PROJECT MUSE®

Finding Images in an Online Public Access Catalogue:
Analysis of User Queries, Subject Headings, and Description
Notes / Le repérage d'images dans un catalogue en ligne à
accès libre : analyse des requêtes des utilisateurs, des
vedettes-matière, et des notes descriptives

Youngok Choi, Ingrid Hsieh-Yee

Canadian Journal of Information and Library Science, Volume 34,
Number 3, September/septembre 2010, pp. 271-295 (Article)

Published by University of Toronto Press

DOI: <https://doi.org/10.1353/ils.2010.0004>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/390926>

Finding Images in an Online Public Access Catalogue: Analysis of User Queries, Subject Headings, and Description Notes

Le repérage d'images dans un catalogue en ligne à accès libre : analyse des requêtes des utilisateurs, des vedettes-matière, et des notes descriptives

Youngok Choi* and Ingrid Hsieh-Yee
School of Library and Information Science
Catholic University of America
620 Michigan Avenue N.E.
Washington, DC 20064
*Corresponding author

Résumé : Cette étude examine les requêtes d'utilisateurs à la recherche d'images dans un catalogue en ligne à accès libre et la façon dont les vedettes-matière et les notes descriptives de la Library of Congress ont été utiles à ces requêtes. Il apparaît que les auteurs de recherches utilisent peu les opérateurs booléens; que le domaine thématique et le format ont eu une influence sur le développement et la reformulation des requêtes; et que ni les vedettes-matières ni les notes descriptives n'ont servi utilement aux requêtes des utilisateurs. L'analyse des termes et des concepts utilisés dans les requêtes des utilisateurs a révélé que ceux-ci ont tendance à chercher selon des concepts thématiques et selon le format physique des images. Les concepts thématiques utilisés dans leur recherche par les utilisateurs finaux et ceux utilisés par les indexeurs professionnels se sont avérés similaires, mais les deux groupes ont utilisé des vocabulaires différents. Suit une discussion des implications de ces différences de vocabulaire pour l'indexation et la recherche d'images.

Mots-clés : analyse des requêtes d'images, recherche d'images, extraction d'images, indexation des images, vedettes-matière LC, notes descriptives pour l'extraction d'images, catalogue en ligne

Abstract: This study examined user queries for images in an online public access catalogue and how Library of Congress subject headings and description notes supported such queries. It found searchers made light use of Boolean operators; subject domain and format seemed to influence query development and reformulation; and neither subject headings nor description notes supported user queries well. Analysis of terms and concepts in user queries revealed that end users tended to search by subject concepts and the physical format of images.

While the subject concepts used by end users for searching and the subject concepts used by professional indexers for indexing were found to be similar, different vocabularies were used by these two groups. The implications of vocabulary differences for image indexing and searching were discussed.

Keywords: image query analysis, image search, image retrieval, image indexing, LC subject headings, description notes for image retrieval, online catalog

Introduction

Images are increasingly important for research as well as for entertainment in our society. Many digital image databases, image search engines, and commercial stock-image websites offer access to images on the Web. In addition, libraries, archives, museums, and other cultural institutions digitize image collections for user access (Palmer et al. 2007). Images can be retrieved by keywords that describe objects and themes presented in images (such as *boat, love, peace*). The objects and themes are referred to in the literature as “concepts,” which are important for image retrieval, and this type of retrieval is categorized as “concept-based” image retrieval (Attig, Copeland, and Pelikan 2004; Enser 2000; Frost et al. 2000). On the other hand, images can be retrieved by visual attributes such as colour, shape, texture, and other image properties, and this type of retrieval is categorized as “content-based” image retrieval (Enser 2008; Smeulders et al. 2000). Content-based image retrieval, however, has had limited success (MacDonald and Tait 2003) and searching images by words remains the preferred method (Eakins, Briggs, and Burford 2004). As a result, search engines, digital image collections, and online catalogues continue to rely heavily on words to retrieve images for users.

Representing image content in words is challenging because images often lack textual clues and bear many different meanings (Panofsky 1962; Rafferty and Hilderley 2007; Shatford-Layne 1994). In addition, an image has attributes associated with its creation as well as subjects, genre, and physical production. The complex nature of images has posed special challenges for image indexing. Cataloguers typically rely on subject headings and textual notes to represent subject content and describe image attributes. Key subject matters of an image are represented by controlled vocabularies, keywords, or content notes, while additional descriptive metadata such as title and creator are added. As a result, the success of end-user image retrieval depends on how well subject headings and description notes in cataloguing records match user search queries. In

order to represent subject content of images to support end-user searching, image cataloguers and indexers need to understand which kinds of semantic content—of an image (such as people, time, event) or about an image (such as the context of its creation)—affect users' search decisions, and which aspects of these images influence interpretation of the image. It was reported that queries reflect elements of special interest to users (Aula 2003), so a potentially fruitful approach to understanding end users' image search behaviour is to focus on their queries. The purpose of this study was to examine characteristics of user search queries for image retrieval and to compare the retrieval effectiveness of subject headings and description notes in supporting user image queries in an OPAC system.

Research questions

The study was designed to address four research questions:

1. What are the characteristics of image query terms and query formulation?
2. How effective are Library of Congress subject headings in matching image queries in an online catalogue?
3. How effective are subject description notes in matching image queries in an online catalogue?
4. To what extent are concepts in image queries represented by Library of Congress subject headings and subject description notes?

Related studies

In an effort to understand how users search for images, researchers have focused on the characteristics of users' description of images (such as Armitage and Enser 1997; Chen 2001; Choi and Rasmussen 2003; Collins 1998; Eakins, Briggs, and Burford 2004; Fidel 1997; Hasting 1995; Jorgensen 1998). These studies have found that when users look for images, they often use terms of "ofness"—that is, terms describing objects (such as a baby) in an image rather than what an image is about ("aboutness") (Shatford-Lane 1994), and they sometimes qualify their queries with period, location, action, and other aspects. This pattern was also observed in studies of digital image retrieval. For example, Frost

et al. (2000) found that subject matter or concepts in images were more important than names of artists, and users tended to conduct a subject search when looking for images. The study by Yoon (2006) found that denotative (objects in an image) and connotative attributes (abstract, subjective, or interpretive meanings in an image) were important across the overall search process and users employed diverse denotative and connotative terms to find a satisfactory image for a given task, while colour itself did not have critical impact during representation and selection. These studies underscore the importance of objects and concepts embedded in images for image retrieval. Perceptual characteristics of an image seem to be of less value for searchers.

Users may express their interest in a query format in a way different from simply describing images in natural words. In an experiment to examine users' queries when searching for imaginary pictures based on textual description without images present, Hollink et al. (2004) found that participants used more specific terms and less abstract and perceptual terms in a query than in an image description. They also discovered that subject domain of an image influenced image description in natural language and query construction in a keyword-based search engine. Other researchers observed differences when users described images with keywords and in sentences (Greisdorf and O'Connor 2002; Laine-Hernandez and Westman 2006; Rorissa 2008). These studies suggest that query construction is different from natural description. It is therefore useful to examine user search queries to identify elements and concepts that are important to searchers.

Several studies have focused on image searching and search behaviour on the Web (Fukumoto 2004; Goodrum, Bejune, and Siochi 2003; Goodrum and Spink 2001; Jorgensen and Jorgensen 2005; Pu 2005). Their findings suggest that Web users type in short queries not only when searching for textual information, but also when searching for visual information, and query modification is important in image searching (Goodrum and Spink 2001; Jorgensen and Jorgensen 2005). User queries differ between search engines targeted primarily at the public and image sites for image experts. Pu (2005) found that users' image queries tended to focus on unique searches such as personal name searches rather than in textual queries. These studies suggest that there is a difference in search interests and thus query types between textual search and image search.

It is known that subject indexing or textual description supports image retrieval. Choi and Rasmussen (2002) found that textual information associated with an image was important for relevance assessment; also, date, title, notes, and subject descriptors were considered most helpful. In Hung's (2005) study, journalism students relied on textual descriptions when judging relevance in general and subjective image search tasks. Matusiak's study (2006) found something similar in that the majority of participants discovered that hyperlinked subject headings within records enabled them to find more images on the listed topic.

Previous studies show that specific objects in an image and subject concept represented in an image are important for image searching. However, there has been little research on the extent to which subject headings and indexed textual content notes support users' image search. In short, studies of user queries for image searches and comparison of queries to subject description are helpful for providing insights into users' expression for visual material searches. Findings of such studies will suggest guidance for development of searching tools and indexing process.

Methodology

The study used SurveyMonkey to create a simulated search environment and collected user queries with a Google-like search box. Thirty-three students and recent graduates of a library and information science program responded to listserv invitations and took part in the study. To shed light on how users developed queries to search for similar images, two target images—a photograph of immigrants in Ellis Island and a picture of a cancer-awareness poster—were selected from the Library of Congress Prints and Photographs Online Catalog. For each target image, participants were given a brief scenario and five spaces to record up to five search queries they would use to look for similar images (see figures 1 and figure 2 for screenshots of images with scenarios). The study selected two target images because digital image collections developed by libraries contain images of cultural or historical significance, and bibliographic records of such digital images are often integrated into online public access catalogues (OPACs) to facilitate search and access. The two images were selected intentionally from different subject domains (medicine and history) and formats (photo and poster) in order to determine if subject domain and physical format were related to user queries and query formulation. Earlier transaction log analysis of user queries found

Figure 1: Screenshot of a scenario and a photograph of immigrants on Ellis Island

Figure 2: Screenshot of a scenario and a poster about cancer awareness

that the average number of queries per search session was 3.36 queries (Goodrum and Spink 2001) and 2.1 (Jorgensen and Jorgensen 2005). With these numbers in mind, the study asked participants to provide up to five queries they would use to complete the search. To remove potential interference of the search interface in query formulation, participants submitted queries without any interaction with an online catalogue. It is acknowledged that using a convenience sample and only two selected images limits the generalizability of the study findings.

After completing queries, the 33 participants reported their familiarity with each search topic on a 5-point scale (1 = very low and 5 = very high). All participants had low familiarity with the two search topics. The average level of familiarity was 2.03 for the immigration history photograph and 1.77 for the cancer poster, suggesting participants were novices in the given domain areas. They also reported limited cataloguing and reference service experience, with about 86% having no cataloguing experience or via coursework only, and 53% with no reference experience or with coursework only.

The study adopted a method of query analysis used in Jorgensen and Jorgensen's study (2005) to examine characteristics of query formulation (see tables 1 and 2). Since the 2005 study focused on queries of image professionals and this study focused on queries of non-experts, the application of their query analysis method helped to shed light on the differences between these two groups in how they searched for images.

Table 1. Types of query terms

Immigration History Photo	Queries	%	Cancer Poster	Queries	%
Proper nouns	87	26.05	Concept terms	128	49.81
Noun	68	20.36	Visual construct	77	29.96
Visual construct	66	19.76	Proper nouns	36	14.01
Concept terms	58	17.37	Adjective	4	1.56
Adjective	20	5.99	Noun	3	1.17
Date terms	17	5.09	Unknown	3	1.17
Verb	11	3.29	Verb	2	0.78
Unknown	7	2.10	Date terms	2	0.78
Total	334	100		257	100

Table 2. Query formulation strategies

	Immigration History Photo				Cancer Poster				
	Q1→Q2 (N = 31)	Q2→Q3 (N = 29)	Q3→Q4 (N = 14)	Q4→Q5 (N = 3)	Q1→Q2 (N = 30)	Q2→Q3 (N = 28)	Q3→Q4 (N = 16)	Q4→Q5 (N = 7)	Total
Item									
Formulation									
Replace	28	25	10	2	28	24	15	7	139 (71.28%)
Eliminate	7	3	0	1	3	3	3	0	20 (10.26%)
Add	4	10	2	0	7	6	2	1	32 (16.41%)
Parallel*	0	1	3	0	0	0	0	0	4 (2.05%)
Total	39	39	15	3	38	33	20	8	195 (100%)
Conceptual change									
Related	0	2	3	2	10	5	3	2	27 (28.42%)
Broad	9	4	1	0	5	3	3	2	27 (28.42%)
Narrow	4	9	3	0	4	5	4	1	30 (31.58%)
Synonym	2	1	1	0	3	2	1	1	11 (11.58%)
Total	15	16	8	2	22	15	11	6	95 (100%)

* Change of term order or change of preposition

To analyze query formulation, one member of the research team and a graduate assistant independently coded all queries submitted by participants using Jorgensen and Jorgensen's method. Then coding results were compared and discussed until a consensus on each query was reached.

To assess how effective Library of Congress (LC) subject headings were for users, user queries were searched against the LC online catalogue. This online catalogue was selected for the project because of the size of the LC collection, the range and quality of its contents, and the fact that the LC online catalogue provides access to its prints and photographs collection. Another advantage is that the LC OPAC supports keyword searching of subjects and contents notes. Figure 3 shows the relationship of retrieved sets and the focus of the study. Results of subject-heading (SH) searches and description-note searches were processed with Boolean NOT as follows:

- Items uniquely retrieved by LCSH = LCSH-retrieved items – note-retrieved items – title-retrieved items
- Items uniquely retrieved by description note = note-retrieved items – LCSH-retrieved items – title-retrieved items

Figure 3: Analysis focus

Results

User queries

For the immigration history photo, 33 participants provided 110 queries, and 31 participants provided 112 queries for the cancer poster (see table 3). The most common Boolean operator used in the queries was AND, accounting for 33.78%, while other operators were rarely used. The use of quotation marks for a phrase search accounted for about 25% of the total queries (see table 4). The average number of terms per query for all the searches ($N = 220$) was 3.12 terms. A mean of terms per query for immigration history was 3.08 and a mean of terms per query for the cancer poster was 3.21.

Figures 4 and 5 show term frequency among multiple queries submitted by the participants. The majority of initial queries were composed of two terms (42.4% for the immigration history photo, 35.5% for the cancer poster). The majority of the second queries were composed of three terms (38.7%, 26.7%). The majority of the third queries were composed

Table 3. Collected search queries

	Query 1	Query 2	Query 3	Query 4	Query 5	Total
Immigration history photo	33	31	29	14	3	110
Cancer poster	31	30	28	16	7	112
Total	64	61	57	30	10	222
	(28.83%)	(27.48%)	(25.68%)	(13.51%)	(4.50%)	(100%)

Table 4. Boolean operators usage

	AND	OR	NOT	Nesting/ Multiple	Phrase	Truncation
Immigration history photo (N = 110)	37	0	1	6	26	5
Cancer poster (N = 112)	38	3	0	5	29	5
Total* (N = 222)	75 (33.78%)	3 (1.35%)	1 (0.45%)	11 (4.95%)	55 (24.77%)	10 (4.50%)

* Some queries included multiple Boolean operators. In this study, single operator's occurrence was counted separately. The percentage represents occurrence of an operator out of the total collected queries.

Figure 4: Term frequency for immigration history photo

of three terms (44.8%, 32.14%). The majority of the fourth queries were composed of four terms (42.86%) and five terms (31.3%). It is interesting to note that when the participants modified their query over time, the number of terms in a query increased. Data on term frequency in query modification suggested that when users modified queries, they tended to add more terms to an initial query to narrow or broaden the search elements. Query modification strategies are further reported in table 2.

The study categorized query terms to examine the kinds of visual content expressed by them. Eight categories were used: proper nouns (such

Figure 5: Term frequency for cancer poster

as *Ellis Island*), common nouns (mostly objects, such as *immigrants*, *steamship*), adjectives (descriptive modifiers such as *holding*, in *holding facilities*), visual construct (type of images, type of image format, such as poster, graphics, or visual materials), concept terms (nouns and noun phrases representing a conceptual theme, such as *immigration*, *cancer care*), and unknown or undecipherable terms.

For the immigration history photo, proper nouns, nouns, visual construct, and concept terms accounted for 84%, while for the cancer poster, concept terms, visual construct, and proper nouns accounted for 94% (see table 1). Since the tasks used in this study were searching for visual resources, frequent use of format terms was logical: image format terms accounted for 20% and 30% for the two images respectively. About half (46.4%) of term types for the immigration history photo search were proper nouns and nouns representing location and objects in the photo and its scenario, while the other half (49.8%) of term types in the cancer poster were concept terms representing the conceptual theme of *cancer care*, which was available in the scenario. Participants seemed to adopt search terms from a task scenario or a given image. As other experimental studies found, study participants tended to use terms in a task assignment (Marchionini 1989; Zhang 2008).

Previous studies reported that when people looked for similar images seen before, their representation or interests would be in objects depicted in the image (Greisdorf and O'Connor 2002; Hollink et al. 2004). This study partially supported this finding in that when participants were asked to construct a query for the immigration history photo, a specific

scene, proper nouns and nouns representing specific objects (such as immigrants, a building), or location (such as Ellis Island) associated with the image were used more frequently. However, when they were asked to search for images representing a conceptual theme (the poster of cancer awareness), more concept terms were used. This finding suggests that tasks and the type of image that users are asked to search for play a role in formulation of the search query (see table 1).

Studies have reported that users change their queries frequently while searching for information (Goodrum and Spink 2001; Rieh and Xie 2006; Vakkari, Pennanen, and Serola 2003). This study compared changes to participants' queries from a sequentially previous query to analyze query modification strategies. The most frequent strategy was to replace one term with another (71.28%) (see table 2). This finding seems similar to that of Jorgensen and Jorgensen's study in that changing one or more of the terms was the most frequent approach. When changing a term, the participants attempted to change to narrow terms (31.58%), broad terms (28.42%), or related terms (28.42%). For the immigration history photo, change to broader terms occurred nine times and was the most used search move of the four types of changes (Related, Broad, Narrow, and Synonym). It seemed that the initial users' queries for the immigration history photo was specific and became broader in subsequent searches. On the other hand, for the cancer poster, change to related terms was the most frequent moves. It seems that domain and genre of images affect strategies for query reformulation. A possible influence of domain and genre of images in types of terms used in queries was also observed in this study (see the above section).

OPAC search results

Effectiveness of LCSH and description notes

Participants submitted 110 queries for the immigration history photo and 112 queries for the cancer poster. The queries were run against subject heading fields and description notes in the Library of Congress online catalogue to assess how well subject headings and description notes supported user queries. Queries with multiple Boolean operators were reformulated to accommodate the advanced search interface of the online catalogue, resulting in a total of 126 queries and 120 queries for the photo and the poster respectively. Each query was searched against subject heading fields and description note fields. A total of 252 searches were performed for the photo topic and 240 searches for the poster topic.

Figure 6: Average Number of Items Retrieved

The researchers used the NOT operator in the online catalogue to process initial search results in order to identify items that were retrieved from query term matches in LC subject headings and items that were retrieved from query term matches in description notes. The online catalogue, however, could not remove items retrieved from keyword matches in the title field automatically, so search sets with up to 100 hits were manually reviewed to remove items with user query terms in the title field. Analysis of random samples found the number of such items to be very small.

For both images, user queries searched against LC subject headings produced a much larger number of items than when searched against description notes. For the immigration history photo, LCSH searches produced an average of 282 items while note searches produced 26; for the cancer poster, LCSH searches produced an average of 386 items while note searches produced 100 items (see figure 6). Searchers seemed to have better luck matching their query terms to LCSH than to description notes. However, the large sets were often the result of users using broad terms, such as *cancer* and *United States*, which then led to high recall but questionable precision. The high recall of LCSH searches, furthermore, becomes less of an asset when searches with zero hits are taken into account.

Figure 7 shows that 72% and 75% of the LCSH searches resulted in empty sets; in contrast, 58% and 59% of description note searches resulted in empty sets. The high percentages of empty sets revealed that subject headings and description notes could support user search queries less than half the time. These data underscored the challenge for image searchers: the match between their vocabulary and the metadata provided

Figure 7: Number of Empty Sets Retrieved

by professional cataloguers, such as subject headings and description notes, was very limited.

No interaction between search mode and subject/format

To determine whether the subject matter (history vs. medicine) or format (photo vs. poster) contributed to the different results of LCSH searches and description note searches, a two-way between group ANOVA test was performed. The test found the interaction between subject domain and format was not significant ($F = 0.016$, $p = .8994$). The test result enabled the researchers to further analyze the results of searches by the main effects—LCSH and description notes—with confidence.

Retrieval differences between LCSH and description notes

Our t tests found the differences between the two modes of searches statistically significant (see table 5), meaning that differences were not

Table 5. t tests of search differences for both images

Value	Mean Difference	df	t Value	p
Immigration history photo LCSH vs. description notes	256.8	125	2.138	.03
Cancer poster LCSH vs. description notes	285.6	119	2.300	.02

due to chance. In other words, subject headings and description notes supported end user queries differently. Description note searches produced fewer empty sets than subject heading searches, indicating description notes were able to match user terms better than subject headings. Subject heading searches, on the other hand, retrieved larger search results than description note searches, probably because searchers used broad terms.

Concept analysis

While neither subject headings nor description notes seemed to match user query terms well, it was not clear whether the mismatch was related to conceptual differences between searchers and professional indexers. To address this issue, concepts in user queries and concepts in subject headings and description notes were compared. For the purpose of analyzing concepts in user queries, the study used Ranganathan's *personality, matter, energy, space, and time* (PMEST) as the framework. PMEST has had a long history of application in knowledge organization and subject analysis, and the five fundamental categories provided a helpful structure for identifying concepts from the two target images. Ranganathan's *personality* (P) refers to actors, *matter* (M) refers to things being acted on, *energy* (E) refers to actions and events, *space* (S) refers to place, and *time* (T) refers to time. In the study's framework *P* represented actors, such as people depicted in images and agencies sponsoring actions or activities. *M* represented resources and materials on which actions were presented, including inanimate objects in images. *E* represented actors' activities and themes of the images. *S* represented place and environment where actions and events took place. *T* represented the time when actions and events took place. Using PMEST as the five facets of interest, the researchers identified 12 concept categories for the photo and nine concept categories for the poster and used them for query analysis. Tables 6 and 7 present the concept categories under PMEST.

To cluster user concepts, terms for the same concept were grouped together and the occurrences of terms under each concept were recorded. Of the 110 queries for the immigration history photo, concepts from four queries could not be sorted into the 12 concept categories and were not included in the final analysis. These excluded concepts were *National Archives*; *travel, move, relocate*; *country of origin*; and *shipping, docking companies*. Of the 106 queries, 265 concepts were identified. The number of concepts ranged from 1 to 5, the average number of

Table 6. Concept groups identified for analysis of queries for immigration history photo (12 concept groups)

Facet	Concept Group
Personality	Immigrants Women Children
Matter	Picture Bridge Building Mood
Energy	Immigration history
Space	Ellis Island United States
Time	20th century Winter

Table 7. Concept groups identified for analysis of queries for cancer poster (9 concept groups)

Facet	Concept Group
Personality	US Public Health Service American Society for Control of Cancer
Matter	Exhibit Picture Poster Visual materials
Energy	Cancer Cancer care Health care
Space	NA
Time	NA

concepts was 2.5, and queries with 2 concepts were the most popular approach. As table 8 shows, *Ellis Island*, the setting of the photo, and *Picture*, the format category, represented 45% of participants' concepts, while thematic categories such as immigration and immigrant represented 33% of users' concepts. *Time* (19th century) and *space* (United States) categories represented 18% of the concepts, but objects in the photo such as buildings and women were not heavily used for searching.

Table 8. Concepts used in queries for immigration history photo (total concepts used: 265)

Concepts	Number	%	PMEST
Ellis Island	61	23.0	Space (image of and about)
Picture	59	22.0	Matter (format)
Immigration history	47	18.0	Energy (image about)
Immigrants	41	15.0	Personality (image of and about)
United States	34	13.0	Space (image about)
20th century	11	11.0	Time
Bridge	9	3.0	Matter (object in image)
Building	2	0.8	Matter (object in image)
Women	1	0.4	Personality (object in image)
Children	–	–	Personality (object in image)
Winter	–	–	Time
Mood	–	–	Matter

Of the 112 queries for the poster, the concepts of 4 queries were not represented by the nine concept categories established by the researchers for data analysis. The excluded concepts were *relay for life*, *life expectancy*, *National Institutes of Health*, and *Memorial Sloan Kettering*. Of the 108 queries, 200 concepts were identified. The number of concepts ranged from 1 to 5, the average number of concepts was 1.9, and queries with 2 concepts represented the most popular approach.

User search queries drew heavily on the theme and format of the poster. As table 9 indicates, thematic categories—*cancer*, *cancer care*, and *health care*—represented 53% of user concepts. The format categories (37% total)—*visual materials*, *poster*, and *picture*—were also favoured by image searchers. The agencies sponsoring the campaign were used in some search queries as well. What tables 8 and 9 suggest is that image searchers use themes and material formats often, and it is important for indexers to bear this in mind when they organize images for access.

In addition, user concepts were compared with LC subject headings and the description note prepared by LC for the photo. Table 10 presents these professionally provided metadata. LC's subject headings reflect the practice of assigning the most specific headings and constructing subject heading strings according to the special syntax of Library of Congress

Table 9. Concepts used in queries for cancer poster (total concepts used: 200)

Concepts	Number	%	PMEST
Cancer care	66	33	Energy (image about)
Visual materials	38	19	Matter (format)
Cancer	31	16	Energy (image about)
Poster	29	15	Matter (format)
US Public Health Service	14	7	Personality (sponsor)
American Society for Control of Cancer	10	5	Personality (sponsor)
Health care	7	4	Energy (image about)
Picture	5	3	Matter (format)
Exhibit	—	—	Matter (format)

Table 10. LC subject headings and description notes for the two images

Image	LC Subject Headings	LC Description Notes
Ellis Island (photo)	<p>Ellis Island Immigration Station (N.Y. and N.J.)—1900–1910</p> <p>Immigrants—Ellis Island (N.J. and N.Y.)—1900–1910. lctgm</p> <p>Immigration stations—Ellis Island (N.J. and N.Y.)—1900–1910. lctgm</p> <p>Emigration & immigration—Ellis Island (N.J. and N.Y.)—1900–1910. lctgm</p> <p>Ellis Island (N.J. and N.Y.)—1900–1910</p> <p>Form/Genre:</p> <p>Half-tone photomechanical prints—1900–1910. gmGPC</p>	<p>Emigrants coming up the boardwalk from the barge, which has taken them off the steamship company's docks, and transported them to Ellis Island. The big building in the background is the new hospital just opened. The ferry seen in the middle of the picture runs from New York to Ellis Island.</p>
Cancer (poster)	<p>United States Public Health Service—Public relations—1930–1940</p> <p>Cancer—1930–1940. lctgm</p> <p>Health care—New York (State)—New York—1930–1940. lctgm</p> <p>Form/Genre:</p> <p>Posters—1930–1940. gmGPC</p> <p>Screen prints—Color—1930–1940. gmGPC</p>	<p>Poster encouraging early treatment for cancer</p>

Table 11. Top query concepts and coverage by LCSH and description notes

Image	Top 4 Concepts Used	% of Queries Using Top 4	Concepts Covered by LCSH	Concepts Covered by Description Notes
Ellis Island (photo)	Ellis Island (61) Picture (59) Immigrant history / immigration (47) Immigrants (41)	78.5	Ellis Island immigrants picture	Ellis Island emigration
Cancer (poster)	Cancer care (66) Visual materials (38) Cancer (31) Poster (29)	82.0	Cancer poster	Cancer treatment poster

subject headings. The subject headings reflect users' concepts of Ellis Island, immigrants, and immigration, but the geographic qualifiers are more specific than the users' term (*United States*) and the qualifiers are presented in the standard abbreviated form of the *Anglo-American Cataloguing Rules*, which users are unlikely to use. The chronological subdivision is also presented in a special syntax and covers a specific period that is narrower than users' terms, such as *19th century*. The format concept is represented by a technical term *halftone photomechanical prints* from the *Thesaurus for Graphic Materials II*.

The subject headings assigned to the poster are similarly problematic for users. They represent the theme of cancer, but a broader term, *health care*, is used for cancer care, and all the headings have geographic or chronological subdivisions. User queries reflect an interest in using the concept of visual materials for searching (19%), and many terms are used for this concept. LC's genre term (*poster*) reflects the object exactly, but another term, *screen prints*, from the *Thesaurus for Graphic Materials II* seems too technical for searchers to match.

LC's description notes use natural language. The note on the photograph provides informative detail and includes concepts such as Ellis Island and New York. In spite of its brevity, the note on the poster sufficiently captures the purpose of the poster and includes three concepts important to users: poster, cancer, and treatment.

As table 11 indicates, the top four user query concepts were used in more than 78% of the queries and covered the main themes and format of the

two images. Conceptually, LC subject headings and description notes covered the top four query concepts fairly well.

Discussion

Analysis of user queries revealed an interesting picture of how users searched for images in this study. On average, each query included three search terms. The average of search terms found in our study is more than that in previous studies: 2 terms in Jorgensen and Jorgensen's (2005) study and 1.48 terms in Westman and Oittinen's study (2006). The main difference between these three studies may have something to do with the background of the participants. Participants in this study were not image experts and had little subject background in the two topic areas, whereas the two earlier studies used image professionals and journalists.

Furthermore, participants used format terms often but made little use of Boolean operators, other than the Boolean AND, and quotation marks for phrases. The study confirmed the occurrence of frequent query modification in image search and found that strategies for term formation and query modification seemed to be related to the types of images sought and the subject domain of the images. When searched in subject heading fields, user queries retrieved more empty sets than when searched against description notes, suggesting difficulty in finding user search terms in professionally assigned subject headings. LCSH searches, however, produced much larger sets than description note searches.

The study found that users tended to search by concepts about the themes and format of the desired images. Term variation, however, presented a serious challenge for searchers and image indexers. Although concepts important to users were covered by LC subject headings and description notes, user query terms were infrequently found in these metadata. Assistance in term selection and the capability to cluster expressions for the same concept would improve precision and recall of image retrieval.

Implications for image indexing

In organizing images for access, professional indexers typically consider what an image is *of* (for example, immigrants at Ellis Island) and what an image is *about* (for example, immigration). They pay attention to the

context of an image as well as its content. Specifically, who, what, when, where, and why are considered for indexing purposes. The subject headings and description notes assigned to the two images selected for the study cover their content and context well; and analysis of user queries confirms the validity of focusing on content and context, because image searchers frequently use concepts about themes. Searching user queries in the LC online catalogue, on the other hand, helps identify two major areas for improving image indexing.

The principle of indexing specificity and the syntax of subject headings

Using the most specific subject heading strings can improve precision but carries the risk of causing empty sets because searchers unfamiliar with LC subject headings may have difficulty using terms in subject heading strings like *Emigration & immigration—Ellis Island (N.J. and N.Y.)—1900–1910*. OPACs that support subject keyword searches will retrieve something for users when they search by *immigration* or *Ellis Island*. However, the syntax of subject heading strings, the state abbreviations, and the specific period pose challenges for searchers who may enter New York or 19th century to search for these concepts. To accommodate users' tendency to search by broad terms and natural language, it would be helpful to provide some support, such as a pull-down menu of terms to help users select appropriate geographic and chronological terms. Another approach is to complement professionally assigned meta-data such as *health care* with more precise user tags such as *cancer care* or *cancer treatment*.

Specific and technical genre terms

Terms from the *Thesaurus for Graphic Materials II* are specific and precise in representing the technical nature of form or genre. Such terms may be useful to image specialists but their value is lost to non-expert image users. It would be helpful to present format or genre in terms that are more accessible to non-specialists or build in a transparent cross reference system to take users who search by an unauthorized format term to the authorized one.

Implications for image searchers

The fact that LC subject headings and description notes effectively cover the top four concepts in user queries indicates that indexers and searchers

have similar ideas on the concepts that are important for searching. The biggest challenge lies in devising ways to match user search terms to professionally assigned metadata like subject headings and description notes, or including some user terms to complement professional metadata. It is fine to include technical terms to address the needs of image specialists, but an image catalogue or databases designed for users in public or academic libraries would demand less technical terms (such as *poster*) to support non-specialist searchers.

Helping searchers to find the right terms for a concept is challenging, but Web search engines and some websites have implemented promising supports. For example, when a search term such as *immigration* is being entered, Google and Amazon will open a window of expressions with the same word stem to show users the range of options. Search interface can present related terms for searchers to identify the most precise terms for their queries. Search interface can ask searchers questions to clarify the concept they have in mind (“Do you mean ‘bank’ as a financial institution?”). Search terms can be presented in facets for searchers to select and combine as needed. Since users are highly interested in assigning user tags on the Web, another approach is to support social tagging in an image catalogue or database and design ways to automatically categorize user tags by concept category. User tags in categories related to a new searcher’s interest can be added to the pool of terms to improve the number and relevance of items retrieved.

Conclusions

As user interest in digital image grows and more digital images are available, it becomes more urgent for information professionals involved in image organization to address challenges in image retrieval. Because Web search engines, online catalogues, and search engines of digital collections continue to rely heavily on text for image retrieval, it is critical to examine how users choose terms in their queries. This study examined characteristics of user queries for images in an OPAC and investigated how well Library of Congress subject headings and description notes could support end user search queries. The study found the average number of search terms to be 3.12 and low use of Boolean operators. The subject domain and the format of images seemed to influence query development, types of terms selected, and query reformulation.

The study also found that neither subject headings nor description notes were very successful in supporting user queries for images, and that when

user search terms matched subject headings, the result sets tended to be very large, making it necessary for users to sift through a lot of items to find relevant ones. Through query term analysis and concept analysis, the study discovered that end users tended to search by concepts related to the themes and format of desired images. End users and indexers were able to identify key concepts for searching and indexing purposes respectively and that professionally prepared metadata such as subject headings and description notes covered users' concepts well. However, exact vocabulary matches were highly problematic mainly because controlled metadata tended to be somewhat artificial and sometimes technical, rendering it difficult for end users to match them in queries. While indexing practice may not need to change drastically, image managers will want to build in support to bridge users' vocabulary to controlled metadata to facilitate the retrieval of images.

References

- Armitage, Linda, and Peter G.B. Enser, 1997. Analysis of user need in image archives. *Journal of Information Science* 23: 287–99.
- Attig, John, Ann Copeland, and Michael Pelikan. 2004. Context and meaning: The challenges of metadata for a digital image library within the university. *College & Research Libraries* 65 (3): 251–61.
- Aula, Anne. 2003. Query formulation in Web information search. In *Proc. IADIS International Conference WWW/Internet 2003*, ed. P. Isaias and N. Karmakar, 1: 403–10. Algarve: IADIS.
- Chen, Hsin-liang. 2001. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology* 52 (3): 260–73.
- Choi, Youngok, and Edie Rasmussen. 2002. User's relevance criteria in image retrieval in American history. *Information Processing and Management* 38 (5): 695–726.
- . 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.
- Collins, Karen. 1998. Providing subject access to images: A study of user queries. *American Archivist* 61: 36–55.
- Eakins, John P., Pam Briggs, and Bryan Burford. 2004. Image retrieval interfaces: A user perspective. In *Proceedings of the Third International Conference on Image and Video Retrieval*, ed. P. Enser, Y. Kompatsiaris, N.E. O'Connor, A.F. Smeaton, and A.W.M. Smeulders, 628–37. Berlin: Springer-Verlag.
- Enser, Peter. 2000. Visual image retrieval: Seeking the alliance of concept-based and content-based paradigms. *Journal of Information Science* 26 (4): 199–210.
- . 2008. Visual image retrieval. *Annual Review of Information Science and Technology* 42: 3–42.

Fidel, Raya. 1997. The image retrieval task: Implications for the design and evaluation of image databases. *New Review of Hypermedia and Multimedia* 3: 181–99.

Frost, Carolyn Olivia, Bradley Taylor, Anna Noakes, Stephen Markel, Deborah Torres, and Karen M. Drabenstott. 2000. Browse and search patterns in a digital image database. *Information Retrieval* 1 (4): 287–313. <http://hdl.handle.net/2027.42/45984>.

Fukumoto, T. 2004. An analysis of image retrieval behavior for metadata type and Google Image Databases. *Proceedings of International Conference on Computers in Education, 1921–7*. Tarrytown, NY: Pergamon.

Goodrum, Abby A., Matthew M. Bejune, and Antonio C. Siochi. 2003. A state transition analysis of image search patterns on the Web. In *Proceedings of the Second International Conference Image and Video Retrieval (CIVR 2003)*, ed. E.M. Bakker, Thomas S. Huang, Michael S. Lew, Nicu Sebe, and Xiang Zhou, 281–90. Berlin: Springer-Verlag.

Goodrum, Abby A., and Amanda Spink. 2001. Image searching on the EXCITE Web search engine. *Information Processing and Management* 37: 295–312.

Greisdorf, Haward, and Brian O'Connor. 2002. Modeling what users see when they look at images: A cognitive viewpoint. *Journal of Documentation* 58 (1): 6–29.

Hasting, Samantha K. 1995. Query categories in a study of intellectual access to digitized art images. *Proceedings of the 58th Annual Meeting of the American Society for Information Science* 32: 3–8.

Hollink, L., A. Th. Schreiber, B.J. Wielinga, and M. Worring. 2004. Classification of user image descriptions. *International Journal of Human-Computer Studies* 61 (5): 601–26.

Hung, Tsai-Youn. 2005. Search moves and tactics for image retrieval in the field of journalism: A pilot study. *Journal of Education Media & Library Sciences* 42 (30): 329–46.

Jørgensen, Corinne. 1998. Image attributes in describing tasks: An investigation. *Information Processing and Management* 34: 161–74.

Jørgensen, Corinne, and Peter Jørgensen. 2005. Image querying by image professionals. *Journal of American Society for Information Science and Technology* 56 (12): 1346–59.

Laine-Hernandez, Marie, and Stina Westman. 2006. Image semantics in the description and categorization of journalistic photographs. In *Proceedings of the 69th Annual Meetings of the American Society for Information Science* 43 (1): 1–25.

MacDonald, Sharon, and John Tait. 2003. Search strategies in content-based image retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 80–7. New York: ACM.

Marchionini, Gary. 1989. Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science* 40 (1): 54–66.

- Matusiak, Krystyna K. 2006. Information seeking behavior in digital image collections: A cognitive approach. *Journal of Academic Librarianship* 32 (5): 479–88.
- Palmer, C., O. Zavalina, and M. Mustafoff. (2007). Trends in metadata practices: A longitudinal study of collection federation. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'07)*, 386–95. New York: Association for Computing Machinery.
- Panofsky, Erwin. 1962. *Studies in iconology: Humanistic themes in the art of the Renaissance*. New York: Harper & Row.
- Pu, Hsiao-Tieh. 2005. A comparative analysis of Web image and textual queries. *Online Information Review* 29 (5): 457–67.
- Rafferty, Paul, and Rob Hilderley. 2007. Flickr and democratic indexing: Dialogic approaches to indexing. *Aslib Proceedings* 59 (4/5): 397–410.
- Rieh, Soo Young, and Hong Xie. 2006. Analysis of multiple query reformulations on the Web: The interactive information retrieval context. *Information Processing and Management* 42: 751–68.
- Roissa, Abebe. 2008. User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. *Information Processing and Management* 44 (5): 1741–53.
- Shatford-Layne, Sara. 1994. Some issues in the indexing of images. *Journal of the American Society for Information Science* 45 (8): 583–8.
- Smeulders, A.W.M., M. Worring, S. Santini, A. Gupta, and R. Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2): 1349–80.
- Vakkari, Pertti, Mikko Pennanen, and Sami Serola. 2003. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing and Management* 39: 445–63.
- Westman, Stina, and Pirkko Oittinen. 2006. Image retrieval by end-users and intermediaries in a journalistic work context. *Proceedings of the 1st International Conference on Information Interaction in Context* 176: 102–10. New York: ACM.
- Yoon, Jungwon. 2006. An exploration of needs for connotative messages during image search process. In *Proceedings of the 69th Annual Meetings of the American Society for Information Science*, ed. A. Grove, 43 (1): 1–19.
- Zhang, Yan. 2008. The influence of mental models on undergraduate students' searching behavior on the Web. *Information Processing and Management* 44: 1330–45.