Collection-Level User Searches in Federated Digital Resource Environment

Oksana Zavalina
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel, Champaign, IL 61820. Email: zavalina@uiuc.edu

As part of a federation project providing integrated access to over 160 digital collections, we are studying how collections can best be represented to meet the needs of diverse user communities. This paper reports preliminary transaction log analysis results from that project on subject representation of the digital collections. The findings reveal prevalence of the broadly defined subject search at the collection level, and the lack of semantic similarity between the user queries and the GEM controlled vocabulary terms used for collection description. Based on the actual search data, it is recommended that the 2nd group of entities in the FRBR model be updated to cover ethnic/national groups and classes of persons. The paper proposes definitions of the two major collection-level search types — known-item and subject — and formulates research questions for further investigation into subject access to federated collections.

Introduction

Subject access to collections and their contents has been a focus of attention in the LIS field for decades. A number of catalog use studies have been conducted in attempts to better understand the role of subject description and the problems users face while searching for information on a particular topic, with transaction log analysis as a method of these studies (e.g., Matthews, Lawrence, & Ferguson, 1983; Larson, 1991B). However, issues of subject access in federated collections, where the "unit of analysis" is a collection rather than an item search, have not yet been investigated. This paper reports preliminary results of this kind of analysis performed on the IMLS Digital Collection Registry transaction log dataset.

The IMLS Digital Collections and Content (DCC) project began at the University of Illinois at Urbana-Champaign in January 2003. After developing collection description metadata schema the DCC project has created a collection registry (hereafter referred to as the Registry) of 169 digital collections funded through the Institute of Museum and Library Services National Leadership Grant (NLG) and built by cultural heritage institutions since 1998. Collections funded through the Library Services and Technology Act (LSTA) grant have been included since 2006. An item-level repository has been created and made public in 2006; digital content from 58 NLG-funded digital collections has been harvested into the repository to date. The types of digital content of the Registry include image (in 80% of collections), text (68%), physical object (29%), sound file (20%), interactive resource (10%), moving image (7%), and dataset (4%). Broad areas of social studies and arts constitute major subject strengths of collections in the Registry.

The Registry is indexed with the Gateway to Educational Materials (GEM) subject scheme created to describe digital objects in the Gateway for Educational Materials repository — a National Library of Education initiative to expand educators' access to Internet-based lesson plans, curriculum units and other educational materials. In part due to a high national and international reputation gained by GEM the subject scheme's application goes beyond its original educational domain. The scheme is considered suitable for browsing databases in more general cultural heritage domain. It consists of twelve "level 1" broad subject headings: Arts, Educational Psychology, Foreign Languages, Health, Language Arts, Mathematics, Philosophy, Physical Education, Religion, Science, Social Studies, and Vocational Education, each of which has between 12 and 29 narrower "level 2" headings under it. The second level subject headings for Philosophy and Religion replicate ERIC Thesaurus "Narrower Terms" for these two broad subjects. Several of the "level 2" GEM subject headings — Careers, History, Informal education, Instructional issues, Process skills, and Technology — are facets applicable to each of the twelve broad subject categories. Digital resource developers participating in the Registry are required to provide top-level GEM subjects in their collection descriptions. Use of alternative subject headings for collection description is not required but supported by the metadata schema.

Results from recent DCC survey and interview data show that digital resource developers are not completely satisfied with the GEM subject scheme use for collection level description. Most of them point to a particular drawback — lack of breadth and depth in topic coverage, especially at the top level of the

subject hierarchy. The absence of standardization in name authority is a recognized deficiency of the digital library architecture (Sutton, 2004), and GEM subject scheme is a good example of this problem: neither name nor place subject are represented in it.

This study is aimed to measure suitability of the GEM subject scheme for describing the diverse collections in the Registry and compare it with the same indicators obtained for the alternative controlled vocabularies. The following criteria were adopted from the literature on subject scheme evaluation (Cochrane, 1986; Larson, 1991A, etc.): 1) diversity of topics covered (breadth and depth of subject coverage), 2) syndetic structure of the subject scheme, 3) heading structure, 4) currency of subject headings, 5) availability of links between scheme's subject headings and subject terms from other controlled vocabularies. Based on IMLS-registry-specific observations, this list of general criteria for measuring subject scheme suitability to collection level description was expanded by adding three criteria dealing with semantic similarity between: 1) GEM subject terms and keywords used by Registry searchers, 2) GEM subject terms and alternative subject terms used in collection level description of specific collections, and 3) GEM subject terms used in collection level description and subject headings used in item- level description within specific collections. This paper concentrates on semantic similarity measures comparing user keywords extracted from the Registry transaction logs and the subject terms in three different controlled vocabularies — GEM, Library of Congress Subject Headings (LCSH), and Art and Architecture Thesaurus (AAT). The LCSH was selected since almost half of the digital collections participating in the Registry are using for item-level description and according to our survey results it is being considered by some of the digital resource developers as an alternative to GEM for collection-level description. The AAT was selected as another plausible alternative for describing cultural heritage collections. A number of collections participating in the Registry are currently using AAT for item-level descriptions; moreover, it is a controlled vocabulary of a narrower scope than LCSH, but substantially more detailed than GEM.

Although a significant volume of research has been dedicated in the LIS literature to the two major types of search within collections (subject and known-item) (e.g., Krikelas's overview of catalog use studies, 1972; Lee, Renear, & Smith, 2006), no research has been done yet with the focus on specifics of search types in federated collections, at the collection level. Our interests are in the correlation between the search type/category and the degree of semantic match between user search terms and controlled vocabulary terms. Comparing Functional Requirements for Bibliographic Records (FRBR) set of entities with the actual kinds of user searches has been a part of this research question. Another goal of this study has been providing general description of the searches made by users in the Collection Registry: 1) the ratio of subject and known-item searches, and 2) typical query profile in terms of length, frequency of query use, etc.

Methods

The major dataset used in this study was IMLS Collection Registry transaction logs — a Microsoft Access file that covered a period of 7 months, between February 2005 (when the Registry was first made publicly accessible) and September 2005. The initial transaction log file consisted of over 100,000 records, but after exclusion of the searches and browsing made by web crawlers and Registry testers its size was reduced to approximately 19,000 records. Each record contained information on the IP address the query originated from, date and time of access, webpage visited within the Registry, raw query string, etc. The transaction log was manually processed to extract all the keyword search query strings — a total of 936. Preserving the context of a search has been considered an important factor for categorizing searches and finding semantic matches with the controlled vocabulary terms. Therefore, the decision was made not to parse queries into separate words or even further — into stems. Minimal processing of the queries was undertaken: plurals were truncated and grouped together with the singulars [morphological variants] (e.g., "Indians" and "Indian", "clipper ships" and "clipper ship"); both correct and misspelled versions of the same words (e.g., "Antarctica" and "antartica", "immigration" and "imigration") were considered the instances of the same query. The stop word list included all prepositions, conjunctions and articles.

At the first stage of analysis, general descriptive statistical procedures were used: search frequencies and the number of words were calculated for each query, averaged for the whole sample and for each search category separately. The major part of this stage was qualitative analysis **—** categorizing the user queries into seven broad search categories derived from the Functional Requirements for Bibliographic Records

(1998). Seven out of ten FRBR entities that can serve as subjects of a work were adopted for this study: *work*, *person*, *corporate body*, *concept*, *object*, *event,* and *place*. The definitions of each entity and examples given by FRBR were followed as guidelines for distinguishing between the categories. In essence, the seven categories are characterized by FRBR as:

1. "*work:* a distinct intellectual or artistic creation" (FRBR, p. 16)
2. *"person*: an individual; encompasses individuals that are deceased as well as those that are living" (p. 23)
3. *"corporate body*: an organization or group of individuals and/or organizations acting as a unit; encompasses organizations and groups of individuals and/or organizations that are identified by a particular name…" (p. 24)
4. *"concept*: an abstract notion or idea; encompasses a comprehensive range of abstractions that may be the subject of a *work*: fields of knowledge, disciplines, schools of thought (philosophies, religions, political ideologies, etc.), theories, processes, techniques, practices, etc. A *concept* may be broad in nature or narrowly defined and precise" (p. 25)
5. *"object*: a material thing; encompasses a comprehensive range of material things that may be the subject of a *work*: animate and inanimate objects occurring in nature; fixed, movable, and moving objects that are the product of human creation; objects that no longer exist" (p. 26)
6. *"event*: an action or occurrence; encompasses a comprehensive range of actions and occurrences that may be the subject of a work: historical events, epochs, periods of time, etc."(p. 27)
7. "*place*: a location; encompasses a comprehensive range of locations: terrestrial and extra-terrestrial; historical and contemporary; geographic features and geo-political jurisdictions"(p. 27).

The FRBR *expression, manifestation* and *item* entities have not been adopted as categories for this analysis — although the cataloging has been traditionally performed for the manifestation level, it is virtually impossible to detect from the transaction log data alone what exactly the user is searching for: an abstract work, its particular expression, manifestation or item. Therefore, in our classification of the Registry queries, *work* category is broader than FRBR *work* and covers any intellectual or artistic creation that has a title attribute, including the digital *collections* that are members of the Registry.

The FRBR *person* entity is currently limited to individual persons but in the process of data analysis presented in this paper it was discovered that at least two other entities — supersets of individual persons — are widely used in actual searches and should be added to its second group of entities: *ethnic/national groups* (e.g., "Irish Americans"," Sioux Indian", "Basque"), and *classes of persons* (e.g., "children that are abused", "prisoners", "country people"). These two additional entities, along with *family* entity (Zeng & Salaba, 2005; FRANAR, 2007) were incorporated into the analysis. The rare occasions of fictitious characters has been treated on the basis of "what they would be if they really existed". For instance, Don Quixote would be an *individual person.* The TV series' character Alf, on the other hand, is a creature, just as a dog or a squid, thus a FRBR *object*.

To achieve consistency in distinguishing between search categories in less straightforward cases, unspecified institutions (e.g., "library", "archive", "can company", "prison") were categorized as *concepts*, while the more specifically named ones (e.g., "Icy Hot Bottle Co.", "library + Moorhead") as *corporate bodies* or *objects* respectively. Some queries presented a real challenge for classification: "books" and "tools" are just two of them, categorized as *objects*, although they could as well be FRBR *concepts*. As any categorization, our approach is inevitably subjective, which constitutes one of the limitations of this study. Another limitation of applying FRBR entity-relation schema — as probably any other framework — for categorization of subject searches lies in the ambiguity and polysemy of the actual queries further discussed in the Findings and Discussion section.

The queries that were entirely ambiguous as to which search category they belong to (e.g., "aF", "beyond", "LU+65") or the intent of the search (e.g., "google", "GEM") were grouped together in the *unknown* search category.

The second stage of analysis included matching actual user queries from the Registry transaction logs to subject terms in three controlled vocabularies — GEM, LCSH, and AAT; results were totaled and averaged for the whole sample and for each of the eleven search categories separately. OCLC Connexion database features — LCSH authority file search and Web Dewey search for editorially mapped LCSH headings — were used for matching user queries with LCSH. Because GEM is not a structured thesaurus, analyzing related, broader and narrower term matches across the three subject schemes was impossible. Only exact and synonymous matches [semantic variations] (e.g., "inoculation" —"vaccination", "raffles"—"lotteries") were considered semantic similarities. Abbreviated queries were matched with the full terms in controlled vocabularies, e.g., "ilgwu" with "International Ladies' Garment Workers' Union", "WW1" with "World War, 1914-1918", "polio" with "poliomyelitis". The order of search terms in the query was ignored for analysis [syntactic variations] (e.g., "French art" was matched with "Art, French", "children that are abused" with "abused children"). Endings [morphological variations] were also disregarded, as long as they did not affect the meaning of the words (e.g., "automated speech recognition" was matched with "automatic speech recognition"). Both preferred terms and — whenever available — variant terms in a controlled vocabulary were considered legitimate matches. Simple user queries were in some cases matched with compound LCSH subject headings, for instance "housing for shipyard workers" was matched with "Shipbuilding industry—Employees—Housing", "photographs of river" — with "Rivers—Photographs".

Coders other than author of this paper were not employed formally and therefore the intercoder reliability rate was not calculated and reported in this paper. However, author coding results have been discussed by the two independent groups: Spring 2006 "Data Analysis in LIS" seminar at the University Of Illinois Graduate School of Library and Information Science, and the metadata roundtable hosted by the IMLS Digital Collections and Content project. The user search terms that spurred the most discussion and upon which the common agreement have not been reached are presented in the Search Categories and Types section of this paper. This limitation of the preliminary analysis will be remedied in further analysis for which formal research procedure will be adopted, at least three coders will be involved, and intercoder reliability rate will be measured.

Findings and Discussions

Search Categories and Types

The first stage of analysis demonstrated that almost three-quarters of all searches made in the Registry are distributed between four broad search categories: *object* (24%), *concept* (21%), *place* (15%), and *individual person* (13%).

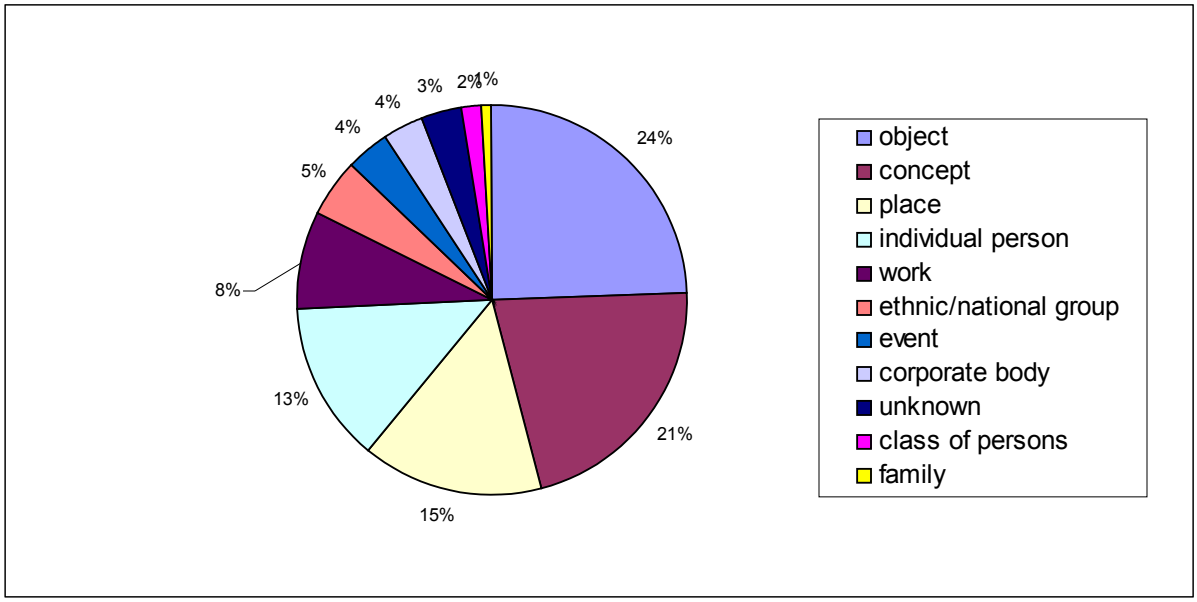

Figure 1. Unique search terms distribution by search categories

The remaining 27% fell under *work, corporate body, event, ethnic/national group, class of persons, family,* and *unknown* search categories. The low level of *event* searching is surprising, since most

of the historical searches would be the searches for *events*. The figures presented in this section show results in regards to *unique search terms* — sets of specific user query instances.

Because of the very nature of *concept, object, place*, and *event* (as defined by FRBR, 1998), these cannot fall under the widely-used LIS definition of the general known-item search type — "a search for some item for which either the author or title is known" (American Library Association, 1958), alternatively defined as *known-work* search (Yee & Layne, 1998). Since *family*, *ethnic/national group* and *class* of *persons* cannot be considered authors of the work, these searches do not belong to the known-item search type either. Therefore *concept, object, place*, *event, family*, *ethnic/national group* and *class* of *persons* search categories can be legitimately considered subject searches, which, broadly defined, includes both controlled- and uncontrolled-vocabulary searches with an intent to find information on a particular subject/topic/discipline/area. As demonstrated by Figure 2 below, subject searches constitute at least 72% of all unique search terms.

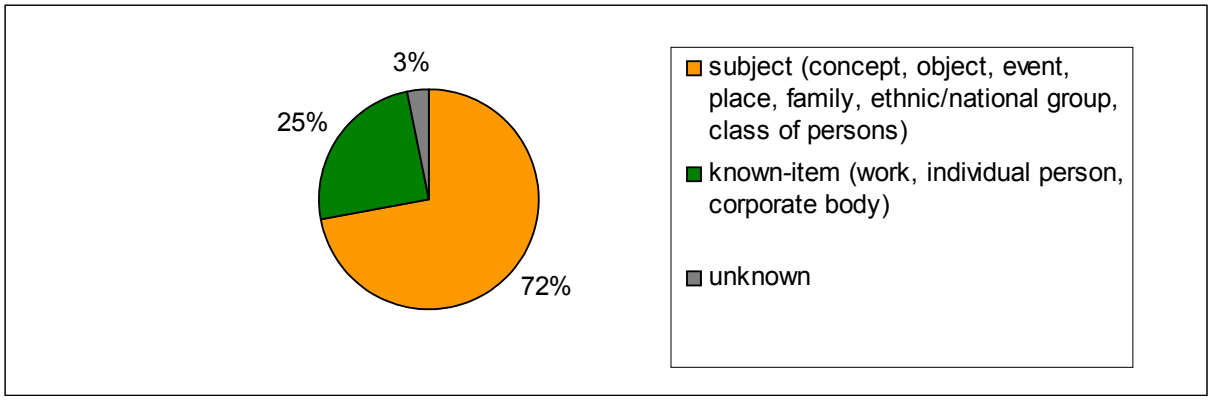

Figure 2. Unique search terms distribution by major search types

Although the number of federated digital collections has been growing recently, as is the creation and use of collection registries, no attempt to conceptualize known-item and subject searches specifically at the collection level has been documented in LIS literature. In our operational definition, searches where the user queries either the title or the author — individual or corporate — of the digital collection belong to collection-level known-item search type; all the other searches in the Registry belong to a collection-level subject search type.

The majority — sixty-seven percent — of the *work* searches were searches for a specific digital collection title, its identifiable portion, or in one case collection URL, thus belonged to a known-item search type. Since the rest of the *work* searches were for specific item-level titles, and therefore at the collection-level search can be treated as subject searches, the distribution of the two major search types — subject and known-item — has been revised as shown in the Figure 3.
If the users were to be interviewed, some of the *item-level work* searches would be found to be performed with intent to find the specific known item. However, possible distortion of the results is evened out by the fact that some of the *individual person* or *corporate body* searches can turn out to be conducted with the aim of finding information on a specific subject. In general, the conservative technique applied here tends to slightly inflate the known-item search type numbers and underestimate subject search type numbers. Nevertheless, the subject search prevalence is beyond question.

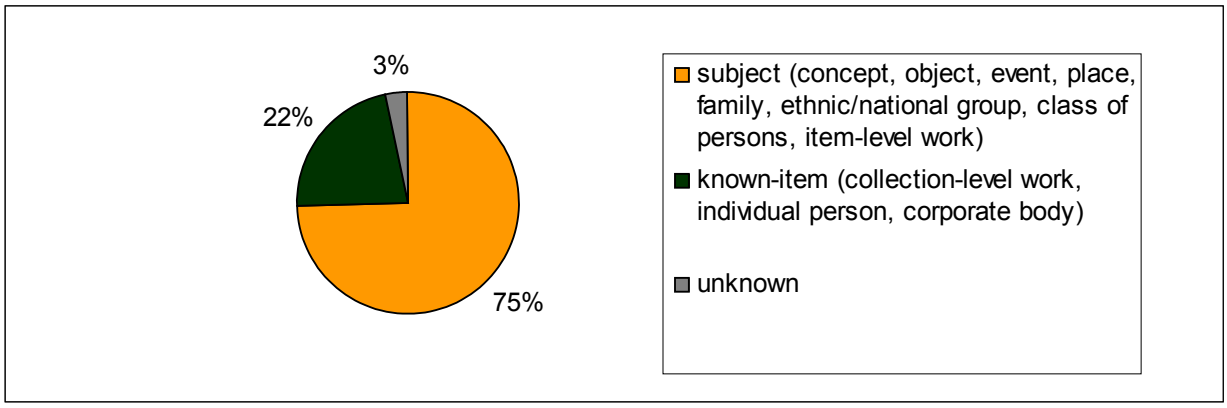

Figure 3. Unique search terms distribution by major search types (adjusted for work category)

The prevalence of a subject search (75%) demonstrated by this analysis remains in agreement with the results of the 1982 large-scale Council for Library Research (Matthews, Lawrence, & Ferguson, 1983) study of online catalog use, which radically changed the accepted understanding of the previously lower degree of subject search performed by patrons — 59% of all searches. Compared to the earlier transaction log studies of online catalog use (e.g., Larson, 1991B), including the Council for Library Research study itself, the relative value of subject search as shown by the current study is much higher, which can be explained by at least two reasons:

1. a general shift towards subject searches in a world where abundance of publication makes it less and less possible to know the title or author of the specific item
2. a conceptual difference between collection-level and item-level searches, which implies a trend towards increased levels of subject search in federated collection registries.

The DCC project team's ongoing research into how search type distribution in the Item-level Repository and Collection Registry correlate with each other will help to shed light on these and other possible reasons for dramatic subject search growth.

It should be noted here that the actual searches conducted by users in the Registry rarely could be categorized "strictly" as any one of the search categories, and sometimes presented a real challenge in determining which category was the major component of a query. This fact could be attributed both to polysemy and ambiguity of the user queries and to incompleteness of the FRBR model selected for categorization. Some of the interesting examples found in this transaction log data are discussed below:

- "Amusement park". As an abstract idea of amusement parks this query might be categorized as a *concept* search. On the other hand, amusement parks are physical structures created by people — *objects* in FRBR definition. Even asking a user what (s)he meant when making this search would not clarify this ambiguity in some cases. This actual search might have even been for a specific institution, thus a *corporate body*. Examples of similar queries from the dataset studied include "Ballrooms", "Highways", "interstates", "detroit+historical +museums" (the latter is also inseparable from a specific *place*, as is "library Moorhead").
- "Industrial models". The very word "models" implies a *concept* search, as modeling requires conceptualization. On the other hand, industrial models are physical structures created by people to assist in specific industrial processes; therefore this search can also be categorized as an *object*. "Lesson+plans" and "dissertations" are similar examples from other realms — education and academia rather than industry.
- "Landscape" is something that exists in the nature, or alternatively can be created by people, thus an FRBR *object*. However, the possibility exists for it to be classed as a *concept* too, if a user is searching for literature on landscapes and landscaping as a discipline.
- "Letters+from+19th+century" is a pretty straightforward example of an *object* search. However, it is qualified by a specific time period — a FRBR *event*.

- "asian+American" is an *ethnic/national group* search. However, it is inseparable from two *places* — Asian and American continents. Based on observations from this dataset, an *ethnic/national group* search is often defined through *place*. Similarly, "children+that+are+ abused" is also a *class of persons* defined by *event* of abuse rather than by *place*.
- "henry+fordmuseumand+greenfiel+village" is a specific *corporate body*. However a *person* (Henry Ford) and a *place* (Greenfield Village) are integral parts of this query.
- "don+quijote" is both a fictitious character created by Cervantes and a phrase widely known as a title of his work — although in fact it is just a part of the work's title. "Tom+Sawyer" is another example of this type of a query where categorization entirely depends on the user intention, which cannot be determined from the query itself. If the user searched for a book, it was a *work* search, while if the user searched for its character it was either a *concept* (something abstract that does not exist and never physically existed), or a *person* if we follow the logic of "what it would be if it existed". Similarly, a query "blimp" can be categorized either as a *person* (fictitious character Colonel Blimp) or an *object* (type of airship) search.
- "Civil rights movement" might be classified as an *event* — a complex entity which according to Functional Requirements to Subject Authority Records Working group (Zeng & Salaba, 2005) is a combination of place and time. But where is time and place in this specific query? It may equally refer to various times and places, e.g., 1950s United States, or 1960s France, or 1970s Soviet Union, or 2000s China. Does the absence of explicit or implicit qualifiers make it a *concept*? "Census" seems to belong to the same group of examples.

Typical user query profile

In regards to the typical IMLS Digital Collection registry query profile, the first stage of analysis demonstrated that user keyword queries vary in complexity and length. The number of words in each query ranges from 1 to 7, with the vast majority consisting of one or two words, as can be seen in Table 1 below. 53 % of the user queries were single-word queries. The average query length constituted 1.67 words per query.

Table 1. Query length distribution

Number of words (excluding stop words) in user query	Frequency	Percentage
1	362	53.08
2	220	32.26
3	70	10.26
4	23	3.37
5	4	.58
6	2	.29
7	1	.15

Table 2. Unique search term use distribution

Number of times unique user query used	Frequency	Percentage
1	541	79.33
2	95	13.93
3	22	3.23
4	13	1.91
5	2	.29
6	3	.44
7	1	.15
9	2	.29
10	1	.15
11	1	.15

13 1 .15

Studies of transaction logs typically also look at search term use frequency. For the sample of queries analyzed in this study, the average frequency of term use was rather low — 1.4. In fact, in over 79% of cases, the unique search term was used only once (Table 2). Quantitative characteristics of the typical user query in the Registry seem to be in agreement with results of the transaction log studies done on OPACs.

Figure 4 below illustrates correlation between the search category, average frequency of term use, and average number of words per query. The highest search term use frequency was recorded for *ethnic/national group* category — 1.70 — and the lowest for *unknown* category — 1.23. The highest average number of words per query was recorded for *corporate body* — 2.5 — and the lowest for *family* search category — 1.00 words per query.

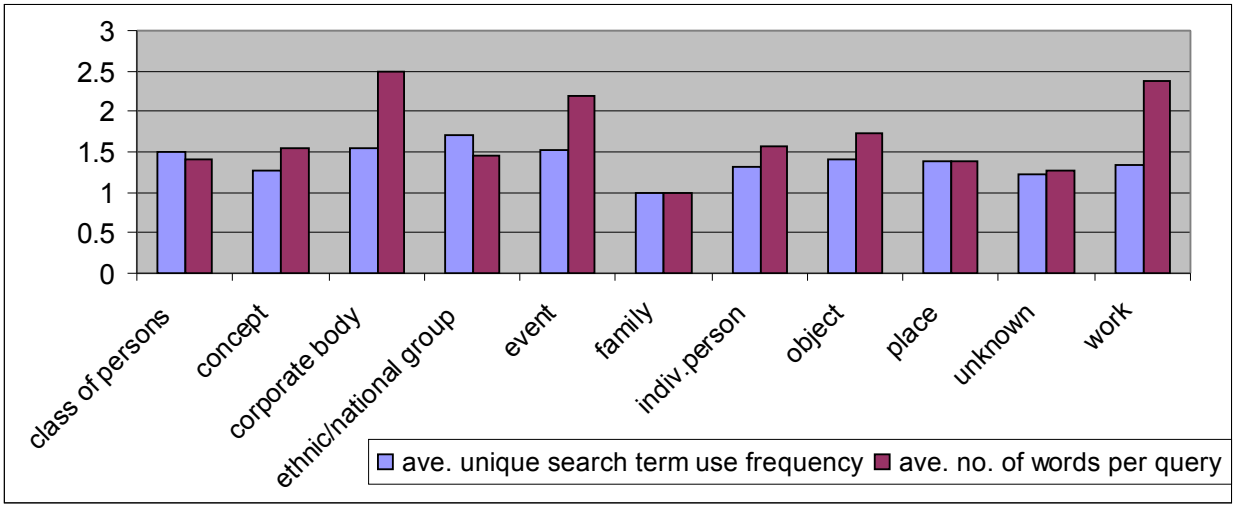

Figure 4. Unique search term use and number of words per query by the search category

Semantic similarity measures

At the second stage of analysis, the number of matches for user search queries in the three controlled vocabularies — GEM subject scheme, LCSH, and AAT — was compared for each unique search term (a set of instances of a specific user query), for each search category, and for the whole sample of queries. A total of 15 matches — 2.2% out of 682 unique search terms — were found in GEM subject scheme. A total of 495 matches — 72.6% — were found in LCSH. The AAT matched 179 unique search terms — 26.3% of user queries.

Table 3. Semantic match between the user queries and three controlled vocabulary terms

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/national 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**

As can be seen from the Table 3 above, the only user search category that GEM had matches to was *concept*, while LCSH had matches to all the categories, including a couple of *unknown* searches. The Art and Architecture Thesaurus terms matched mostly *concepts* and *objects*, with no matches at all in *corporate body, place* and *work* search categories.

The lack of semantic similarity between the user search terms and the GEM subject terms is best of all explained by the extreme broadness of this subject scheme, which might still be suitable for browsing but does not satisfy search functionality at the collection level.

There is no widely shared notion of the digital collection even among collection creators and managers (Lee, 2000; Palmer et al., 2006); much more confusion likely exists among the end-users of federated collection repositories. Such ambiguity can cause unjustified preciseness and narrowness in collection-level search terms selected by the Registry users, who are not making a distinction between searching for items in collection and searching for collections in a collection registry. Whatever the reason, the mismatch between the GEM subject scheme and actual user searches at the collection level is obvious.

The well-developed, up-to-date, flexible and faceted AAT, which seems to be especially suitable for describing cultural heritage materials and collections, performed better but still matched only slightly over a quarter of user search terms, possibly due to the fact that it does not include name and place authority files. A better means would be to incorporate broader Getty Thesaurus framework.

LCSH demonstrated the highest level of semantic match with user queries. These results are in line with some earlier studies (e.g., Carlyle, 1989) which found strong match at a concept level. Although matching most of the user terms, LCSH still leaves over 27% unmatched. This subject scheme was the most effective in matching *places* and *concepts*, while *works* remained the least matched; only about a half of the *corporate bodies* and *events* were covered by LCSH terms. The reason may lie in the general inflexibility of LCSH — a large scheme that is extremely hard to keep up-to-date. A vivid illustration is the absence of terms such as “learning standards” in the LCSH authority file.

However, as can be seen from the Table 4 below, LCSH on its own (without any overlap with AAT or GEM) covers 48% of the user search terms. Only 12 terms matched in AAT were not also matched in LCSH, while all the terms matched in GEM were also matched in LCSH. Slightly over one quarter of user search terms were not matched in any of the three controlled vocabularies.

Table 4. Semantic match in single vs. multiple controlled vocabularies

Search category	**unique search terms**	**GEM alone**	**GEM and LCSH**	**GEM and AAT**	**LCSH alone**	**LCSH and AAT**	**AAT alone**	**All**	**none**
class of persons	12	0	0	0	3	7	0	0	2
Concept	146	0	5	0	44	69	5	10	13
corporate body	24	0	0	0	17	0	0	0	7
Ethn./nat. group	33	0	0	0	14	15	0	0	4
Event	25	0	0	0	7	2	1	0	15
Family	5	0	0	0	4	0	0	0	1
Object	166	0	0	0	56	62	6	0	42
individual person	90	0	0	0	72	0	0	0	18
Place	103	0	0	0	98	0	0	0	5
Unknown	22	0	0	0	4	0	0	0	18
Work	56	0	0	0	7	0	0	0	49
TOTAL	**682**	**0**	**5**	**0**	**326**	**155**	**12**	**10**	**174**

Conclusions

These study results demonstrate a high level of subject searching at the collection level which is unusual for catalog use / transaction log analysis studies. Further investigation is needed into the reasons of such prominence of subject search, including collection of the data through interviews and observations of the Registry users. The user conceptualization of the collection-level search and its possible difference from the concept of the item-level search also needs investigation.

A productive next step for the DCC project will be to explore which combination of vocabularies would optimally represent digital collections in the Registry as well as in other cultural heritage domain federated collections. Although LCSH has demonstrated strong results, none of the three controlled vocabularies in this study fully represents the subjects of diverse collections in the Registry, or at least user expectations towards these subjects. Additional study needs to investigate more flexible than LCSH controlled vocabularies of the moderate scale, which, unlike GEM or AAT, represent a wide variety of search categories. To compensate for deficiencies of transaction log analysis think-aloud protocol observation of the users searching the Registry should be incorporated into further analysis to provide insights into users' motivations and intentions in selecting search terms.

It has been noted (e.g., Bates, 2002) that the larger the size and complexity of the collection the higher the level of sophistication of the subject scheme is required for adequate description. A strong semantic match to user queries offered in this study by a traditional library subject scheme — Library of Congress Subject Headings — supports this principle for the federated digital resource environment and for collection level description. This suggests that to provide adequate search functionality federated collection developers will need to retain very narrow subject scope — which is highly unlikely — or to place significant efforts into selection and testing of highly-developed subject schemes for collection-level description. Although LCSH was not significantly complemented by the two other controlled vocabularies in matching user search terms in this study, the combination of two or more standardized controlled vocabularies for subject description at the collection level shows promises for facilitating subject access to collections in the federated environment.

Acknowledgements

This research was supported by a 2002 IMLS NLG Research & Demonstration grant. Project documentation is available at http://imlsdcc.grainger.uiuc.edu/.

References

Bates M. (2002). After the dot-bomb: getting Web information retrieval right this time. *First Monday*, *7*(7). Retrieved December 27, 2006, from: http://www.firstmonday.org/issues/issue7_7/bates/index.html.

Carlyle, A. (1989). Matching LCSH and user vocabulary in the library catalog. *Cataloging and Classification Quarterly, 10*(1/2), 37-63.

Cochrane, P. (1986). *Improving LCSH for Use in Online Catalogs*. Colorado Springs: Libraries Unlimited.

IFLA Study Group on the Functional Requirements for Bibliographic Records, & International Federation of Library Associations and Institutions. (1998). *Functional Requirements for Bibliographic Records: final report.* München: K.G. Saur.

IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR) (2007). *Functional Requirements for Authority Data: A Conceptual Model*. 2nd draft. Retrieved April 12, 2007, from http://www.ifla.org/VII/d4/FRANAR-ConceptualModel-2ndReview.pdf

Jackson, S. (1958). *Catalog use study: director's report*. Chicago: American Library Association. Cataloging and Classification Section. Policy and Research Committee.

Krikelas, J. (1972). Catalog use studies and their implications. *Advances in Librarianship, 3,* 195-220.

Larson, R. (1991A). Between Scylla and Charybdis: Subject searching in online catalogs. *Advances in Librarianship, 15,* 175-236.

Larson, R. (1991B). The decline of subject searching: long-term trends and patterns of index use in an online catalog. *Journal of the American Society for Information Science, 42*(3), 197-215.

Lee, H. (2000). What is a collection? *Journal of the American Society for Information Science, 51*(12), 1106-1113.

Lee, J., Renear, A., & Smith, L. (2006). Known-item searching: Variations on a concept. *Proceedings of the 69th ASIS&T Annual Meeting*, 3-8 November 2006, Austin, Texas.

Matthews, J., Lawrence, G., & Ferguson, D. (Ed.) (1983). *Using online catalogs: A nationwide survey*: *A report of a study sponsored by the Council on Library Resources.* New York: Neal-Schuman.

Palmer, C., Knutson, E., Twidale, M., & Zavalina, O. (2006). Collection definition in federated digital resource development. *Proceedings of the 69th ASIS&T Annual Meeting*, 3-8 November 2006, Austin, Texas.

Sutton, S. (2004). Building an education digital library: GEM and early metadata standards adoption. In. D. Hillmann, E. Westbrooks (Ed.), *Metadata in Practice,* 1-15, Chicago: American Library Association.

Yee, M., & Layne, S. (1998). *Improving online public access catalogs.* Chicago: American Library Association.

Zeng, M., & Salaba, A. (2005). Toward an international sharing and use of subject authority data. *FRBR Workshop*, OCLC, 2005. Retrieved April 7, 2006, from http://www.oclc.org/research/events/frbr-workshop/presentations/zeng/Zeng_Salaba.ppt.