

Collection-level search analysis

Oksana Zavalina
IMLS Digital Collections and Content project
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

2007 LITA national forum

Subject Representation in the IMLS DCC Collection Registry

- Gateway to Educational Materials (GEM) subject headings (at least one GEM Topic Terms heading is required)
- Optional -- subject headings from alternative scheme(s) (e.g., LCSH, AAT, locally-developed)
- Geographic coverage headings (use of [Getty Thesaurus of Geographic Terms](#) is strongly recommended).

Subject Representation in collection records

The screenshot shows a web browser window with the following content:

Collection Information

TITLE: New York Public Library's Picture Collection Online

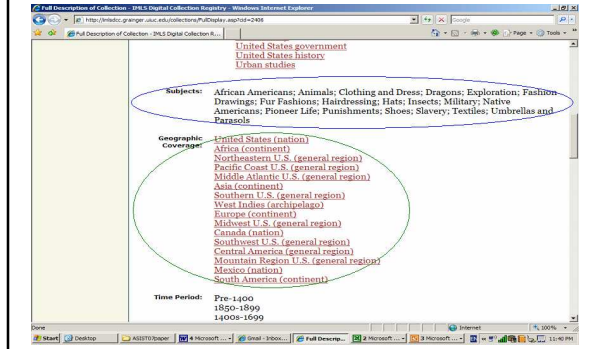
URL: <http://picturecollection.nypl.org>

Description: The Picture Collection Online is an image resource site for those who seek knowledge and inspiration from visual materials. It is a collection of 50,000+ digitized images from books, magazines and newspapers as well as original photographs, prints and postcards, mostly created before 1929.

GEM Subjects:

- Arts
 - History of art
 - Photography
 - Popular culture
 - Visual arts
- Language Arts
 - Alphabet
- Science
 - Biology
 - Entomology
 - History of science
 - Natural history
- Social Studies
 - Anthropology
 - Native history

Subject Representation in collection records



Research questions

- What are the quantitative characteristics of the search session and typical user query in the IMLS DCC Collection Registry?
- What are the typical search categories in the IMLS DCC Collection Registry?
- What is the distribution of the two major search types (known-item and subject search) in the IMLS DCC Collection Registry?

Research questions (cont'd)

- How suitable are GEM Topic Terms for describing diverse collections in the IMLS DCC Collection Registry compared to alternative controlled vocabularies?
 - semantic similarity measures
 - user keywords extracted from transaction logs
 - subject terms in 3 different controlled vocabularies — GEM, Library of Congress Subject Headings (LCSH), and Art and Architecture Thesaurus (AAT).

Dataset

- Transaction log
 - MS Access file
 - 7 months-worth of searches
 - 19,000 records
 - 936 keyword search query strings

Dataset

- Minimal data processing:
 - exclusion of web crawlers' automatic queries and searches made by the Collection Registry testers (number of records reduced from 111,000 to 19,000)
 - keyword query string extraction
 - morphological variants
 - no query parsing
 - stop words: prepositions, conjunctions and articles
 - grouping exact same and morphologically variant queries into *unique search term* (682)

Methods

- Transaction log analysis:
 - descriptive statistics:
 - Keyword search and subject browsing
 - Query length
 - Frequency of query use
 - qualitative content analysis:
 - assigning user keyword queries to 10 search categories
 - assigning user keyword queries to 2 search types
 - matching user keyword queries with terms in 3 controlled vocabularies

Methods: 10 search categories

- 7 FRBR entities:
 - **work** [collection as a work; any intellectual or artistic creation that has a title attribute]
 - *(individual) person*
 - **corporate body**
 - **concept**
 - **object**
 - **event**
 - **place**
- FRANAR/FRAD entity
 - **family**
- Additional categories:
 - **class of persons** (e.g., "abused children", "prisoners")
 - **ethnic/national group** ("Irish Americans", "Sioux Indian")
 - **unknown**

Methods: 2 search types

- searches where the user queries either the title or the author — individual or corporate — of a digital collection belong to **collection-level known-item** search type;
- all the other searches in the Collection Registry belong to a widely defined (both controlled and uncontrolled-vocabulary) **collection-level subject** search type.

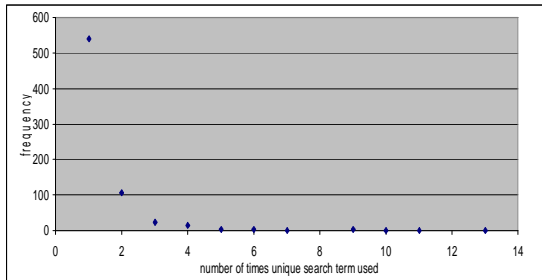
Methods: similarity measures

- exact matches
- synonymous matches (semantic variants)
- near-exact matches:
 - syntactic variants
 - morphological variants
 - acronyms
- NO broader and narrower terms matches

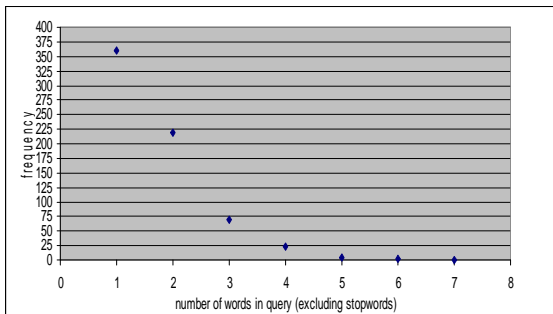
Findings: Keyword Search and Subject Browsing in IMLS DCC Collection Registry

- 476 sessions include keyword searching or subject browsing
- 1 to 18 keyword queries per session (average of 2)
- Keyword search queries represent 4.77% of all queries in the Collection Registry
- Subject browsing represents 7.71% of all queries in the Collection Registry

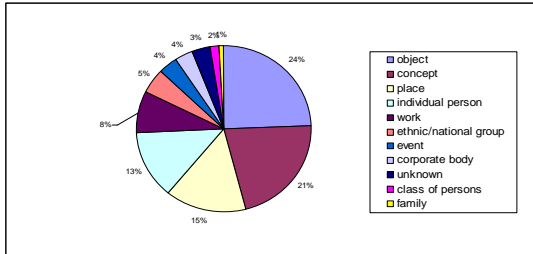
Findings: typical user query



Findings: typical user query

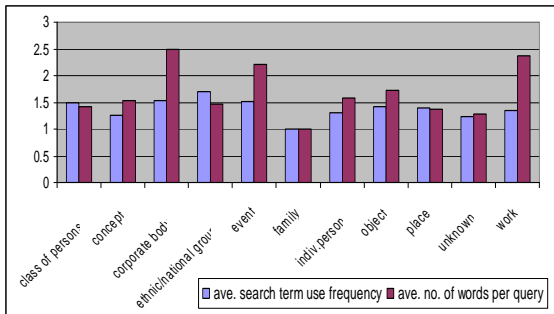


Findings: search categories

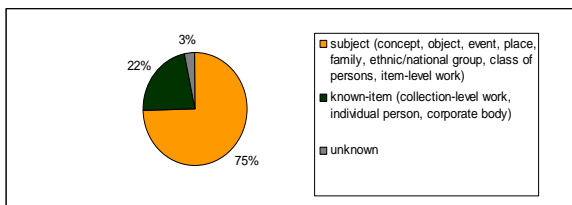


- Object, concept, place, individual person are heavily used
- surprisingly low level of event searching (4%)

Findings: typical user queries by search categories



Findings: search types



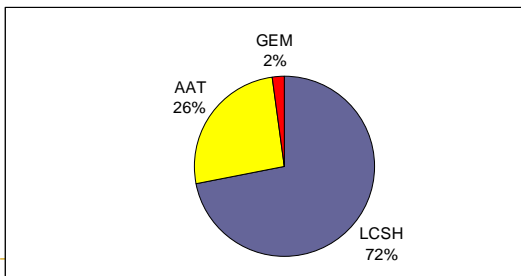
Findings: search types

- Prevalence of subject search
- Higher than usually observed level of subject searching
- Possible reasons:
 - general shift towards subject search in modern world
 - conceptual difference between collection-level and item-level search
 - other?

Findings: semantic similarity

search category	unique terms	GEM match	GEM match, %	LCSH match	LCSH match, %	AAT match	AAT match, %
concept	146	15	10.27	127	86.99	85	58.22
corporate body	24	0	0	17	70.83	0	0.00
event	25	0	0	9	36.00	3	12.00
object	166	0	0	118	71.08	69	41.57
class of persons	12	0	0	10	83.33	7	58.33
ethnic/nat'l group	33	0	0	29	87.88	15	45.45
family	5	0	0	4	80.00	0	0.00
individual person	90	0	0	72	80.00	0	0.00
place	103	0	0	98	95.15	0	0.00
work	56	0	0	7	12.50	0	0.00
unknown	22	0	0	4	18.18	0	0.00
TOTAL	682	15	2.20	495	72.58	179	26.25

Semantic similarity between collection-level user search terms and terms in three controlled vocabularies



Conclusions

- Unusually high for catalog use / transaction log analysis studies level of subject searching
- Strong semantic match to user queries offered by a traditional library subject scheme — Library of Congress Subject Headings
- Combination of two or more standardized controlled vocabularies for collection-level subject description may be beneficial for Collection Registry.

Further research

- Reasons of subject search prominence (interviews and observations of the Registry users)
- User conceptualization of the collection-level search and its possible difference from the concept of the item-level search
- Investigate more flexible than LCSH moderate-scale controlled vocabularies, which, unlike GEM or AAT, represent a wide variety of search categories
