same controlled vocabularies are used.

However, many digital library systems do not realize the full potential of controlled vocabularies. Most controlled vocabularies have some sort of syndetic structure which describes relationships between terms in the vocabulary. These structures generally follow the ANSI/NISO standard Z39.19-1993: *Guidelines for the Construction, Format, and Management of Monolingual Thesauri* (ANSI/NISO, 1993). Perhaps most frequently used are equivalence relationships, commonly expressed as "Use" and "Used for." This is the very heart of a controlled vocabulary, pointing synonyms or near-synonyms for a concept to a single authorized heading that should be used to describe the concept in a structured environment. An unauthorized term that points to an authorized term is known as a "lead-in term." This relationship is also used in authoritative lists of personal, corporate, and geographic names, directing users from all versions of the name to one "preferred" version. Some systems, including some library catalogs such as the Sirsi Unicorn and Endeavor Voyager products, make use of this relationship by either informing an end-user of a preferred term when they enter a non-preferred term, or less often by simply taking the user directly to records matching a preferred term from any non-preferred term.

Another pair of commonly used relationships in controlled vocabularies are hierarchical relationships, usually expressed as "Broader term" and "Narrower term." These relationships are used frequently in controlled vocabularies for topical, form, and genre terms, but much less often in vocabularies for personal, corporate, and geographic names. Traditional cataloging practice dictates the assignment of the most specific term in a hierarchy appropriate to an item. This practice makes the implicit assumption that the hierarchy of the vocabulary will be available for use when searching. However, most search systems do not make use of this hierarchy, so the responsibility for locating and using appropriate terms from the vocabulary falls to the searcher. Thus if a user searches for "sports," a system that does not exploit the hierarchical relationships of the controlled vocabulary would only retrieve images that have been assigned the specific term "Sports." This assignment would only occur when an image represents either a generic view of sports and no more specific term is available, or multiple sports are present in the image and no single sport is the real focus. For the user to gain access to images of specific sports, she must first either guess as to appropriate headings or determine the controlled vocabulary in use and consult it to determine which specific sports terms are available. She must then perform a search on each of the terms identified (or combine them in a Boolean OR query if available).

One other relationship is common among controlled vocabularies, an associative relationship, usually expressed as "Related term." This relationship denotes that one term is a valid heading for use (and may have Used for, Broader term, and Narrower term relationships to other terms), but that a second term, also a valid heading for use, is in some way related to the first. The cataloger or searcher may find the second term also of value when the first is appropriate. Very few systems make use of this information to aid in retrieval.

Metadata for the Cushman Collection

The Cushman collection came with a great deal of descriptive information, in the form of extensive notations on the slide mounts and in accompanying notebooks (see Figures 1-3) by Mr Cushman himself. He commonly provided the date the photograph was taken, an extraordinarily precise location (often an exact street addresses), the names of people pictured, and other details. He also provided descriptions of the scenes in his own words, such as "Airing the quilts. Farm house off US 150 N.W. of Terre Haute." (http://purl.dlib.indiana.edu/iudl/archives/cushman/

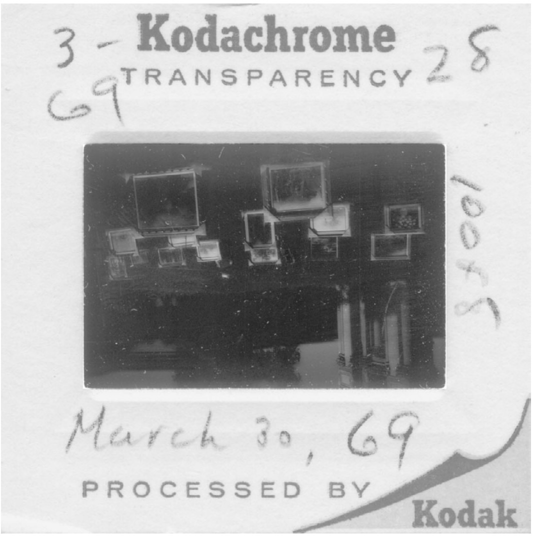


Figure 1.
Mr Cushman's notations on a slide mount

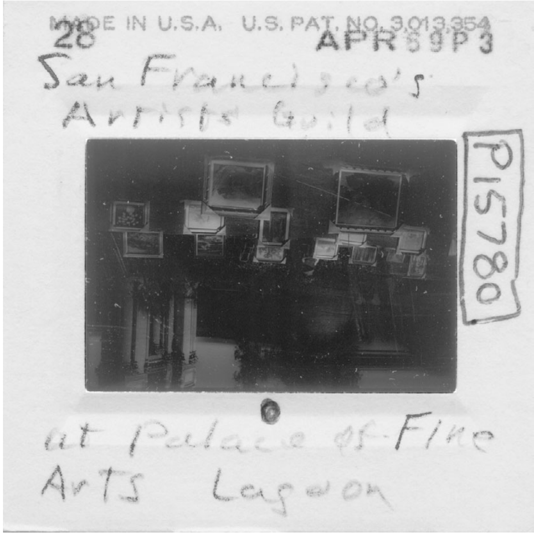
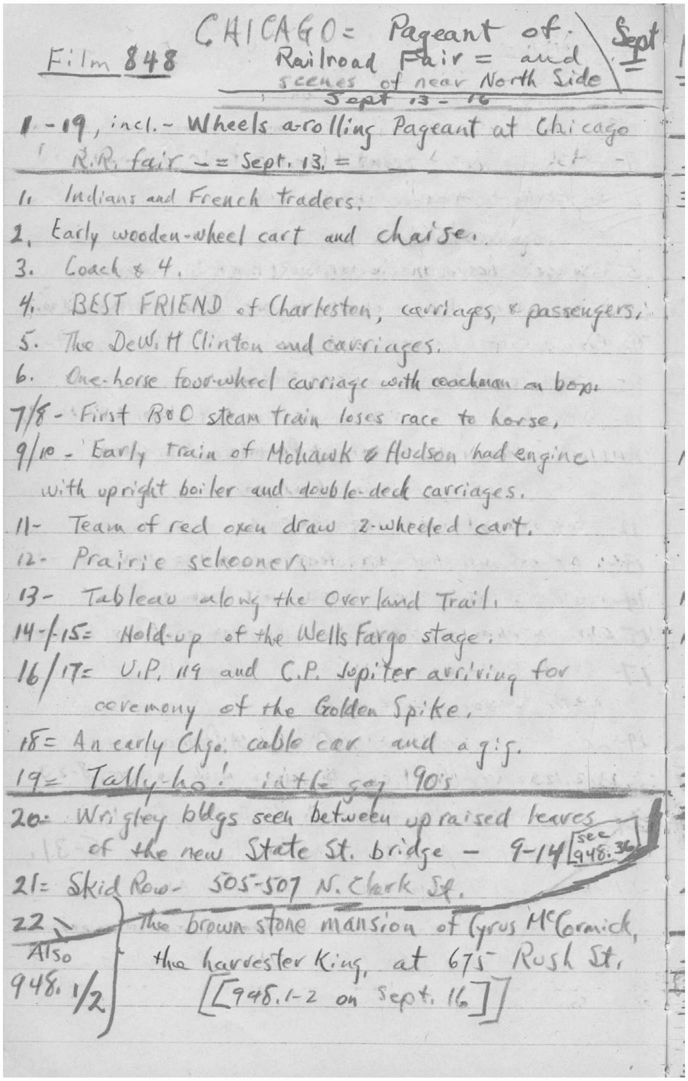


Figure 2.
Mr Cushman's notations on a slide mount, reverse side



Film 848

CHICAGO = Pageant of Railroad Fair = and scenes of near North Side

Sept 13 - 16

Sept I

1 - 19, incl. - Wheels a-rolling Pageant at Chicago R.R. fair - = Sept. 13. =

1. Indians and French traders.
2. Early wooden-wheel cart and chaise.
3. Coach & 4.
4. BEST FRIEND of Charleston, carriages, & passengers.
5. The DeWitt Clinton and carriages.
6. One-horse four-wheel carriage with coachman on box.

7/8 - First B&O steam train loses race to horse,

9/10 - Early train of Mohawk & Hudson had engine with upright boiler and double-deck carriages.

11 - Team of red oxen draw 2-wheeled cart.

12 - Prairie schooner.

13 - Tableau along the Overland Trail.

14-/-15= Hold-up of the Wells Fargo stage.

16/17= U.P. 119 and C.P. Jupiter arriving for ceremony of the Golden Spike.

18= An early Chgo. cable car and a gig.

19= Tally-ho! in the gay '90's

20= Wrigley bldgs seen between upraised leaves of the new State St. bridge - 9-14 [see 948.36]

21= Skid Row - 505-507 N. Clark St.

22 Also 948. 1/2 — The brown stone mansion of Cyrus McCormick, the harvester King, at 675 Rush St. [[948.1-2 on Sept. 16]]

Figure 3.
Mr Cushman's notations in a notebook

P03026) These descriptions are invaluable in giving us an insight into Cushman's view of the world. They are also, however, rife with abbreviations, misspellings, and factual errors, which in any given case may or may not have been deliberate.

To supplement Mr Cushman's own words for viewing and searching by end users, we employed catalogers to add controlled information to the metadata record for each image. We entered dates in a standardized form to facilitate date browsing. For personal and corporate names, we used forms from the Library of Congress Name Authority file (LCNAF) when they were available, and forms conforming to AACR2r when they were not. For place names, catalogers used the Getty's Thesaurus of

Geographic Names (TGN) (www.getty.edu/research/conducting_research/vocabularies/tgn/), information from the US Geological Survey, online map databases, and other resources to identify standard country, state or province, county, city, neighborhood, and other geographic feature names based on the information Cushman provided. We entered each of these place names into separate fields in our metadata records; the hierarchical structure of TGN was used only by the cataloger for research purposes, not by our search and browse application. Based on the scene depicted in a slide, our catalogers also assigned one or more genre terms from the Library of Congress' Thesaurus for Graphical Materials II: Genre & Physical Characteristic terms (TGM II) (http://lcweb.loc.gov/rr/print/tgm2/).

Many of our catalogers' efforts were spent assigning topical terms to images from the Library of Congress' Thesaurus for Graphical Materials I: Subject terms (TGM I) (www.loc.gov/rr/print/tgm1/). Since the Cushman collection is by a single artist, and is in a single medium (photography), the greater scope of a vocabulary such as the Getty's Art & Architecture Thesaurus (AAT) (www.getty.edu/research/conducting_research/vocabularies/aat/), which includes facets for visual resources by form, function, location, medium/technique, and subject, was not necessary. We, like the Library of Congress, "…concluded that neither LCSH nor AAT were appropriate for large, general collections of visual materials" (Alexander and Meehleib, 2001, p. 191). Therefore TGM I, which includes a syndetic structure indicating relationships between terms, was our vocabulary of choice for use by catalogers to assign topical headings to images in the collection.

Despite the obvious appropriateness of TGM I for the Cushman collection, and its pervasive use as a vocabulary for describing visual materials by catalogers, it has some weaknesses as a vocabulary supporting end-user searching and browsing. First, and perhaps most significantly, the lead-in vocabulary is insufficient. In addition to lacking entries for common words pointing to an authorized synonym, there are two persistent types of omissions of lead-in vocabulary. TGM I terms for nouns are generally plural forms of words, following the Z39.19 thesaurus standard (ANSI/NISO, 1993, p. 6). However, TGM I does not provide singular forms of terms as lead-in vocabulary. Trained and knowledgeable human users can make up for this shortcoming; however, most end-users of search systems never realize topical headings are created according to a controlled vocabulary, much less take the action necessary to learn how terms are created within the vocabulary. Similarly, without the assistance of lexical databases, search systems cannot by themselves add access from singular forms of words to their plural equivalents if the thesaurus lacks the explicit connection.

TGM I lead-in vocabulary is also consistently lacking in cases where the authorized heading is a compound term. The TGM I thesaurus commonly uses headings made up of multiple terms connected with an ampersand, for example, "Nails & spikes." Headings of this sort are used when two terms are not exact synonyms, but are close enough in meaning that it is appropriate in the vocabulary to use one heading for both concepts. TGM I, however, does not consistently provide lead-in term access from the second word in these headings to the complete heading. A system that provides keyword searching of subject headings would partially alleviate the access problem caused by this lack of lead-in vocabulary, but a corresponding lack of precision would accompany this gain in recall.

TGM I's broader and narrower term structure is also incomplete. When new terms are added, they are frequently not adequately connected to other terms in the thesaurus. For example, in early 2004, the term "School attendance" was added. This term has "Used for" references from "Absenteeism (School)" and "Truancy (School)." References from other synonyms that a user might enter, such as "Absences," were not created. No broader or narrower term relationships were added, despite numerous possibilities including "School children," "Students," and "Schools."

Two final weaknesses of TGM I are worth mentioning, both of which stem from its origin as a tool for catalogers rather than for end-users. Like many thesauri and other types of controlled vocabularies, the authorized headings chosen for TGM I tend to be in formalized language that bears very little similarity to search terms entered by typical end-users. The TGM I term for "Communes," for example, is "Collective settlements." TGM I also does not fare well for use as a subject browsing structure. Of the nearly 6,900 authorized terms in TGM I as of February, 2004, over 1,300 do not have a broader term. A hierarchical subject browse would ideally have a very small number of highest-level terms – Yahoo! (http://dir.yahoo.com/), for example, has 14. The top-level terms that do exist in TGM I also suffer from the same formalized language problem as the rest of the thesaurus. An end-user would have to choose the obscure term "Natural phenomena" to browse down the hierarchy to "Clouds" or "Moonlight."

Design toward integration

Our design team was committed to ensuring that the broad audience for the Cushman web site, ranging from the casual browser to the urban American history scholar, could intuitively access the images and metadata abundant in this online resource. We believed that in order to facilitate user exploration of the web site, the collection's underlying metadata and, more importantly, its relational structure, needed to be exposed. To validate this belief, we investigated current online resources using controlled vocabularies, conducted a review of published research, and implemented a series of user studies to understand the extent to which we should expose the controlled vocabulary structure.

Review of existing online resources

Few current online resources provide a browse and search interface that integrates existing descriptions and controlled vocabularies as a core part of the discovery functionality. At most, the typical resource attempts to expose controlled vocabularies either by providing rudimentary access via a long alphabetical list of terms, or by providing a separate search interface to a thesaurus. One example of such a site is the US Steel Gary Works Photograph Collection, 1906-1971 (Steel collection) (www.dlib.indiana.edu/collections/steel/), an image collection documenting the building of the Gary Works steel mill and the corporate town of Gary, Indiana, released by the Indiana University Digital Library Program in February 2002. TGM I topical headings were assigned to images in the Steel collection and were organized into an alphabetical subject browse on the web site, but the structure of the thesaurus as a whole was not used to aid in image discovery for end-users.

Usability studies of the Steel collection revealed that users seldom consulted the long list of preferred subject headings provided. Instead, they would enter a topical term in hopes that it would match the controlled term or a term in the description. Often, users' searches yielded results, but when they searched for unauthorized synonyms such as "doctors," they would receive zero results since "physician" was the authorized term. Findings such as these led us to further investigate the utility of incorporating the structure of the controlled vocabulary into the Cushman Web application.

Within the academic research community, we found experimental online image collections with accompanying research literature that exemplify the user experience we considered emulating in the Cushman web site. The University of Michigan's SI Art Image Browser[1] was designed to test how classification schemes can be utilized to organize browse and search options as well as augment discovery. The SI Art Image Browser relies on the Getty's Art and Architecture Thesaurus (AAT) to provide users with broad categories, called facets, to assist in discovery. A key feature of the SI Art Image Browser is a hierarchical browse based on the AAT's structure, allowing users to easily expand or refine a result set. The University of California, Berkeley's Flamenco online images research project (http://bailando.sims.berkeley.edu/flamenco.html) provided additional inspiration for our Cushman design team. Flamenco provides a hybrid browse and search interface based on the underlying metadata structure allowing users to easily explore vast domains (Hearst *et al.*, 2002, p. 45). Like the SI Art Image Browser, Flamenco relies on the AAT's structure for browsing hierarchically. The Flamenco project also supports combining categories or facets while browsing as a way to easily refine or broaden the result set (Hearst *et al.*, 2002, p. 47). While both web sites use the metadata's hierarchical structure to assist browsing, neither integrate the syndetic structure of the thesaurus to supplement searching.

Related research

For some time, scholars have been researching the benefits and detriments of keyword and controlled vocabulary searching, and most have found that users would benefit from information retrieval systems that expose thesauri (Fidel, 1991a, b; Dubois, 1984, 1987; Soergel, 1999). Current advocates of integrating metadata structures with retrieval systems are extending the findings of their predecessors with user-informed operational theories and experimental applications (Greenberg, 2001a, b, 2004; Tudhope *et al.*, 2001, 2002). Following Fidel's approach of studying users with actual information needs, Jane Greenberg conducted an iterative user-based study in which she collected from users real-time queries posed to databases with underlying thesauri (Greenberg, 2001a, b). These queries were later mapped to terms in a controlled vocabulary and their place in the vocabulary's syndetic structure in order to make recommendations for query expansion methods that would enhance precision or recall, depending on the users' search goals. Greenberg determined that synonyms and narrower terms of a query were good candidates for automatic retrieval by the system while related and broader terms are better candidates for user-initiated retrieval (Greenberg, 2001b, pp. 494-6).

Research that focuses on image discovery reveals that image seekers rely predominantly on textual labels, especially topical headings, as access points (Choi and Rasmussen, 2002, p. 507). As a result of a two-part study, Choi and Rasmussen highly recommended in-depth cataloging of images following established metadata standards and controlled vocabularies to ensure increased access. Researchers from the Getty

Research Institute have also explored the value of integrating thesauri structures when searching for images, and as a result, have proposed concepts such as "search assistants" that present users with broader, narrower and related terms in various ways as the user engages in discovery (Baca, 2003; Harpring, 1999).

Usability studies

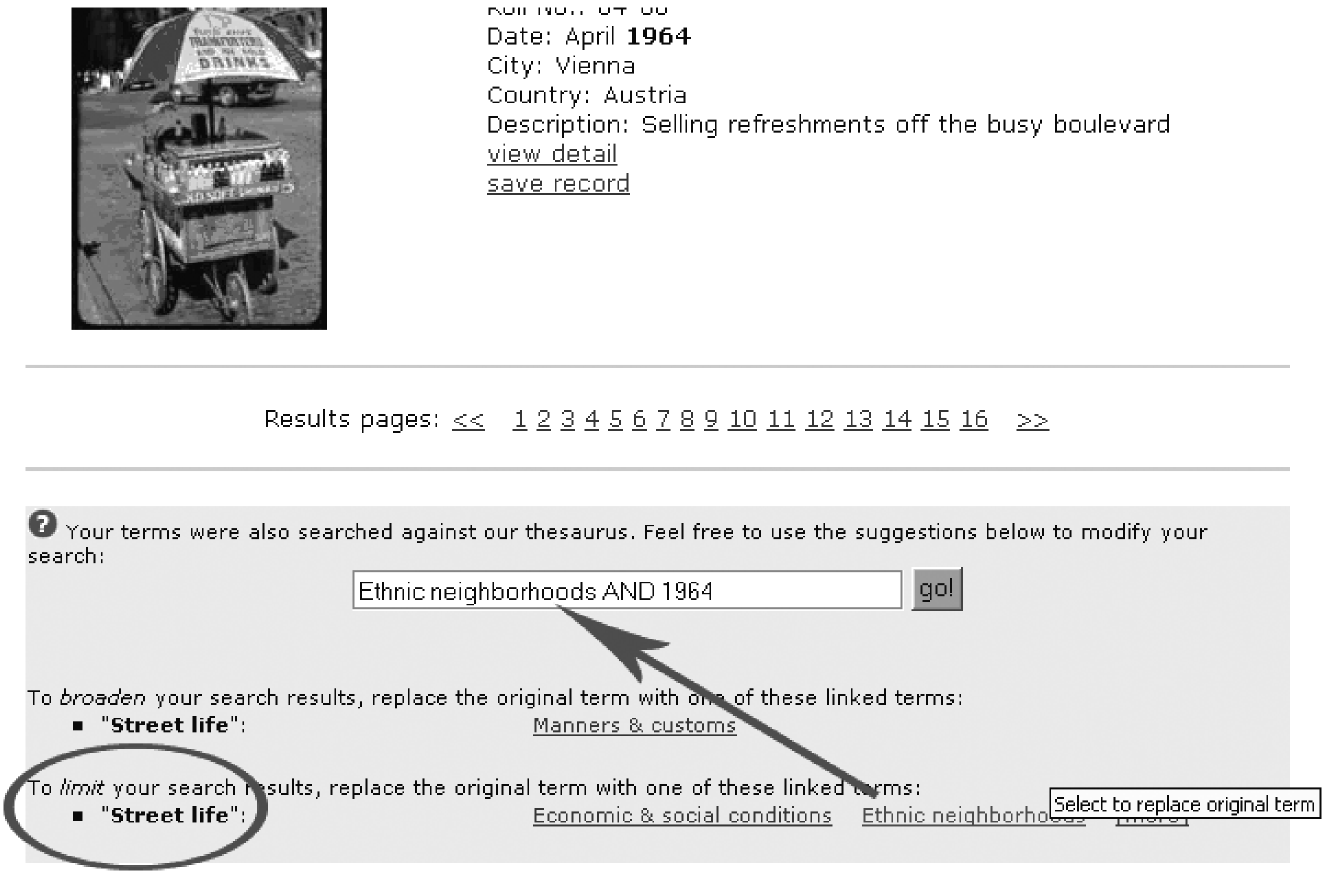
To build on what we had learned from existing online collections and published research, we conducted a series of user studies at various stages of the Cushman project development cycle. Evaluations were performed on HTML prototypes for user feedback. The prototypes presented were illustrative, not prescriptive, which allowed the user to explore alternative presentations of and interactions with the controlled vocabulary. These prototypes evolved into the Cushman web site now available.

The walkthrough study. In January and February of 2003, a thorough evaluation of an early prototype web site for the Cushman collection was conducted with the help of representative users. The study focused on many aspects of the site, including its ability to promote successful searching by exposing the TGM I structure to end-users. In this prototype (see Figure 4), search results generated topical suggestions based on the user's query, but these were not available from subject browsing. The browse options prototyped were by date, genre, and location. The ability to combine browse categories, for example, images of covered bridges in Indiana, was not available in this version of the prototype.

The group walkthrough methodology as outlined by Randolph Bias[2] was adopted for this user study. Walkthroughs were performed with both groups and individuals. The group walkthrough entailed the individual completion of seven browse and search-related task scenarios based on paper versions of the prototyped screens. Discussion followed regarding the participants' individual approaches to the tasks as well as general comments about the prototype. The individual walkthrough followed a similar format. Instead of task scenarios, individual participants spent part of their time demonstrating how they actually discover and use images in their teaching and research.

Six participants were recruited: a special collections librarian who oversees several photograph collections on campus, two graduate students from the School of Library and Information Science who were also part-time reference assistants, a professional photographer, and two faculty members from the history and art history departments. The faculty members were part of the individual walkthrough study. Their collective areas of expertise ranged from popular and urban history to photography. All participants were regular users of the internet, and all have consulted online image collections on a regular basis either in reference librarian work, for lectures and research, or for general interest.

All participants completed background questionnaires, which collected data regarding their research and subject areas, computer expertise, and exposure to similar online image resources. Satisfaction questionnaires were also completed, which gave the participants an opportunity to anonymously comment on the presentation and organization of the prototype web site. The task scenario responses were coded and measured against ideal paths for completion of the stated task. Because the methodology was discussion-based, the bulk of the data collected was qualitative.



Roll No.: 64-86
Date: April **1964**
City: Vienna
Country: Austria
Description: Selling refreshments off the busy boulevard
view detail
save record

Results pages: << 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 >>

Your terms were also searched against our thesaurus. Feel free to use the suggestions below to modify your search:

Ethnic neighborhoods AND 1964 go!

To *broaden* your search results, replace the original term with one of these linked terms:

- **"Street life"**: Manners & customs

To *limit* your search results, replace the original term with one of these linked terms:

- **"Street life"**: Economic & social conditions Ethnic neighborhoods [more]

Figure 4.
Prototype of search suggestions used in the group walkthrough study

Overall, the participants reacted positively to the prototype. All the participants touched upon the importance of supporting refinement while searching. Their suggestions correlate with research that overwhelmingly shows that online browsing and searching is an iterative process; users search and repeatedly revise their search until they encounter what they need (Choo *et al.*, 2000). Participants suggested several types of refinement, including the ability to search within a retrieved result set or a saved result set, and the presentation of clustered results like those seen in the Vivisimo (http://vivisimo.com/) and Teoma (www.teoma.com/) search engines. Refining within a result set using a search box is becoming standard practice for most digital collections; however, refining using links, rather than just a search box, is less common.

The librarians and reference assistants shared many examples of queries posed by their patrons such as, "I am looking for photos of covered bridges in Indiana," which revealed the need to support the combination of browsing categories. The ability to combine location and subject browse categories as well as other categories such as year and genre would facilitate such access.

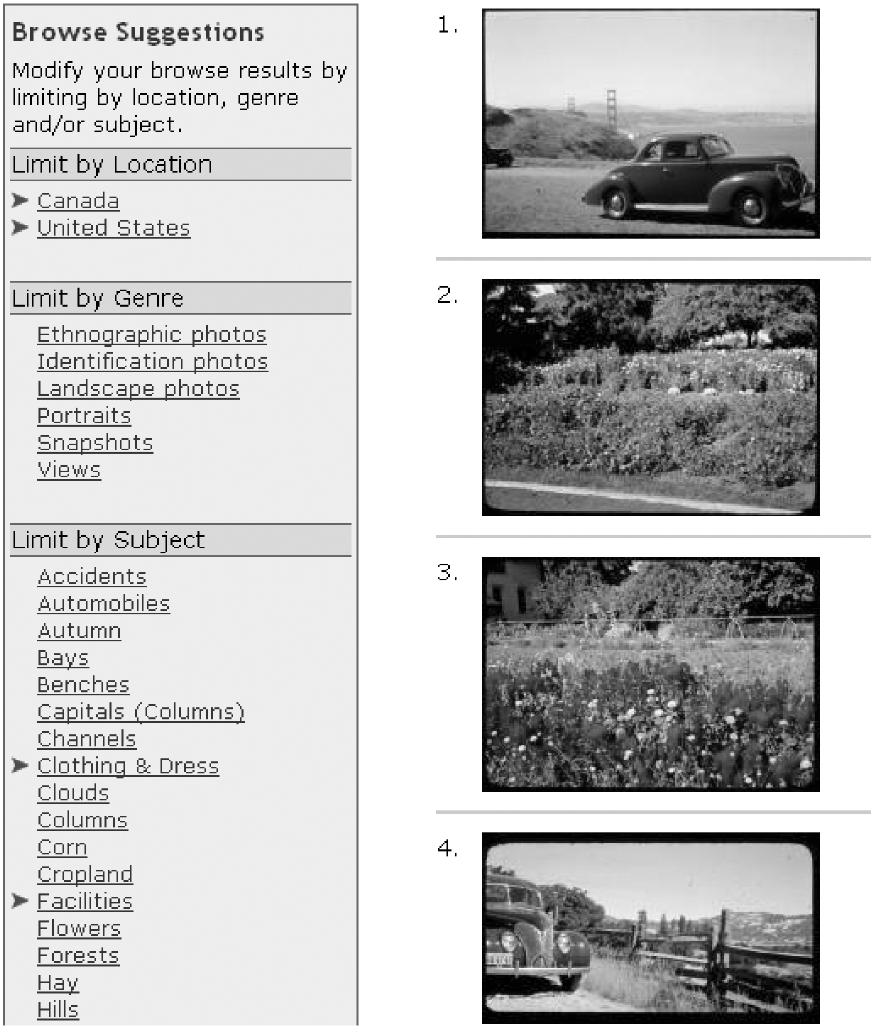
Five out of the six participants welcomed search suggestions based on the TGM I thesaurus structure offered by the prototype. Many felt that suggestions would reveal more about the subject matter represented within the collection as well as provide helpful cues for formulating better or new queries. The majority of participants who welcomed the suggestions were either accustomed to e-commerce web sites with similar functionality, such as Amazon, or understood the functions of thesauri. These results led us to believe that search suggestions based on broader and narrower terms in TGM I should be explored further in later prototypes and user tests.

Several participants expressed concerns with word choice, as is evident by this recurring question: "Do I type US, USA, United States or America to generate hits?" This concern, which also emerged in our Steel collection usability findings, supported our recommendation to integrate the controlled vocabulary's cross-reference structure into the search engine on the Cushman site, thereby allowing users to access images described with synonyms or variant names.

The task-scenario study. The goal of the task-scenario study was primarily to evaluate browse and search functionality by presenting participants with a set of representative tasks to complete. The tasks were created to explore further findings from the preceding walkthrough study and to determine if Greenberg's optimal query expansion processing methods matched our users' expectations for automatic and user-initiated browse and search. Following Greenberg's (2001b, p. 496) findings, the prototype for this test processed a topical query as follows:

- all lead-in vocabulary was automatically mapped to the preferred term;
- records using narrower terms were automatically retrieved; and
- user-initiated options (browse and search suggestions) supported navigation up and down the hierarchy.

The prototype for this test evolved from a static HTML version to a PHP scripted prototype web site. Much of the functionality required to test faceted browsing and search suggestions was simulated; the prototype was not actually functional. In this prototype (see Figure 5) the suggestions were moved to the left-hand side of the screen from their location at the bottom in the first prototype, with clearer, more concise



Browse Suggestions

Modify your browse results by limiting by location, genre and/or subject.

Figure 5.
Prototype used in the task-scenario study, which supports faceted browsing with linked suggestions displayed on the left side of the page

labeling. The ability to combine browse categories was also supported. The topical suggestions generated on the browse results page, derived from the TGM I thesaurus, displayed only the broadest term in the TGM I hierarchy to test how users navigate the subject hierarchies. For example, rather than display the specific term used in indexing – for example, "Suits" or "Scarves" – the topical suggestions displayed the broader term, "Clothing & dress."

Ten participants, including Indiana University staff, faculty and students and the general public, provided feedback. The staff participants all had experience working with images in some manner. The remaining participants were not consistent users of online image collections, although one manages a photo album online and the others occasionally visit specific image-oriented web sites. Nine of the ten participants were regular users of the internet.

Seven multi-part tasks were designed to test browse and search functionality in the PHP prototype site. Four of the seven tasks were measured for level of completion: pass easy, pass difficult and fail. The remaining tasks were talk-aloud tasks and therefore were not measured according to benchmarks.

The results of this usability study affirmed that the evolving Cushman prototype was increasingly serving the needs of our users. Most users successfully completed the tasks pertaining to the browse and search suggestions. The problems that did surface were all tied to ambiguous or absent labeling and visual cues in the prototype design.

Two out of ten participants preferred searching as their primary means of access to the Web site. Throughout the evaluation these two participants felt the browse and search suggestions were "in the way," yet they were nonetheless able to complete tasks successfully using their preferred mode of discovery: the search box. The remaining participants felt that the suggestions were helpful. Most agreed that they informed users as to the extent of the collection. Based on these results, we recommended that the final web site support a structured form of browsing and searching with linked suggestions as well as keyword searching and refining so that our users can choose the discovery path they prefer.

Providing subject browsing options only from the highest level of the TGM I hierarchy proved problematic in this test. Participants were asked to find images of botanical gardens in California. They were asked to begin the task by browsing by location, but once in the results page they were free either to search or use the linked suggestions. Those who were comfortable using the suggestions to navigate the content were stumped; they kept looking to no avail for "botanical gardens" or "gardens" in the subject list of suggestions. Despite an attempt by the prototype to indicate visually that a given subject had other terms beneath it in the subject hierarchy, users did not intuitively choose a broader term to navigate down to a narrower one. Because all images of botanical gardens were hierarchically under the browse suggestion "Facilities," eight out of ten participants failed this task. Based on these results, we concluded that the subject search and browse suggestions should provide links to the actual terms used in cataloging and not only to the broadest term in the TGM I hierarchy.

Enhanced interface for discovery

Our research and user testing showed that integrating the TGM I thesaurus structure into the Cushman web application was an effective way to enhance discovery for our users. The functional requirements designed for the application were intended to perform this integration without overwhelming users with thesaurus lingo or overly complex hierarchies. We wanted to ensure that our users would always have suitable discovery options to accommodate varying browse and search styles. Finally, we wanted to structure the use of the TGM I thesaurus to guarantee that users always retrieved results when using the search or browse suggestions.

Browse

The main browse access points into the Cushman collection are: year, genre, subject and location. Users can either browse these categories individually or in combination (known as faceted browsing). Categories with inherent hierarchical structures, such as subject and location, allow navigation down the hierarchy once the facet is added during browsing. Figures 3-5 illustrate a faceted browse. The user must enter from one of the four pre-identified browse categories. In this case, the user has chosen to browse by location (see Figure 6). The results page (see Figure 7) shows images of the chosen location, Innsbruck, Austria. Also provided on the result page are a series of links grouped by the remaining browse categories: year, genre and subject. The user selects "automobiles" from the subject facet (see Figure 8). At this point, she has refined her original result set from 143 images to 27. She could continue to refine by genre or date, or change direction altogether and remove any of active categories to broaden the current result set.

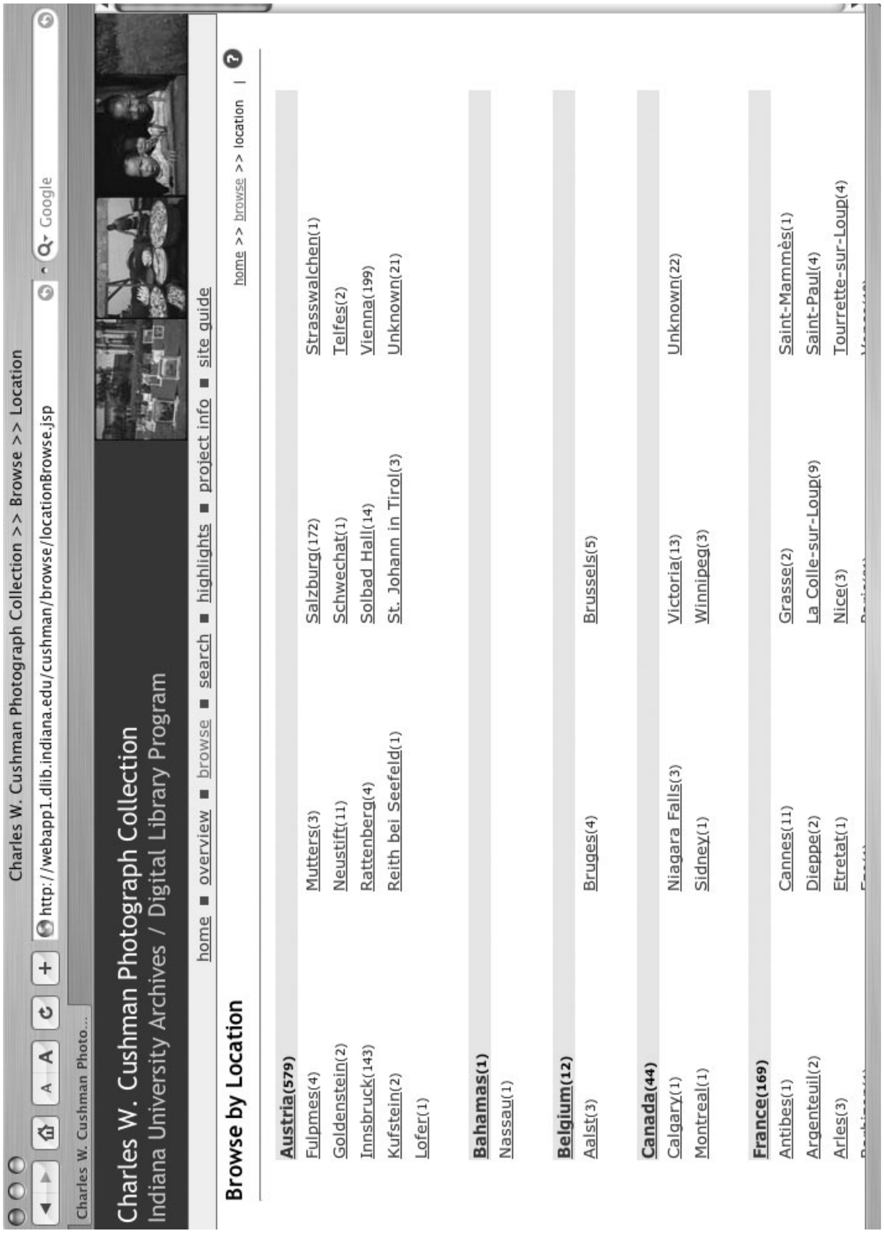
Search

Owing to the extensive descriptions provided by Charles Cushman himself in addition to the TGM I descriptors, an integrated search approach was recommended that sends a user's query to search against both the descriptive fields in our database and the complete TGM I thesaurus (see Figure 9). Should the user's topical search term find a match in the thesaurus, the results page provides search suggestions so that users can broaden or refine a result set based on the TGM I structure. These search suggestions assist users in reformulating their queries in context by allowing for topical term-for-term replacement. For instance, if a user queries for "gardens and California," she will be offered narrower term suggestions (for example, botanical gardens and Japanese gardens) and broader term suggestions (for example, facilities) related to her original term, "gardens." The suggestion selected will replace her original term in the query. The user can always return to her original result set by navigating up or down the hierarchy.

In addition to providing users with suggestions for topical terms entered, our functional specifications indicated that the search engine should perform two automatic behaviors: mapping of lead-in terms to authorized terms and retrieval of all narrower terms. For instance, if a user searches for "cars," which is not the authorized term in TGM I, we can still provide the user with results by mapping the query to the preferred term, "automobiles." If a user queries for images of "sports," she expects to see images that relate to specific sports such as baseball and basketball. Because the most specific term available was assigned to each image, narrower term expansion using the TGM I syndetic structure is necessary to meet this user expectation.

Technical implementation

The Cushman collection web site is a three-layer, database driven web application developed in Java (http://java.sun.com/), using Java Servlet (http://java.sun.com/products/servlet/) technology, Java Server Pages (JSP), (http://java.sun.com/products/jsp/) Java Database Connectivity (JDBC) (http://java.sun.com/products/jdbc/), and the Struts Java Web application framework (http://jakarta.apache.org/struts/). The front end is developed with HTML, CSS, JSP and Struts tags. The back end is an Oracle 9i database. The middle layer is a locally-developed server application on top of Struts. The site is served using the open source Tomcat application server (http://jakarta.apache.org/tomcat/) and Apache HTTP server (http://httpd.apache.org/) software. The photograph images are stored as JPEG files on the web server and delivered via HTTP.



Charles W. Cushman Photograph Collection >> Browse >> Location

http://webapp1.dlib.indiana.edu/cushman/browse/locationBrowse.jsp

Google
Charles W. Cushman Photo...

Charles W. Cushman Photograph Collection
Indiana University Archives / Digital Library Program

home ■ overview ■ browse ■ search ■ highlights ■ project info ■ site guide

Browse by Location

home >> browse >> location |

Austria(579)

Fulpmes(4)	Mutters(3)	Salzburg(172)	Strasswalchen(1)
Goldenstein(2)	Neustift(11)	Schwechat(1)	Telfes(2)
Innsbruck(143)	Rattenberg(4)	Solbad Hall(14)	Vienna(199)
Kufstein(2)	Reith bei Seefeld(1)	St. Johann in Tirol(3)	Unknown(21)
Lofer(1)			

Bahamas(1)

Nassau(1)			

Belgium(12)

Aalst(3)	Bruges(4)	Brussels(5)	

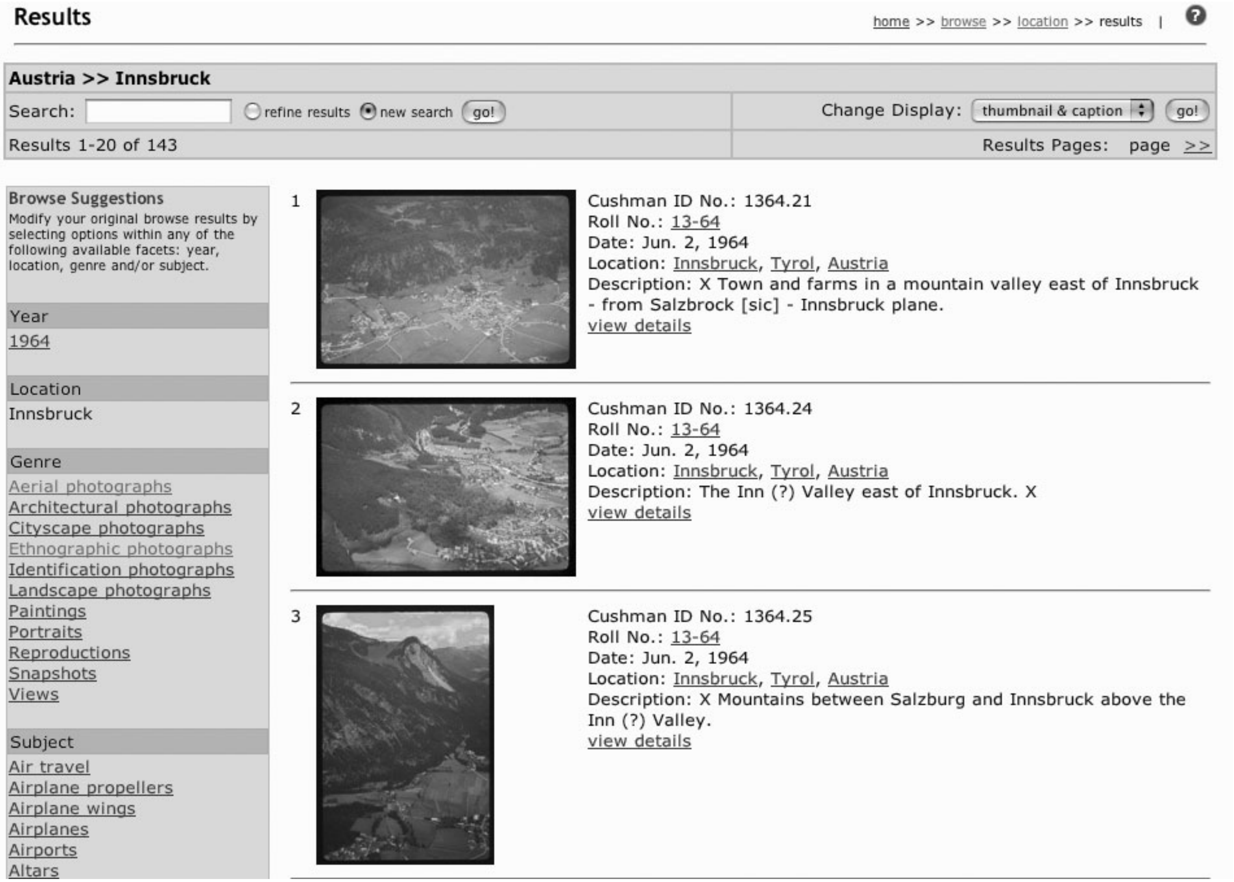
Canada(44)

Calgary(1)	Niagara Falls(3)	Victoria(13)	Unknown(22)
Montreal(1)	Sidney(1)	Winnipeg(3)	

France(169)

Antibes(1)	Cannes(11)	Grasse(2)	Saint-Mammès(1)
Argenteuil(2)	Dieppe(2)	La Colle-sur-Loup(9)	Saint-Paul(4)
Arles(3)	Etretat(1)	Nice(3)	Tourrette-sur-Loup(4)

Figure 6.
All initial browse screens for year, genre, subjects and location serve as menu pages with an organized list of every option under each category. This is Browse by Location



Results

home >> browse >> location >> results | ?

Austria >> Innsbruck

Search: ○ refine results ● new search go!

Change Display: thumbnail & caption go!

Results 1-20 of 143

Results Pages: page >>

Browse Suggestions

Modify your original browse results by selecting options within any of the following available facets: year, location, genre and/or subject.

Year

1964

Location

Innsbruck

Genre

Aerial photographs
Architectural photographs
Cityscape photographs
Ethnographic photographs
Identification photographs
Landscape photographs
Paintings
Portraits
Reproductions
Snapshots
Views

Subject

Air travel
Airplane propellers
Airplane wings
Airplanes
Airports
Altars

1

Cushman ID No.: 1364.21
Roll No.: 13-64
Date: Jun. 2, 1964
Location: Innsbruck, Tyrol, Austria
Description: X Town and farms in a mountain valley east of Innsbruck - from Salzbrock [sic] - Innsbruck plane.
view details

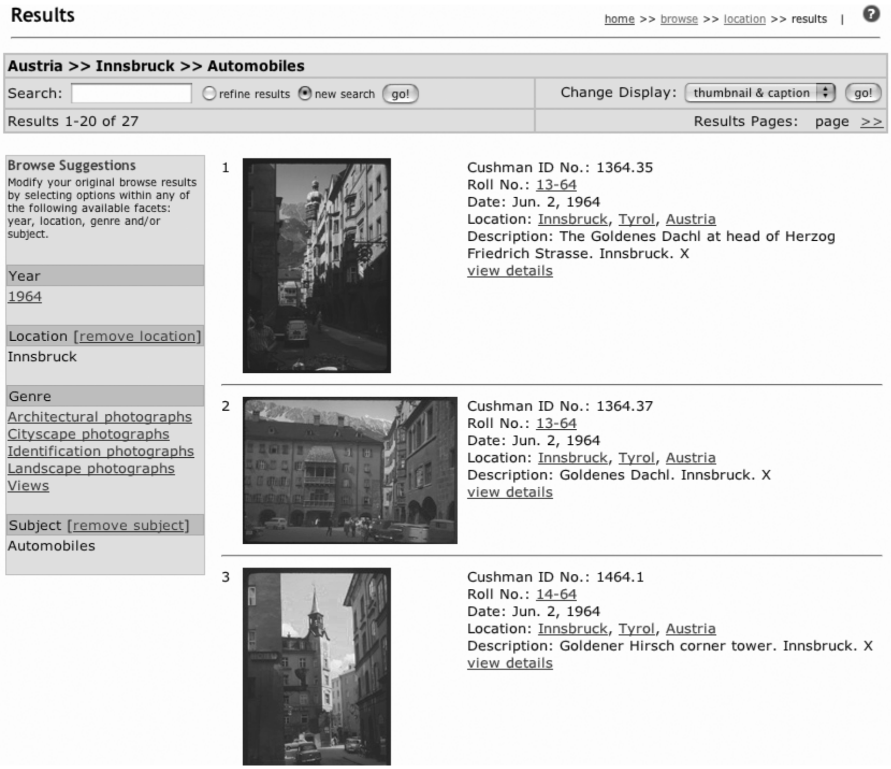
2

Cushman ID No.: 1364.24
Roll No.: 13-64
Date: Jun. 2, 1964
Location: Innsbruck, Tyrol, Austria
Description: The Inn (?) Valley east of Innsbruck. X
view details

3

Cushman ID No.: 1364.25
Roll No.: 13-64
Date: Jun. 2, 1964
Location: Innsbruck, Tyrol, Austria
Description: X Mountains between Salzburg and Innsbruck above the Inn (?) Valley.
view details

Figure 7.
Faceted Browse example: results and browse suggestions for Innsbruck, Austria



Results

home >> browse >> location >> results | ?

Austria >> Innsbruck >> Automobiles

Search: ○ refine results ● new search go!

Change Display: thumbnail & caption go!

Results 1-20 of 27

Results Pages: page >>

Browse Suggestions

Modify your original browse results by selecting options within any of the following available facets: year, location, genre and/or subject.

Year

1964

Location [remove location]

Innsbruck

Genre

Architectural photographs
Cityscape photographs
Identification photographs
Landscape photographs
Views

Subject [remove subject]

Automobiles

1

Cushman ID No.: 1364.35
Roll No.: 13-64
Date: Jun. 2, 1964
Location: Innsbruck, Tyrol, Austria
Description: The Goldenes Dachl at head of Herzog Friedrich Strasse. Innsbruck. X
view details

2

Cushman ID No.: 1364.37
Roll No.: 13-64
Date: Jun. 2, 1964
Location: Innsbruck, Tyrol, Austria
Description: Goldenes Dachl. Innsbruck. X
view details

3

Cushman ID No.: 1464.1
Roll No.: 14-64
Date: Jun. 2, 1964
Location: Innsbruck, Tyrol, Austria
Description: Goldener Hirsch corner tower. Innsbruck. X
view details

Figure 8.
Faceted Browse example: results for automobiles in Innsbruck, Austria

Database design – built for searching

The blueprint used to create a database design for the Cushman web site was a specifications document that defined the desired behaviors for the search and browse features of the site. This document was the result of the rigorous research, prototyping, and usability testing described above, and was essential in creating a database model that could satisfy the goals set forth by the interface and functional designers. In addition to the specifications document, the metadata themselves were carefully analyzed to determine the natural relationships that existed among the descriptive and controlled vocabulary elements.

The resulting Oracle database was designed with an emphasis on search performance, ease of indexing and querying, and the integration of controlled vocabulary features. When designing databases, typical practice is to normalize the model fully using strict database design techniques. The tradeoff to this, however, is that it is often difficult to aggregate data from fully normalized models for text searching or data mining applications. Building indexes and queries for this type of access often requires creating composite views of data in order to search various entities in the model simultaneously. The more complex the data model is, the more complicated the query statements will have to be to traverse tables when indexing and querying. Since text queries usually involve multiple search terms, the query syntaxes become even more complicated if complex statements have to be constructed for each

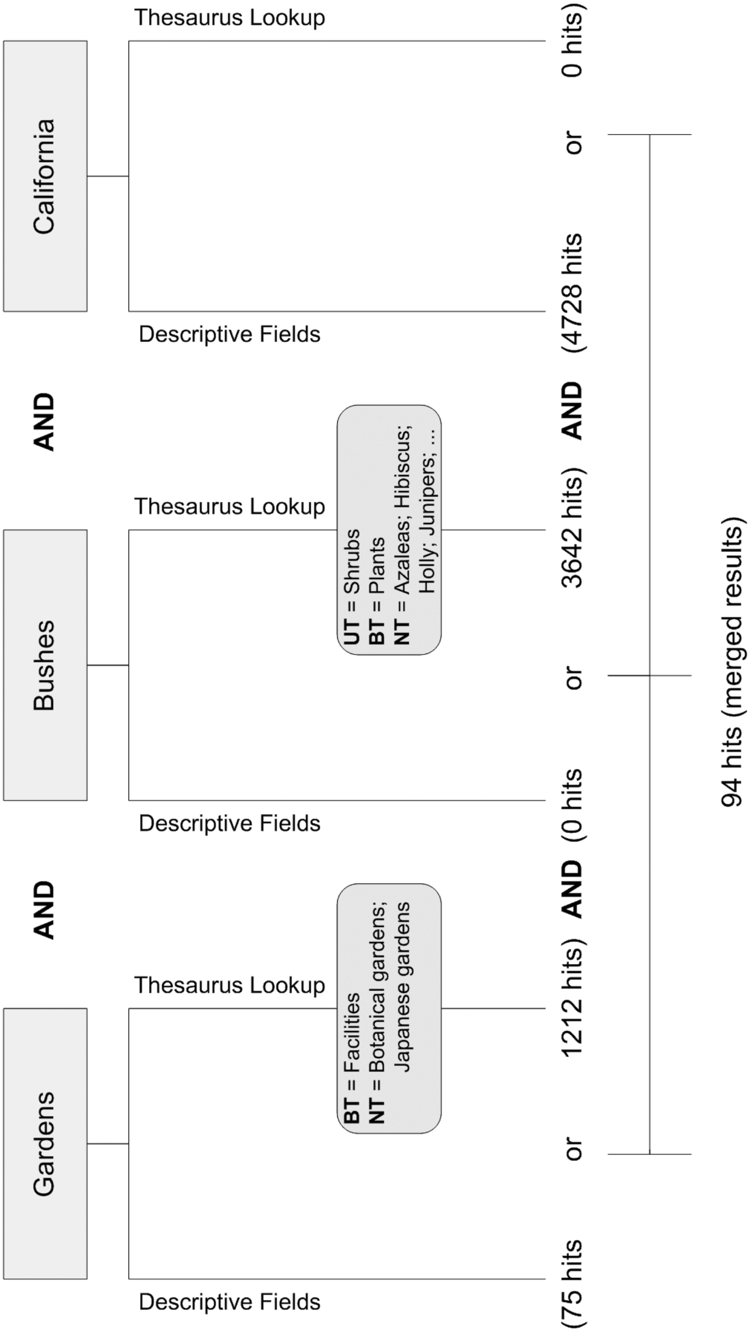


Figure 9.
The integrated query process in which both the descriptive fields in the database and the TGM I fields are searched

term and then further combined together to search all of the desired tables and columns. Such statements are not only difficult to assemble, but could potentially perform poorly.

With this in mind, relationships for the Cushman web site data model were kept simple, and an effort was made to keep related child elements just one database table away from the master image table. The resulting model, illustrated in the Entity Relational Diagram (ERD) shown in Figure 10, is simple, relatively flat, and similar to a typical parent/child model of an XML schema.

In this design, the “cushroll” table represents rolls of film and has a one-to-many relationship with “cushimage,” which represents images. This is a simple XML-like parent/child relationship where images are repeating elements within a roll, and the roll element itself repeats to represent all rolls of film in the collection. Each image element has its own descriptive fields, along with a number of its own repeating child elements used to assign descriptive metadata such as subjects, genres, locations, etc. These relationships are also represented by simple one-to-many relationships in the Oracle database schema, as illustrated above. With this simplified design, text searchable indexes could then easily be designed to represent individual elements or combinations of elements.

Text searching and controlled vocabulary in Oracle

Text searchable indexes with Oracle Text

Oracle’s solution for facilitating text searches in applications is a suite of technologies consisting of special index types, query syntaxes, and stored procedures, known as Oracle Text. The primary purpose of Oracle Text is the ability to define text indexes for columns in tables, which are then consulted when performing case insensitive queries for occurrences of a word or phrase. Without a facility such as this, standard Structured Query Language (SQL) syntaxes have to be used for partial string matching and case conversion, and these result in resource intensive table scans and inadequate query performance.

In this system, Oracle Text indexes exist on each of the metadata elements so that they can be searched individually or in combination as part of an advanced search. There is also an overall keyword index to support the simple search that represents all relevant metadata for each image, regardless of the table or column in which it is stored. To create this index, an Oracle stored procedure was written that gathers and concatenates all child metadata elements stored throughout the relational model that are relevant to a single photograph. When the keyword index is consulted, it is searching in a single location text that was originally scattered throughout the relational model.

To search an Oracle Text index, special keywords supplied as extensions to Oracle’s SQL implementation are used in the predicate part of a selective query, which is the part of the expression that comes after a standard SQL “WHERE” clause to filter records. The most important of these is the “CONTAINS” keyword, which is used to instruct a query to specifically consult an Oracle Text index when searching for a word or phrase. Upon searching for a string in an index using a “WHERE CONTAINS” expression, matching rows are returned in a result set just as they would be in any other query.

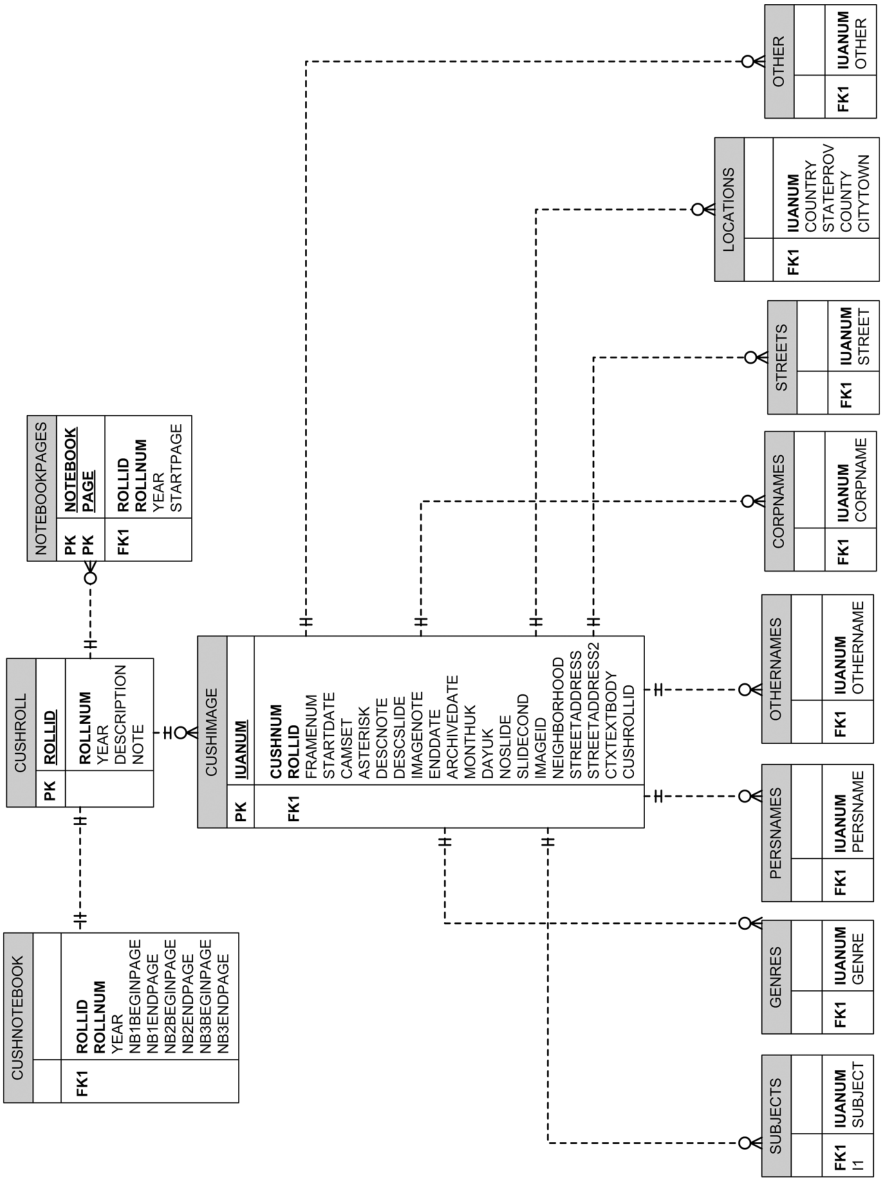


Figure 10.
Entity Relational Diagram (ERD) for the Cushman metadata

Thesaurus support with Oracle Text

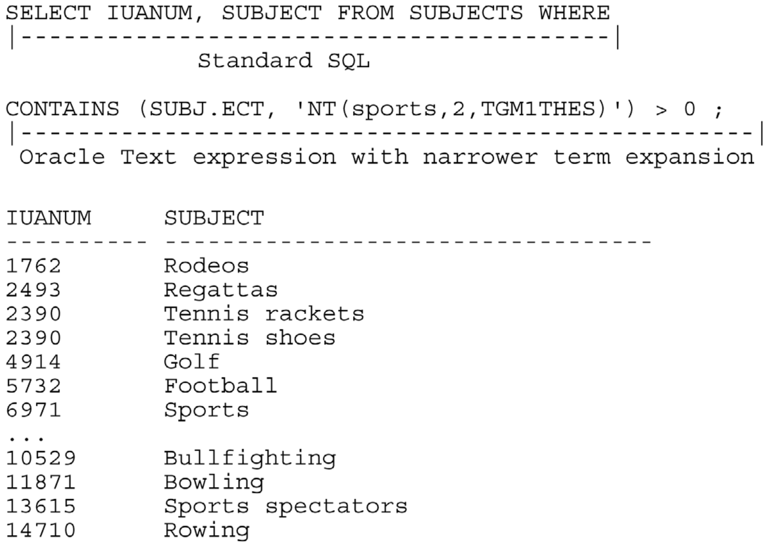
The basic text searching capabilities of Oracle Text were a critical part of developing the search and browse interface for the Cushman Web site, but it was the thesaurus support that made it possible to implement the controlled vocabulary features. Tools within Oracle Text allow for the definition of a custom thesaurus and provide special operators for accessing the thesaurus in SQL queries. As explained previously, the "CONTAINS" keyword is used to search for matching rows using an Oracle Text index. The built-in thesaurus support extends this type of keyword usage to perform dynamic expansion of the search term provided. This expansion is accomplished by looking up the term in a custom thesaurus, retrieving the appropriately related terms, and including them as part of the search string in the original query. Using the features in this manner will return rows where the indexed column contains the original term and/or associated terms in the thesaurus. A custom thesaurus can also be queried and browsed in a stand-alone fashion using stored procedures provided in the Oracle Text package. These can be invoked at any time from within an application, using Oracle Programming Language/SQL (PL/SQL) calls, and will return actual terms in the thesaurus having the specified relationship instead of rows from the database. Special thesaurus operators are available for both methods of access, and have the ability to look up narrower terms, broader terms, related terms, preferred terms, and other relationships conforming to ANSI Z39.19 standards for the construction of thesauri.

The Oracle Text package comes with a default thesaurus, and its content is similar to a basic thesaurus found in modern word processors. Tools are also provided to create and load custom thesauri from user-defined terms and relationships, provided that they conform to a specific format and use certain keyword conventions for relating terms. The TGM I source, available for download (www.loc.gov/rr/print/tgm1/downloadtgm1.html), was converted into the correct format using a Perl script and loaded for use by Oracle Text as a thesaurus.

To illustrate thesaurus access via SQL expansion operators, the query shown in Figure 11 is a simplified example from the Cushman collection, where columns are selected for display from the main image table, an Oracle Text index called "subject" is searched using the special "CONTAINS" keyword, and the "NT" expansion operator is used to find matching rows containing either the string "sports" or its narrower terms (if any).

The sample output in Figure 11 shows that the query produced rows having not only the string "sports" in the subject column, but also rows containing narrower terms as defined in the TGM I thesaurus. If the string "sports" matched a term in the thesaurus, such as "Sports," and several related terms were defined with narrower term relationships pointing to it, then those terms were also included in the search string.

This type of thesaurus access and expansion of narrower terms is used in the Cushman web interface when the simple keyword search is chosen, or when an advanced search is performed on the subject index. However, the faceted browsing features of the site require access to the thesaurus terms themselves and not matching rows. For this kind of access, the PL/SQL functions provided in Oracle Text are used to browse the thesaurus hierarchy to obtain the actual terms having a desired relationship. This functionality is also used to implement the controlled vocabulary



```
SELECT IUANUM, SUBJECT FROM SUBJECTS WHERE
|------------------------------------------|
             Standard SQL

CONTAINS (SUBJ.ECT, 'NT(sports,2,TGM1THES)') > 0 ;
|----------------------------------------------------|
 Oracle Text expression with narrower term expansion

IUANUM     SUBJECT
---------- ----------------------------------
1762       Rodeos
2493       Regattas
2390       Tennis rackets
2390       Tennis shoes
4914       Golf
5732       Football
6971       Sports
...
10529      Bullfighting
11871      Bowling
13615      Sports spectators
14710      Rowing
```

Figure 11.

mapping functions in the Cushman web site. If a user's search term is not an authoritative term used by the catalogers, a simple PL/SQL call using the appropriate Oracle Text function will determine if it has a preferred term in the custom TGM I thesaurus. If the term is already an authoritative term, or if there is no match for it in the thesaurus, then the function merely returns the original term as the result.

Web application development

The web application for the Cushman collection is the middle layer between the front-end interface and the back-end Oracle database. It was built to custom specifications, and performs a number of functions, most notably facilitating and logging searching and browsing.

Java query analyzer

Oracle Text provides powerful text searching and thesaurus support, but it requires a vendor specific SQL syntax and stored procedures to successfully use these functionalities. Therefore, a mechanism for translating end-user queries into syntax appropriate for Oracle Text was needed. While keeping reusability in mind, we developed a configurable query analyzer application as the key module for browsing and searching the Cushman web site.

The query analyzer is used to assist user discovery in three distinct ways. First, it serves as a parser that analyzes the user's input, and translates that input into Oracle specific SQL statements, including syntax to access the TGM I thesaurus. Second, the query analyzer provides a generic interface to handle query texts, so it can be used for both browsing and searching. Finally, the query analyzer offers programming interfaces to facilitate faceted browsing and searching suggestions.

The most important component in the query analyzer is the parser. It accepts the user's query input, and translates Boolean operators, phrases in quotes, grouping with parenthesis, wildcards and field indicators into the SQL syntax required by Oracle Text. This transformation is accomplished through predefined XML-based configuration files. The transformation is defined in a lexical analyzer grammar using Java Compiler Compiler (JavaCC) (https://javacc.dev.java.net/), an open source parser generator and lexical analyzer generator. The parser breaks the user's query into small pieces, uses the configuration files to map these pieces to SQL segments, and then re-combines them for delivery to Oracle Text. When users search on terms in fields in which the controlled vocabulary is not used, the parser creates SQL statements using only the mappings in the configuration file. However, when users search or browse terms as either keyword or subject (which do use thesaurus functions), the parser will call on Oracle stored procedures to determine which of the three conditions apply: if the term entered exists as a preferred term, if it exists as a lead-in term, or if it is absent from the TGM I thesaurus. If the term does not exist in the thesaurus, the term will be passed as-is to an SQL statement. If the term exists as a lead-in term, the parser will fetch both the preferred term and the narrower terms of this preferred term. If the term is a preferred term, the parser will fetch the term's narrower terms. The parser then uses this new list of terms to build the final SQL statement sent to Oracle Text. Thus, the actual search executed includes not only the original query term, but also, if necessary, its preferred term, and all of the narrower terms of the term or its preferred term. This approach performs well, even for terms that have hundreds of narrower terms, such as "People."

The query analyzer also supports replacement, insertion and deletion of terms from searching or browsing query text while maintaining the original structure of the query. These functions, as defined through our user testing experiences described earlier, facilitate search suggestions and faceted browsing within the application. After an initial query, all broader and narrower terms of matched preferred terms are loaded into memory on the server, and are shown as options for an end-user to broaden or refine her search, while keeping the other terms in the original query intact. By maintaining the structure of the original queries during the replacement process, we have made it easier for users to revise and reformulate queries dynamically as they explore the Cushman web site. Dynamic term insertion and deletion also allows the application to support faceted browsing. Adding or removing a facet when a user clicks a browse suggestions link can be accomplished in the web application by simply appending to or removing from the query a term with a field restraint.

Displaying only valid search and browse suggestions

The limitations of TGM I and the way TGM I is used in this project imposed great challenges for the application development. In the database, we use the entire TGM I as the thesaurus for Oracle Text, yet we have only assigned a subset of all terms in TGM I as subject headings to our images. On the web site's browse page, all the subject headings assigned to images in the collection must be listed, as well as all other terms in TGM I which themselves are not among the subject headings assigned, but whose narrower terms are. Therefore all subjects which yield results after narrower term expansion is performed are listed on the subject browse page. Whenever the metadata for the collection is updated, an automated process builds a list of these terms that yield results from TGM I and subject headings in the database, retrieves all records in the database using the terms, and saves the list in an XML file. As a result, in the Cushman web application, there are three lists of terms – the full list of TGM I terms in the Oracle Text thesaurus, subject headings in the subject table in the database, and the list of browsable subjects in a static XML file. The static XML file is transformed into the subject browse page with an XSLT transformation stylesheet.

In addition to its use in generating the subject browse page, this static XML file also helps to filter the search suggestion list on the results page. For example, when the web application generates the search suggestion list for a query of "wrecks," the server first maps the original query to its preferred term, in this case, "Accidents." It then retrieves narrower terms and broader terms for "Accidents." However, many of these terms would return no results because neither they nor any of their narrower terms were used to describe images in the collection. Therefore, we use the list of terms that generate results from the static XML file to filter the list of broader terms and narrower terms. Only terms that exist in both cases will be added to the search suggestion list. This approach successfully ensures that no subject search and browse suggestions links go to zero-hit result sets.

Logging

The Cushman web site is intended not only as a digital image collection, but also as a resource for research on users' searching patterns and their reactions to the application's use of thesaurus relationships. To supplement standard Apache web logs, a logging scheme was developed within the web application to record users' searching or browsing activities with a standard format, which includes an indication of the type of search or browse performed. The logs do not record personal identifiable information. The logging function was developed with the Apache Log4J package (http://logging.apache.org/log4j/docs/index.html), which is an open source Java-based logging library. All logs are in the format:

```
Timestamp [source] query
```

For example, when a user browses for "Animals" on the subject browse page, the URL sends a query of "subject:Animals" to the server. Thus, the log for this query looks like:

```
2004-6-19 12:30:32 [BROWSE] subject:Animals
```

Future work

There are many aspects of the Cushman collection web site and its technical implementation that will be of interest to other institutions that are researching and developing similar features for digital library systems. For the most part, this system uses free and/or open source software, such as Java, Struts, Tomcat, and Apache, and would therefore be easy to share with others. We feel the query parser especially would be useful to other institutions, and will certainly re-use it within our own development environment.

Open source database alternatives

Despite the heavy reliance on open source technology in the Cushman web application, most of the unique features relating to text searching and controlled vocabulary rely heavily upon Oracle and Oracle Text, which are not as pervasive in academic computing when compared with open source databases. For this reason, it might be beneficial to attempt to replicate the controlled vocabulary features using free or open source software solutions. In order to be capable of fulfilling the search and browse requirements for the Cushman web site, a free or open source database solution would have to provide text indexing features. This is possible with both the open source PostgreSQL (www.postgresql.org/) and MySQL (www.mysql.com/) relational database management systems. The current distribution of PostgreSQL includes the Tsearch2 text search module, and MySQL has its own built-in searchable text index type called "FULLTEXT." Both of these are similar to Oracle Text in the way that text indexes are created with specific creation commands and keywords, and also in the way that special syntaxes are used in combination with standard SQL queries to search the resulting index. Implementing the text searching features in either PostgreSQL or MySQL would seem to be largely a matter of minor reengineering and testing, and making appropriate adjustments where necessary.

Recreating the thesaurus and controlled vocabulary features of the system through non-proprietary means, however, presents many technical challenges due to the lack of ready-made solutions in the open source community. Presently, neither PostgreSQL nor MySQL provide thesaurus capabilities in conjunction with their text search features. The Cushman web application design does not necessarily dictate that the thesaurus has to be an integral part of the database and its search engine, so a separate component would be feasible. But, at the time of this writing, there is also little development activity in the realm of stand-alone thesaurus applications from which to borrow.

There are a small number of thesaurus development projects, but they are either still in the early stages of development or are developing in a way that is very specific to a discipline or an institutional need, though they could still be a good starting point to borrow from or contribute to if appropriate. One example of these is the Alexandria Digital Library Thesaurus Protocol/Thesaurus Server (www.alexandria.ucsb.edu/thesaurus/protocol/). This is a Java-based XML/HTTP server implementation that communicates terms and relationship via messages to and from a client. The thesaurus structure itself conforms to ANSI/NISO Z39.19 guidelines.

Another promising avenue for implementing thesaurus support would be in the recently announced Dictionary and Thesaurus API created by IBM and available on IBM's Alphaworks program (www.alphaworks.ibm.com/tech/jadt/). This is a Java API that provides classes for creating and querying custom thesaurus structures using synonyms, acronyms, holonyms, hypernyms and meronyms etc, as defined in a plain text file or an XML file with a certain format. This API, however, does not have support for retrieving preferred terms, broader terms or narrower terms from thesauri, although it does provide a good foundation on which programmers could build such functionality. This API would lend itself well towards creating a standalone Java web application that could facilitate thesaurus lookups for any application.

Reuse of the query analyzer

The query analyzer developed for the Cushman web application was abstracted into a stand-alone component and has already been successfully used in another project within the Indiana University Digital Library Program. In its abstract state, the parser provides support for query parsing, interpretation, and transformation by providing a transparent programming interface to developers. It can translate users' generic queries into standard or vendor-specific SQL statements, XQuery or XPath so we can query native XML databases as well as relational databases. We plan to continue using this parser in other contexts.

Large-scale image repository

As the Indiana University Digital Library Program has grown since its inception in 1997, we have mounted a number of digital image collections. To date, each has operated in its own technical environment, and we currently have no mechanism for cross-collection searching among them. In the very near future, we will be embarking on an initiative to build an image repository for digitized collections at the University. The repository will facilitate archival storage of master images; manage descriptive, technical, and other types of metadata; provide advanced image discovery mechanisms to end-users; offer a standardized delivery interface; integrate collections with teaching and learning applications, including course management software; and facilitate creation of new collections quickly. The user research and development work done for the Cushman web site has provided us with a solid basis on which to begin work on the larger repository.

Conclusion

It is important to note, following this technical discussion, that this work has all been done for the purpose of increasing access to the images in this unique and stunning collection. The sheer breadth of the collection, from farmers in rural Alabama in 1941 (http://purl.dlib.indiana.edu/iudl/archives/cushman/P02551), to the Empire State Building (http://purl.dlib.indiana.edu/iudl/archives/cushman/P11557), to Haight-Ashbury in 1967 (http://purl.dlib.indiana.edu/iudl/archives/cushman/P15512), can take users on extraordinary exploratory journeys. The Cushman site has been both a Yahoo! Pick of the Day (http://picks.yahoo.com/picks/i/20031117.html) and an entry in the Internet Scout Report (http://scout.wisc.edu/Archives/SPT−FullRecord.php?ResourceId = 19123). We hear frequently from users via e-mail and we can study their searches and browses in our application logs. It is, however, through the Web itself that we learn most about the response to the online collection. It has been a frequent citation in Weblogs, including Metafilter (www.metafilter.com/mefi/29540). Weblog entries teach us about users who take images from the site as the basis for making connections to Alfred Hitchcock movies (www.easterwood.org/hmmn/archives/000829.html) or for elaborate storytelling (www.brokentype.com/blog/000210.html#210). These sorts of activities are examples of new uses for image collections in the internet age. While a few favorite images show up frequently in web log entries, for the most part individual bloggers find and comment on images that mean the most to them uniquely. The search and browse suggestions implemented for the Cushman web site are major methods by which these images can be iteratively identified.

The technologies described in this paper enable us to better meet users on their own terms and make the complex nature of image discovery more transparent, allowing users to focus on intellectual exploration and the beauty of the images as visual art. They facilitate use of the Cushman collection far beyond a site with a simple keyword search and no browse feature. Our mission as libraries and archives focuses on providing access to our carefully curated collections. With the Cushman collection, we believe we have created a resource that truly fulfils this goal.

Notes

1. See www.si.umich.edu/Art_History/. This project has been on-going since 1993, with the bulk of the development and research publications peaking around 1995. The functionality currently available to the public has since changed since the site was referenced for the Cushman project in 2003.
2. The method calls for the evaluation of a product in the early design phases by way of individually completing task scenarios that then lead into group discussions about those individual approaches. The group walkthrough method is a hybrid in some respects of a usability lab test and focus group. It takes the positive contributions of each to reveal significant problems in the early development stages of the proposed design. See Bias (1994) for more information.

References

Alexander, A. and Meehleib, T. (2001), "The *Thesaurus for Graphic Materials*: its history, use, and future", *Cataloging & Classification Quarterly*, Vol. 31 Nos 3/4, pp. 189-211.

ANSI/NISO (1993), *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*, Z29.19-1993, NISO Press, Bethesda, MD, available at: www.niso.org/standards/resources/z39-19a.pdf (accessed 28 December 2004).

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

Bias, R.G. (1994), "The pluralistic usability walkthrough: coordinated empathies", in Nielsen, J. and Mack, R.L. (Eds), *Usability Inspection Methods*, John Wiley & Sons, New York, NY, pp. 65-78.

Choi, Y. and Rasmussen, E. (2002), "Users' relevance criteria in image retrieval in American history", *Information Processing and Management*, Vol. 38, pp. 695-726.

Choo, W., Detlor, B. and Turnbull, D. (2000), "Information seeking on the web: an integrated model for browsing and searching", *First Monday*, Vol. 5 No. 2, available at: http://firstmonday.org/issues/issue5_2/choo/index.html (accessed 28 December 2004).

Dubois, C.P.R. (1984), "The use of thesauri in online retrieval", *Journal of Information Science*, Vol. 8 No. 2, pp. 63-6.

Dubois, C.P.R. (1987), "Free text vs. controlled vocabulary: a reassessment", *Online Review*, Vol. 11 No. 4, pp. 243-53.

Fidel, R. (1991a), "Searchers' selection of search keys: I. The selection routine", *Journal of the American Society for Information Science*, Vol. 42 No. 7, pp. 490-500.

Fidel, R. (1991b), "Searchers' selection of search keys: II. Controlled vocabulary or free-text searching?", *Journal of the American Society for Information Science*, Vol. 42 No. 7, pp. 501-14.

Greenberg, J. (2001a), "Automatic query expansion via lexical-semantic relationships", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 6, pp. 402-15.

Greenberg, J. (2001b), "Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 6, pp. 487-98.

Greenberg, J. (2004), "User comprehension and searching with information retrieval thesauri", *Cataloging & Classification Quarterly*, Vol. 37 Nos 3/4, pp. 103-20.

Harpring, P. (1999), "How forcible are right words! Overview of applications and interfaces incorporating the Getty vocabularies", *Archives & Museum Informatics: Museums and the Web*, 12-14 March, available at: www.archimuse.com/mw99/papers/harpring/harpring.html (accessed 28 December 2004).

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K. and Yee, K. (2002), "Finding the flow in web site search", *Communications of the ACM*, Vol. 45 No. 9, pp. 42-53.

Soergel, D. (1999), "Indexing and retrieval performance: the logical evidence", *Journal of the American Society for Information Science*, Vol. 45 No. 8, pp. 589-99.

Tudhope, D., Alani, H. and Jones, C. (2001), "Augmenting thesaurus relationships: possibilities for retrieval", *Journal of Digital Information*, Vol. 1 No. 8, available at: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Tudhope/ (accessed 28 December 2004).

Tudhope, D., Binding, C., Blocks, D. and Cunliffe, D. (2002), "Compound descriptors in context: a matching function for classifications and thesauri", *International Conference on Digital Libraries: Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries, July 14-18*, ACM Press, New York, NY, pp. 84-93.

Visual Resources Association (2004), "Cataloguing cultural objects: a guide to describing cultural works and their images", available at: www.vraweb.org/CCOweb/ (accessed 28 December 2004).

Further reading

Bawden, D. (1993), "Browsing: theory and practice", *Perspectives in Information Management*, Vol. 3, pp. 71-85.

Choi, Y. and Rasmussen, E. (2003), "Searching for images: the analysis of users' queries for image retrieval in American history", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 6, pp. 498-511.

Yee, K., Swearingen, K., Li, K. and Hearst, M. (2003), "Searching and organizing: faceted metadata for image search and browsing", *Conference on Human Factors in Computing Systems: Proceedings of the Conference on Human Factors in Computing Systems, April 5-10*, ACM Press, New York, NY, pp. 401-8.