Cushman Exposed!

Exploiting Controlled Vocabularies to Enhance Browsing and Searching of an Online Photograph Collection

Michelle Dalmau, mdalmau@indiana.edu
Jenn Riley, jenlrile@indiana.edu
IU Digital Library Program

Brown Bag Series

Overview

- Introduction
- Metadata
- Research Overview
- Usability Findings
- Browse and Search Specifications
- Implementation
- Lessons Learned

The Cushman Collection

- Funded with an Institute of Museum & Library Services (IMLS) grant
- ~14,500 color slides taken between 1938-1969
- Held at the IU University Archives
- Site launched October 2003 and March 2004

Looking Back

- U.S. Steel Gary Works Photograph Collection
    - ~2,200 Images
    - Archival descriptions
    - Assigned subject terms from CV
- Subject field search requires referencing the A-Z list of subjects
- Usability studies revealed not using the CV's syndetic structure impacts searching

Metadata for Image Collections

- Advantages to “free-text” descriptions:
  - Preserve photographer’s notations
  - Resembles the user’s language
- Advantages to CV descriptors:
  - More access points
  - Collocation
  - Disambiguation
  - Interoperability

Metadata for the Cushman Collection

- Cushman's description in notebooks and slide mounts
- Dates
- Location
- Names
- TGM I – LC Thesaurus for Graphic Materials: Subject Terms
- TGM II - LC Thesaurus for Graphic Materials: Genre & Physical Characteristics
- TGN – Getty Thesaurus of Geographic Names

TGN: Getty Thesaurus of Geographic Names

- Online browser available
- Data available for licensing for incorporating into a local system
- Current and historical place names
- Hierarchically organized
- Useful as research tool and as structured CV
- Cushman cataloging
- Cushman display

TGM II: Genre and Physical Characteristics Terms

- Online and free downloadable versions available
- Contains over 600 terms
- Poly-hierarchically organized
- We only used 24 TGM II terms
- Multiple genres assigned when appropriate
- More appropriate than AAT for our generalist users
- Cushman cataloging
- Cushman display

TGM I: Subject Terms

- Online and free downloadable versions available
- Contains over 6,300 terms
- Hierarchically organized
- Includes terms for what picture is OF (eg dogs) plus what picture is ABOUT (eg democracy)
- Cushman cataloging
- Cushman display

TGM I: Subject Terms Strengths and Weaknesses

- Strengths include:
  - Pre-defined relationships between concepts
  - Some lead-in vocabulary
- Weaknesses include:
  - Complete syndetic relationships lacking, especially for new terms
  - Language not user-friendly
  - Not enough lead-in vocabulary
  - Form and number of top-level categories not useful for a browse structure

Searching Image Collections: Research Shows

- Complement free-text with controlled vocabulary searching (Fidel, 1991)
- Image retrieval is heavily based on textual labels (Choi & Rassmussen, 2003)
- Query expansion methods based on the CV relationship structures can increase access (Greenberg, 2001/2002)
  - Automatic Expansion: Synonyms and Narrower terms are good candidates for automatic retrieval
  - Interactive Expansion: Broader, Narrower and Related terms are good candidates for user-directed, “manual” retrieval
- Search assistants are helpful (Harping, Getty, 1999)
  - Integration of Getty vocabularies (“a.k.a” and ARThur)

Browsing Image Collections: Research Shows

- Browsing is exploratory – it fosters new connections, innovative use of resources and the ability to easily pursue new paths (Bawden, 1993)
- Browsing is a significant part of image discovery (Choi & Rasmussen, 2002)
- Guided, flexible browsing in context works (Flamenco and SI Art Image Browser projects)

Usability Methods

- Group Walkthrough (prototype excerpt)
  - Paper-based tasks and prototype evaluation
  - 4 participants (mostly librarians)
- Individual Walkthrough
  - Interview and prototype evaluation
  - 2 participants (faculty)
- Task Scenarios (prototype excerpt 1 & 2)
  - On-site task-based testing (14 tasks)
  - 12 participants (staff, students and faculty image users)

Usability Findings Show

- Searching
  - Referencing an A-Z list with no lead-in terms for searching is NOT helpful at all
  - Concerns about word choice (US, USA or America?)
  - Iterative reformulation of queries in context is desired
  - Relevant suggestions are helpful

Usability Findings Show

- Browsing
  - Structure is important
  - Contents should be easily exposed
  - Flexible and combinatorial browsing is desired
  - Browsing cultivates searching

Implementation Specifications

- Search
  - Mapping from lead-in vocabulary
  - Retrieval of all records with narrower terms
  - Integrated search against BOTH “free-text” descriptions and thesaurus
  - User-initiated broadening and narrowing

- Browse
  - Year
  - Genre
  - Subjects (hierarchical)
  - Access via assigned headings with ability to move up and down (pending user studies)
  - Location (hierarchical)
  - Combination of facets

Implementation of the Cushman Web site

- Java using Java Servlet and Java Server Pages (JSP)
- HTML / CSS for interface display
- Oracle 9i, Release 2 database
  - Oracle Text
- Tomcat and Apache HTTP servers
- JPEG images served from file system (PURLS)

Thesaurus-Enhanced Browsing & Searching: Oracle Text

- Link to existing thesaurus or define custom thesaurus
  - Preferred terms
  - Broader terms
  - Narrower terms
  - Related terms
- SQL syntax for using thesaurus to expand database query
- PL/SQL stored procedures for getting information from thesaurus itself

Challenges Using Oracle Text

- Preferred term matches multiple lead-in terms
  - Crops USE Farming; USE Plants
- Phrase matching
  - Military finds Military officers, Military uniforms, etc.
- Qualifiers
  - Cranes vs. Cranes (Birds)
- Punctuation used in TGM terms

Lessons Learned

- Approach to metadata needs to be well-planned and flexible
- Metadata quality control is essential
- Need more data on how people use images
- This stuff is HARD!

But It’s Worth the Effort!

- Enhanced discovery
- Innovative implementation for a production-level collection
- People love the Cushman Collection!

Looking Forward

- Strive to make our collections truly accessible even if only incrementally
- Sustainability of the Cushman approach
- Defining functionality for future image repository for all of our collections

References

- Bawden, D. (1993). Browsing; theory and practice. *Perspective in information management*, 3 (1): 71-85.
- Choi, Youngok and Rasmussen, Edie M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing and Management*, 38: 695-726.
- Choi, Youngok and Rasmussen, Edie M. (2003). Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History. *Journal of the American Society for Information Science and Technology*, 54 (6): 498-511.
- Fidel, Raya. (1991). Searcher's selection of search keys: Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science, 42 (*7): 501-514.

References (con't)

- Greenberg, J. (2001). Optimal QE Processing Methods with Semantically Encoded Structured Thesauri Terminology. *Journal of the American Society for Information Science and Technology*, 52 (6): 487-498.
- Harpring, Patricia. (1999). How forcible are right words!: Overview of applications and interfaces incorporating the Getty vocabularies. *Archives & Museum Informatics*: http://archimuse.com/mw99/papers/harpring/harpring.html
- Hearst, Marti et al. (2002). Finding the flow in web site search. *Communications*, 45: 42-53.
- University of California, Berkeley: Flamenco Project -- http://bailando.sims.berkeley.edu/flamenco.html
- University of Michigan: SI Art Image Browser -- http://www.si.umich.edu/Art_History/

Shout Out!

- Thanks to the Cushman Team comprised of Archives and DLP members especially . . .
    - Randall Floyd (Database Guru)
    - David Jiao (Java Genius)